GEOPOLITICA DELL'INTELLIGENZA ARTIFICIALE _ Aresu, 2024, Feltrinelli[1]

 # Premessa

L'intelligenza artificiale è l'invenzione definitiva dell'umanità. La sua comparsa sulla scena evoca il rischio dell'estinzione del suo creatore. La sua diffusione, come forma di intelligenza generale e superiore, compirà il destino dell'umanità perché porterà al suo superamento.

Visto da vicino, il dibattito sull'intelligenza artificiale chiama in causa alcuni concetti, tra cui: la stessa intelligenza, ciò che sappiamo e ignoriamo del cervello e del nostro pensare; l'idea di un'intelligenza "generale" applicata alle macchine, la sua ricerca e le sue ossessioni; i limiti quantitativi e qualitativi del calcolo; il problema dell'allineamento della tecnologia ai nostri bisogni e ai nostri valori.

L'intelligenza artificiale al servizio dell'umanità, un approccio umano al problema, la centralità del fattore umano: le variazioni sul tema sono innumerevoli. Siamo assillati da questo marketing del concetto di umanità, anche attraverso un florilegio di proposte di governance. Ma cosa vuol dire tutto ciò? E quali sono le sue implicazioni per un mondo radicalmente diviso, segnato dalle tensioni della guerra tecnologica tra Stati Uniti e Cina, che coinvolgono anche le infrastrutture e gli usi dell'intelligenza artificiale, con conseguenti rischi per la sicurezza nazionale?

Non esiste una storia dell'intelligenza artificiale che procede in modo omogeneo. C'è invece un intreccio delle sue storie in cui emerge un protagonista.

Lo spirito del capitalismo, la rivoluzione incessante della struttura economica dall'interno, si manifesta in un ragazzo nato a Taiwan, che si trasferisce negli Stati Uniti a dieci anni, e da uomo indossa sempre un giubbotto in pelle che cela il tatuaggio del logo della sua azienda, NVIDIA, nata nel 1993, di cui è da sempre amministratore delegato e "uomo immagine". Tutti lo chiamano per nome: Jensen.

Quando a cinquant'anni suonati parla a migliaia di appassionati di videogiochi, ricorda che le finali del campionato mondiale del videogioco League of Legends hanno più spettatori delle finali NBA e commenta esaltato: "Amo LeBron ma noi siamo più grandi di LeBron! Siamo più grandi di Jordan!".

La grande trasformazione è iniziata dai videogiochi, ma per ogni mercato potranno essere realizzati prodotti e soluzioni. Ogni mercato potrà essere trasformato, rivoluzionato dall'interno.

Quasi tutti gli attori di questa rivoluzione tecnologica sono immigrati; l'Europa, che non è in grado né di attirare o trattenere i principali talenti, i quali si muovono sempre di più verso il magnete del Nord America, né di costruire e realizzare i componenti delle infrastrutture di calcolo e dei loro mercati di riferimento, come avviene soprattutto in Asia orientale, nuovo motore manifatturiero del mondo, ma non nuovo approdo di ricercatori e imprenditori. Almeno, questo è il percorso prevalente, anche se sono possibili alcune varianti o "eresie".

Il sistema pubblico, incapace di pagare e sostenere un'infrastruttura sempre più legata a soluzioni e potenzialità commerciali, viene progressivamente svuotato, e in un certo senso "schiacciato" verso il monopolio legittimo della violenza, attraverso l'allargamento della sicurezza nazionale, che, da ultimo, si pone in contrasto con gli interessi commerciali.

L'infrastruttura di calcolo alimenta una continua accelerazione, e su di essa si basa il lavoro di programmi sempre più evoluti che sfruttano una nuova disponibilità di dati, resa possibile dalla diffusione e dalla pervasività di una rete con miliardi di creatori e fruitori di testi, immagini e filmati.

 # Parte prima. Dal Kentucky a Lady Gaga
1. Un lavoro vero

Nel 1973, a Oneida, Kentucky, compaiono due ragazzini cinesi. Oneida si trova tra i fiumi e il nulla. I suoi abitanti, negli ultimi censimenti, sono poco più di duecento. Una sola scuola americana si dichiara disponibile ad accogliere i due ragazzini venuti da lontano: il riformatorio di Oneida, Kentucky. Così Jen-Hsun Huang1, a dieci anni, si ritrova

[1] Riassunto per ECCOCI! da Gigi Bacchetta; segnalazione errori: gigi.bacchetta@cgilpiemonte.it

in un tratto sperduto di America, e impara a pulire i bagni cinquant'anni prima del momento in cui la sua azienda, il motore dell'intelligenza artificiale, supererà i 1000 miliardi di dollari di capitalizzazione alla borsa di New York.

Si fa chiamare Jensen: un nome più semplice, per farsi capire non solo dal veterano con cui gli piace chiacchierare, ma anche dal suo compagno di stanza, un diciassettenne uscito dalla galera. Col reduce del Vietnam, Jensen costruisce una specie di intesa. Le parole che gli erano estranee, le parole dell'America, gli entrano in testa, una dopo l'altra.

Quando il padre di Jensen prende la decisione di mandare i figli a New York, la madre regala un dizionario ai due ragazzini, con un compito ben preciso: ogni giorno, Jensen e suo fratello dovranno imparare dieci parole, per prepararsi a New York. All'inverno di Oneida si affianca un altro inverno, in parallelo: quello dell'intelligenza artificiale.

L'inizio della storia dell'intelligenza artificiale dipende dalla sua – difficile – definizione. Come ogni storia, affonda nel mito. Il Prometeo di Eschilo, prima ancora di fornire il fuoco agli uomini, dà loro il numero (arithmos), che rende possibile la misura del fuoco, come quella di ogni tecnica, di ogni macchina. "Per loro scoprii il numero, la prima conoscenza", afferma il Titano. Dopo il mito, è una storia che si colloca nella traiettoria della logica e del pensiero formale, da Aristotele a Gödel, passando per Leibniz e Boole. È una storia di sviluppo di macchine, della riproduzione e dell'automazione di processi mentali e fisici.

I lavori di Alan Turing, John von Neumann e Claude Shannon. Le loro ricerche sono fondamentali, tra l'altro, per definire l'intelligenza artificiale rispetto all'uomo attraverso un test, che Turing chiama "gioco dell'imitazione", nonché per formalizzare l'architettura dei computer e per lo studio scientifico dell'informazione. Ciò avviene mentre l'invenzione del transistor ai Bell Labs, nel 1947, rende possibile l'era digitale, attraverso la creazione di computer sempre più piccoli, veloci ed efficienti. In questo modo, le ricerche di von Neumann sull'architettura delle macchine possono essere riprodotte in modo sempre più accurato nella pratica.

Mentre l'inizio della guerra fredda consente l'aumento delle risorse per la ricerca di frontiera, gli anni cinquanta avvolgono l'intelligenza artificiale in un clima di ottimismo ed elevate aspettative, codificate da alcuni studiosi (John McCarthy, Marvin Minsky, Nathaniel Rochester e lo stesso Claude Shannon) nella cosiddetta "proposta di Dartmouth" del 1955, un manifesto del nuovo campo di studi, basato "sulla congettura che ogni aspetto dell'apprendimento o ogni altro aspetto dell'intelligenza possa in principio essere descritto in modo così preciso da realizzare una macchina in grado di simularlo".

C'è l'idea che la comprensione e riproduzione dell'intelligenza umana attraverso le macchine possa essere risolta nel breve-medio periodo, attraverso la soluzione di alcuni problemi di ingegneria di calcolo.
Nel 1957, lo psicologo Frank Rosenblatt inventa il percettrone, con l'obiettivo di realizzare una macchina in grado di imparare, attraverso le connessioni tra i suoi "neuroni", programmi che imitano le unità fondamentali del cervello umano. L'attivazione della "rete neurale" in risposta ad alcuni segnali indica la possibilità di "apprendimento". I dati consentono di imitare attività percettive come il riconoscimento delle immagini o la scrittura.

La "macchina pensante", il percettrone, secondo Rosenblatt ci mostra "la possibilità di costruire cervelli in grado di autoriprodursi in una catena di montaggio e che saranno coscienti della propria esistenza". Nel 1969 Marvin Minsky e Seymour Papert mostrano i limiti di questa pretesa di apprendimento in un libro chiamato Perceptrons, tra le cui pagine operano una spietata demolizione del processo. Uno dei problemi più significativi è quello della funzione XOR, un tipo di operazione logica.

Per capire di che si tratta, immaginiamo un gioco: se vengono mostrate due carte e si vince solo se una carta è rossa e l'altra è blu, il percettrone non riesce a capire la regola. Non sa gestire situazioni dove la risposta corretta dipende dalla combinazione di più elementi. Chi è incapace di compiere simili operazioni, di giocare a un gioco così semplice, non potrà mai vincere il gioco dell'imitazione di Turing, verso il Santo Graal dell'imitazione umana.

Nel 1973 Lighthill descrive l'ABC dell'intelligenza artificiale facendo riferimento a tre categorie che attribuisce a un "consenso generale" ma che in realtà organizza in modo personale: A (Advanced Automation), B (Bridge, ma anche Building Robots), C (Computer-Based Central Nervous System). L'obiettivo della categoria A è la sostituzione degli uomini con le macchine per vari scopi concreti e utili, di natura sia industriale e militare sia scientifica. La categoria C ha uno scopo biologico: la costruzione di reti neurali come modelli del cervello e del sistema nervoso, per svolgere funzioni umane come il riconoscimento delle immagini e il linguaggio. La categoria B è la base del ragionamento di

Lighthill, e del suo approccio distruttivo, perché la coerenza dell'intero edificio dell'intelligenza artificiale si regge a suo avviso su questo "ponte": la realizzazione di robot in grado di imitare alcune attività umane, di riconoscere oggetti, di risolvere i problemi in modo convincente. "La maggior parte di coloro che sono impegnati nella ricerca sull'intelligenza artificiale e in campi a essa legati confessano un marcato sentimento di delusione per quanto realizzato negli ultimi venticinque anni."

Il campo di ricerca che Lighthill considera inaugurato nel 1947 con gli studi di Alan Turing, al compimento del suo venticinquesimo anno, non offriva particolari prospettive per i successivi venticinque anni. In estrema sintesi: i soldi pubblici potevano andare altrove.
Nel dibattito, trasmesso dalla BBC, McCarthy ammette anche perché ha inventato nel 1955 quell'espressione: per avere fondi. Nel 1952, Claude Shannon gli aveva detto di non utilizzarla in quanto troppo appariscente, in favore di un più semplice "automi", che non riesce ad attirare nemmeno i paper per le loro conferenze. Parlare di "intelligenza artificiale", con "l'obiettivo a lungo termine di raggiungere un'intelligenza di livello umano", serve per attirare attenzione e soldi. È un'intuizione di marketing.

Dopo i primi anni di apprendistato in Kentucky, Jensen e suo fratello sono raggiunti negli Stati Uniti dai genitori. Non a New York: prima nello stato di Washington, poi nell'Oregon, a Portland. Jensen grazie ai suoi risultati scolastici riesce a saltare un paio di classi e si dedica in parallelo a una grande passione: il ping pong.

Nel 1984 l'immaginario delle macchine in grado di ribellarsi agli uomini conosce un'epocale fortuna grazie a James Cameron, con l'inizio della saga di Terminator, nata da un incubo del regista durante un suo soggiorno a Roma. Nella cultura popolare, l'intelligenza artificiale da quel momento in poi sarà Skynet, nel motivo della ribellione delle macchine, del controllo delle macchine. Con la variante di macchine uccise da altre macchine, più o meno in combutta con un'umanità che non può fare a meno delle sue creazioni, non può pensare semplicemente di staccare la corrente.

Jensen, per costruire una corazza nella sua vita americana, ha iniziato in Kentucky a sollevare pesi. E a un certo punto, dopo l'incontro con Lori, comincia a indossare un giubbotto in pelle da motociclista, che diviene la sua armatura, la sua firma. Soprattutto nelle occasioni più formali, come per dire "questo sono io" a ogni suo interlocutore.

Dovete seguire i soldi. Chi costruisce questi modelli di intelligenza artificiale? Li stanno costruendo per dominare le quote di mercato, quindi insegnano loro l'avarizia. O li stanno costruendo a scopi di difesa, quindi insegnano loro la paranoia. Penso che l'uso dell'intelligenza artificiale come un'arma sia il nostro maggiore pericolo," afferma James Cameron nel 2023. "Vi avevo avvisato nel 1984, ma non mi avete ascoltato."

Dopo la laurea di Lori, Jensen si sposta di nuovo. Non a New York, ma nel luogo dove gli studi di ingegneria elettronica divengono imprese e innovazioni: la Silicon Valley, la "valle del Silicio". Così il ragazzo del Kentucky inizia a lavorare per alcune aziende di semiconduttori, prima per AMD, realtà già affermata, e poi per LSI Logic. Quest'ultima ha iniziato la sua attività nel 1981 grazie al finanziamento di Sequoia Capital, società pionieristica del venture capital, il capitale di rischio per finanziare le start-up tecnologiche. Il fondatore di Sequoia, Don Valentine, ha imparato il mestiere sul campo, andando "alla Fairchild Semiconductor Business School", ed è una vera leggenda: in meno di dieci anni, da quando Jensen arriva in Kentucky, aiuta a creare giganti come Atari, Apple, Electronic Arts.

LSI Logic basa il suo successo sulla tecnologia ASIC (Application Specific Integrated Circuit): si tratta di circuiti integrati nati per rispondere non a calcoli generici, ma a scopi precisi, su cui acquisire una particolare specializzazione per venire incontro alle esigenze dei clienti. Tra questi clienti, non ci sono solo le società di elettronica ma anche, e sempre di più, le aziende militari e aerospaziali. La sfida della competizione della difesa diviene anche una sfida microelettronica: quel mercato vale, all'inizio degli anni novanta, il 10% del fatturato di LSI Logic. Il distretto militare di Orange County dipende dalle soluzioni dell'azienda dove Jensen lavora e si fa apprezzare, con quello strano giubbotto in pelle.

Dall'altro lato del Pacifico, a metà degli anni ottanta, il Giappone sfida gli Stati Uniti per il primato nei semiconduttori. Un veterano dell'industria americana, Morris Chang, nel 1987 inizia a realizzare a Taiwan l'idea che cambierà tutta quella filiera: separare la progettazione dalla produzione, con la creazione della Taiwan Semiconductor Manufacturing Company (TSMC).

L'obiettivo è andare oltre lo schema tradizionale: l'architettura inventata da John von Neumann, in cui la memoria registra le informazioni e un processore centrale fa i calcoli in sequenza. "Noi preleviamo un primo numero da una certa postazione di memoria, un secondo da un'altra, li mandiamo nell'unità aritmetica centrale, li sommiamo e inviamo il risultato in qualche altra postazione."

In questo modo, osserva Feynman, la memoria è come "uno schedario in cui le informazioni sono utilizzate molto di rado". Invece, la velocità del calcolo sarebbe senz'altro aumentata da processori in grado di lavorare contemporaneamente: il calcolo parallelo "può abbreviare drasticamente il tempo di risoluzione ogni volta che la dimensione del problema rende necessaria una gran mole di calcoli; e questo vale non solo per i problemi scientifici".

Può esserci un "ingegnere" convinto di sapere come funzioni il cervello e che progetti una macchina in grado di imitarlo. "Ma, attenzione, questo non ci dice nulla su come funziona realmente il cervello, e d'altra parte, per costruire un calcolatore davvero efficiente, non è nemmeno indispensabile saperlo. Non è necessario conoscere il modo in cui gli uccelli battono le ali o la struttura delle loro piume per costruire macchine volanti. Non è essenziale capire il sistema di leve nelle zampe di un ghepardo per fabbricare un'automobile da corsa. Non è quindi necessario imitare nel dettaglio il comportamento della natura per progettare un dispositivo che possa superare, per molti aspetti, la natura stessa.

Le macchine rispondono a un problema, eseguono un compito in modo efficace o meno, e su questo vanno giudicate, senza portare troppo in là l'analogia. Del resto, su questo metro di paragone, il calcolatore è già molto meglio dell'uomo. Allo stesso tempo, è molto peggio. "Un calcolatore può prendere decine di migliaia di numeri e ridarmeli in ordine inverso, sommarli, o fare un mucchio di cose che per noi sono impossibili. D'altra parte, a me basta un'occhiata per riconoscere la faccia di una persona; mentre nessun calcolatore è in grado di farlo, neppure se gli sono state mostrate molte facce e si è cercato di insegnargli a riconoscerle."

All'inizio degli anni novanta, il personal computer sta raggiungendo una diffusione di massa. Microsoft presenta le prime versioni del sistema operativo Windows con grande successo commerciale. Il sogno iniziale di Bill Gates, "un computer in ogni scrivania e in ogni casa", all'inizio sembra assurdo, poi sempre più a portata di mano. Con quali conseguenze? Da un lato, le economie di scala del personal computer possono permettere uno sviluppo senza precedenti per l'hardware e il software, per le aziende che diverranno dominanti in quei mercati. E più in profondità, il personal computer ha il potere di cambiare il modo con cui si comunica, con cui si lavora. Non solo: altri mercati, a prima vista meno evidenti, stanno affacciandosi all'orizzonte della generazione che cresceva col personal computer, per cui quell'oggetto può diventare sempre più familiare e stabilire una nuova relazione con il loro tempo, con le loro giornate.

LSI ha un grande futuro davanti a sé, con quest'opportunità. Ma forse, dietro l'angolo, c'è qualcosa di ancora più grande. Ed è ciò che Jensen inizia a vedere parlando con i suoi clienti di Sun Microsystems quando si ritrovano nel 1992, ovviamente da Denny's. Jensen è ormai diventato l'uomo di riferimento in LSI Logic per i chip dedicati alle workstation di Sun, i computer ad alte prestazioni utilizzati per applicazioni professionali, che al tempo costano 10.000-20.000 dollari e che vendono varie migliaia di esemplari. Non è abbastanza.

Le stesse workstation, tra i loro vari scopi, sono legate anche allo sviluppo di una grafica, ancora rudimentale, per i videogiochi; il personal computer sta diventando questo: la grande massa, l'economia di scala. Malachowsky e Priem convincono Jensen, seduti al tavolino di Denny's: c'è un enorme mercato, un popolo di videogiocatori, pronto a comprare milioni, decine di milioni di giochi, per giocare con i milioni e milioni di personal computer che sarebbero inevitabilmente finiti nelle loro case. Una maggioranza silenziosa che vuole vivere e progettare dentro mondi virtuali, creati da programmatori come John Carmack e John Romero, che proprio nel 1993 con la loro id Software lanceranno lo sparatutto Doom.

Con NVIDIA il sogno di una nuova azienda, il sogno di un nuovo mondo, si spalanca davanti a Jensen; il ragazzo del Kentucky sente il bisogno di parlarne a qualcuno: sua madre. "Mamma, penso di lasciare tutto per fondare un'azienda che faccia schede grafiche per i videogiochi." "Perché invece non ti trovi un lavoro vero?"

Ammazzarsi di lavoro, pulire bagni a dieci anni, fare l'università a sedici, per poi finire a pensare ai videogiochi, mentre si ha la responsabilità di crescere dei figli? Com'è potuto accadere? Cos'è andato storto? Ingurgitare il cibo del fast food, e in particolare la colazione "Grand Slam" di Denny's, gli ha forse rovinato il cervello?

Prima di arrivare a New York, a suonare la campanella della borsa, anche Lam ha dovuto affrontare una visita alla madre. Ha raccontato questa storia commosso al Museo dei Cinesi in America, quando ha ricevuto un'onorificenza per i suoi eccezionali contributi in quella comunità. A quasi quarant'anni Lam decide di fondare la sua azienda, dopo aver lasciato l'Asia per l'Università di Toronto negli anni sessanta e aver lavorato in giganti come Texas Instruments e Hewlett-Packard. Sua madre gli chiede di leggere il business plan. La signora Lam ha una comprensione limitata dell'inglese e non riesce a capire il senso dell'azienda, come del resto qualunque profano che si accosti al segmento dell'industria dei semiconduttori che David Lam intende occupare: la strumentazione per l'incisione al plasma.

Per la loro azienda, Jensen e i due soci hanno anzitutto bisogno di un nome. Partono dalla loro abitudine di chiamare le schede grafiche a cui lavorano "NV", Next Version. L'accostamento inusuale di quelle due lettere fa scattare qualcosa e li porta a cercare sul dizionario le parole che le contengono, tra cui il latino invidia, a cui tolgono la "i" iniziale: NVIDIA. Per la filosofa María Zambrano, un "inferno terrestre" è l'invidia. "Una distruzione che si alimenta da sé: questa sembra essere la prima, originale, definizione di invidia." Ma c'è un'altra distruzione che si alimenta da sé, all'interno della burrasca in cui vive: la distruzione creatrice di Joseph Schumpeter, in cui si colloca l'innovazione. L'invidia è in-videre, guardare di traverso. Ma è uno sguardo: e gli utenti, i videogiocatori, dovranno iniziare a vedere mondi attraverso le schede grafiche.

Le operazioni dell'economia del silicio, fondate sul processore centrale (CPU, Central Processing Unit), attorno a cui già ruota un mondo di calcoli e di prodotti, non funzionavano in modo troppo diverso. Tanti compiti da fare, uno dopo l'altro. Nei limiti dell'architettura fondata da John von Neumann, il segreto è saper svolgere ogni compito nel modo più veloce e con il consumo minore di energia.

Chi gioca vuole avere la migliore esperienza di quel mondo, vuole vederlo al meglio grazie ai poligoni, la grafica che è il risultato del modo con cui i programmi effettuano i loro calcoli, e ha bisogno di uno strumento che lo renda possibile, un nuovo cervello di calcolo, la GPU (Graphics Processing Unit). Chi gioca, chi abita il mondo del gioco, vuole vivere in parallelo, come risposta continua e complementare rispetto agli stimoli che riceviamo.

Come le operazioni che costruiscono i poligoni dei videogiochi che, ammassati l'uno sull'altro, sulla base di circuiti integrati ammassati allo stesso modo l'uno sull'altro, fanno riconoscere al sistema visivo e al sistema nervoso un pezzo di grattacielo, e così attivano il percorso della memoria: immaginate di essere in una grande cucina industriale negli anni ottanta, quando si facevano i biscotti uno alla volta; questo è il calcolo seriale. Con l'avanzare del tempo, intorno agli anni novanta, qualcuno ha avuto l'idea di usare più forni contemporaneamente per cuocere più biscotti in una volta – questo è l'inizio del calcolo parallelo. Ora, NVIDIA entra in scena nel 1999, e come un pasticciere visionario, porta in cucina un robot (la GPU) capace di decorare simultaneamente centinaia di biscotti. Questo ha rivoluzionato non solo la pasticceria ma anche il mondo dei videogiochi.

Prima, i giochi avevano grafiche semplici perché il "pasticciere" (CPU) doveva gestire tutto da solo. Con l'arrivo della GPU di NVIDIA e la sua capacità di calcolo parallelo, improvvisamente si poterono creare mondi di gioco con grafiche complesse e dettagliate, ogni "decorazione" (elemento grafico) veniva elaborata in parallelo per creare un'esperienza immersiva che prima sarebbe stata impensabile.

I soldi alimentano il calcolo tanto quanto la tecnologia. Se c'è un mercato abbastanza grande, qualcuno svilupperà il prodotto. I tre di Denny's hanno visto il mercato. A quel punto, per accelerare il loro gioco, hanno bisogno di soldi. La prima valutazione di NVIDIA è di 6 milioni di dollari ma tra gli investitori c'è il grande nome capace di attirare l'attenzione: Sequoia. Wilfred "Wilf" Corrigan, immigrato dal Regno Unito dopo la laurea in ingegneria chimica all'Imperial College, amministratore delegato di Fairchild Semiconductor, co-fondatore e amministratore delegato di LSI Logic, telefona personalmente a Don Valentine per dirgli che quel ragazzo è così in gamba da meritare i suoi soldi per la sua azienda, qualunque cosa sia. Don Valentine chiarisce subito a Jensen le regole del gioco: "Se perdi i miei soldi, ti ammazzo". Accelerare o morire.

Nel 1994 NVIDIA avvia una partnership strategica con SGS-Thomson (oggi STMicroelectronics). L'azienda italo-francese, rilanciata grazie al lavoro dell'ex manager siciliano di Motorola, Pasquale Pistorio, dà la capacità manifatturiera necessaria (soprattutto con le sue fabbriche a Malta) per la prima scheda portata sul mercato, NV1. L'obiettivo di NVIDIA è offrire un'esperienza tridimensionale per alcuni videogiochi, come Virtua Fighter della SEGA, l'azienda che in quegli anni è impegnata in una "guerra delle console" con la Nintendo. I tecnici di NVIDIA hanno l'occasione di collaborare col leggendario programmatore di SEGA, Yu Suzuki, creatore di Virtua Racing e Virtua Fighter.

1° aprile 1997: non può essere uno scherzo, perché su quel prodotto, RIVA (Real-Time Interactive Video and Animation), si basa il futuro di una società in bilico tra la vita e la morte. RIVA 128, per le sue prestazioni e il rapporto tra qualità e prezzo, distingue NVIDIA sul mercato e contribuisce al suo successo, assieme alle partnership con produttori di PC e di hardware, al marketing, al posizionamento, e al crescente interesse per i giochi 3D e le applicazioni grafiche. NVIDIA inizia a caratterizzarsi per lo sviluppo di software ottimizzati, cerca di migliorare grazie alla risposta dei consumatori. E per la produzione si affida all'azienda che negli anni novanta a Taiwan sta cambiando le regole del gioco: TSMC, fondata da Morris Chang.

L'anno decisivo è il 1999. Sta per arrivare il Millennium bug. C'è una corsa tra l'apocalisse dei dispositivi informatici, prevista per il cambio di data a fine millennio, e l'uscita di Quake 3 Arena, il nuovo sparatutto di id Software, atteso da tutti i videogiocatori. NVIDIA lancia finalmente la GPU, quando un prodotto rivoluzionario, la GeForce 256, arriva sul mercato il 31 agosto. I nerd dell'azienda forniscono una definizione tecnica della GPU: "Un processore a chip singolo con motori integrati di trasformazione, illuminazione, creazione/clipping di triangoli e rendering in grado di elaborare un minimo di dieci milioni di poligoni al secondo". La descrizione è accompagnata dalla promessa che "la GPU cambierà tutto quello che avete visto o sentito sul vostro PC" e dalla sintesi per cui "in sostanza, la GPU vi fornisce, gratis, un realismo davvero straordinario".

All'inizio del 1999 a Wall Street. La lettura del prospetto della quotazione in borsa, con un prezzo di 12 dollari, è ancora oggi istruttiva. L'azienda che vuole costruire i mondi virtuali si incontra col mondo virtuale della finanza, la condizione necessaria per la sua scommessa sul futuro.
L'azienda ha accumulato perdite ogni anno dalla sua nascita fino ai primi tre trimestri del 1997. Anche lo scenario del futuro non sembra troppo roseo: le vendite sono concentrate su un numero limitato di clienti primari (Creative, Diamond, STB) che poi vendono ad altri, sulla base della domanda del mercato. I fattori di rischio di NVIDIA sono molteplici, legati alla dipendenza da clienti e fornitori, nonché all'incertezza geografica. In particolare, "in passato utilizzava ST e attualmente utilizza TSMC per produrre i wafer semiconduttori e altre aziende indipendenti per l'assemblaggio, il test e l'imballaggio". Senza questi procedimenti, senza questa supply chain, NVIDIA non può avere un prodotto da consegnare ai clienti.

NVIDIA è un'azienda americana ma la sua dipendenza dall'estero "per le operazioni di manifattura, assemblaggio e test la espone a numerosi rischi legati alle attività fuori dagli Stati Uniti. Jensen, che ha una sola cosa in mente: Toy Story. Siamo obbligati a parlare di rischi, d'accordo, ma parliamo anche dello sceriffo e dell'astronauta.

Per una grafica 3D interattiva di alta qualità, i processori grafici 3D avanzati richiedono milioni di transistor per elaborare miliardi di operazioni aritmetiche al secondo. Gli attuali processori grafici 3D sono oltre dieci volte più complessi degli acceleratori 2D e paragonabili alla complessità dei microprocessori Pentium di Intel. Tuttavia, nonostante i recenti progressi, la grafica 3D per PC oggi disponibile non è in grado di offrire in tempo reale una grafica della qualità del film Toy Story. Quest'ultima richiedeva oltre 100 potenti workstation e oltre 800.000 ore di computer per eseguire il rendering dei 114.000 fotogrammi del film, con una media di 7 ore per ogni fotogramma. Affinché i PC tradizionali possano fornire questo livello di capacità grafica 3D, le prestazioni dei processori grafici 3D dovranno essere migliorate di molti altri ordini di grandezza. E per avvicinarsi alle prestazioni grafiche del "mondo reale", ben oltre quelle viste in Toy Story, i processori grafici richiederebbero molti altri notevoli miglioramenti.

2. Visioni del paradiso dei panda

Quando Nixon va in Cina nel 1972, Mao gli promette in dono due panda, che avrebbero allietato i bambini dello zoo di Washington. Quando Deng Xiaoping cambia la traiettoria economica della Cina, negli anni ottanta, modifica i termini della diplomazia dei panda: questi animali, orgogliosi della loro pigrizia e golosi di bambù, sarebbero stati affittati, non regalati. Nel mentre, lo studio dei panda diventa importante per la Cina.

La famiglia di Fei-Fei prende, pochi anni dopo, la stessa decisione dei signori Huang: lei deve crescere nella terra delle opportunità, l'America. In questo caso, è il padre ad andare in avanscoperta e preparare il terreno per l'arrivo di sua moglie e sua figlia, non così lontano dai grattacieli di New York: a Parsippany, nel New Jersey, dove c'è una folta comunità cinese e dove il signor Li, nonostante la sua formazione da ingegnere elettronico, si arrabatta con qualche lavoretto.

Per Fei-Fei, sbarcata nel New Jersey a quindici anni, non è facile vivere il declassamento da studentessa brillante, che con la sua curiosità vuole capire i misteri della fisica, a ultima arrivata che non sa esprimersi bene in inglese. Gli altri sembrano seguire un ritmo troppo veloce mentre le parole le affiorano alle labbra lentamente, una dopo l'altra.

Capire e farsi capire: lo scoglio che ogni straniero trova davanti a sé. Col terrore di restare indietro, di buttare al vento il sogno americano. Per Fei-Fei tutto cambia quando incontra il professor Bob Sabella, insegnante di matematica a Parsippany.

La accoglie nel suo ufficio per ripetizioni private, dove lei trova il coraggio di chiedergli qualche consiglio di lettura. Sabella scopre così che Fei-Fei aveva letto un sacco di libri in Cina, anche se ha difficoltà a farsi capire quando ripete in inglese i nomi degli autori. Un giorno, la gita del weekend coi genitori la porta all'Università di Princeton, dove si ritrova davanti al busto del suo eroe, Albert Einstein. Lo scienziato prediletto della bambina di Chengdu amava passeggiare in quel luogo fin dal suo arrivo negli Stati Uniti, nel 1933, all'Institute for Advanced Studies.

Decide di fare domanda a Princeton, dove, con sua grande sorpresa, viene ammessa. Mentre NVIDIA compie i suoi primi passi per migliorare l'esperienza dei videogiocatori, sempre in cerca di soldi per andare avanti, Fei-Fei studia fisica. La sua stella polare è la scienza come professione e come vocazione. Ma la scienza, ogni scienza, non può vivere nell'astratto. La vocazione scientifica sente, che lo voglia o meno, il bisbiglio della storia in cui è gettata. Fei-Fei si laurea proprio nel 1999, lo stesso anno in cui Jensen presenta al mondo la GPU prodotta da TSMC.

La signora Li non è più in grado di lavorare. Coi risparmi accumulati nei lavoretti svolti fino a quel momento, con le risorse dei familiari e con l'aiuto della comunità cinese, Fei-Fei decide di porsi nel solco della mitologia americana dell'innovazione e fondare la sua start-up: una lavanderia. C'è un problema: ha bisogno di 100.000 dollari, ma sommando risparmi e prestiti arriva solo a 80.000. Quando è pronta ad abbandonare questo progetto, Sabella le chiede di accompagnarla di persona nel tragitto da Princeton a Parsippany. A un certo punto ferma la macchina e, con la parlata incerta delle persone introverse, le dice che, dopo aver parlato con sua moglie, hanno preso una decisione: le avrebbero prestato loro gli altri 20.000 dollari per la lavanderia. Quel giorno, tornato a casa, Sabella vede il futuro che ha reso possibile. Non solo questo, ma le incessanti "visioni del paradiso" della fantascienza che abitano la sua mente.

L'intreccio tra la formazione in fisica e l'interesse per le neuroscienze e l'informatica, che caratterizza l'ultimo periodo dei suoi studi, la porta ad avvicinarsi a due importanti programmi di ricerca in altre grandi università degli Stati Uniti. Il primo si trova al MIT (Massachussets Institute of Technology), luogo fondamentale per gli studi di intelligenza artificiale, dove opera dagli anni settanta uno scienziato genovese, Tomaso Poggio, che ha affrontato in modo pionieristico la visione artificiale. Il secondo programma, al Caltech di Pasadena, storica università di Richard Feynman, è gestito da un professore di Padova, Pietro Perona. Per una curiosa ragione, entrambi i programmi sono guidati da italiani.

La sua tesi di dottorato del 2005 testimonia il percorso compiuto negli anni precedenti: il doppio obiettivo di comprendere meglio la visione umana e di costruire macchine in grado di vedere. Quando guardiamo una fotografia, vediamo un'immagine complessa composta da colori e forme. Quando un computer guarda la stessa immagine, processa i dati in forma binaria, come sequenza di bit. In una slide sul divario semantico della visione realizzata da Fei-Fei12, a sinistra c'è quello che vediamo, un gatto, a destra c'è una serie di bit, 0 e 1. I bit delle immagini sono organizzati come una griglia di quadratini colorati, i pixel, unità di informazione dell'immagine. Ogni pixel rappresenta le tessere che costituiscono il mosaico dell'immagine. Ogni pixel ha un colore specifico, determinato da una combinazione di valori di rosso, verde e blu. Questi colori, combinati in varie intensità, creano l'intera gamma di colori che vediamo. Ciò che il computer "percepisce", nella sua caverna, è la mappa di pixel colorati. Noi riconosciamo il gatto e le sue caratteristiche (baffi, occhi, pelo), mentre il computer percepisce i pixel e va addestrato a interpretare modelli e combinazioni di pixel per "vedere" il gatto. Ovviamente, i gatti che noi incontriamo non si presentano nello stesso modo, nella stessa griglia, nella stessa organizzazione, ma in diverse posizioni e variazioni, e questo accresce la complessità.

La visione artificiale, all'inizio, si basa su algoritmi progettati per compiti molto specifici, che hanno portato alcuni risultati. Per esempio, il riconoscimento ottico dei caratteri (OCR), in cui il computer viene programmato per riconoscere lettere e numeri stampati in condizioni di illuminazione stabile. Questi metodi sono stati utili per il controllo qualità di alcuni procedimenti industriali, in cui gli algoritmi aiutano a supervisionare gli standard della produzione. Anche in questi casi, il cammino verso la precisione è pieno di ostacoli. Le variazioni delle immagini che per noi sono naturali, dovute all'illuminazione e alle angolazioni, sembrano impedire ogni duttilità e flessibilità dei modelli. Non è facile uscire dalla caverna dei pixel, anche se forse "i pixel sono l'interfaccia universale". La difficoltà della visione artificiale ci fa apprezzare la meraviglia della visione umana, l'immediatezza con cui rivela la sua verità. Fei-Fei, nella sua ricerca, si concentra sui passaggi successivi: qual è il contenuto essenziale che vediamo della scena, "che cos'è esattamente che percepiamo e comprendiamo quando guardiamo il mondo"? Rispondere a questa

domanda è fondamentale per stabilire cosa debbano apprendere le macchine. I soggetti degli esperimenti, studenti e dottorandi del Caltech, guardano alcune immagini di prova e devono scrivere quello che vedono, basandosi su alcune categorie che costituiscono un ordine e una gerarchia delle immagini. Gli esperimenti sono volti a mostrare la percezione dell'essenza (gist) della scena, che nel lavoro di Fei-Fei è l'insieme delle informazioni immediatamente percepite, che vantano un "privilegio" nell'elaborazione visiva.

Nel nostro apprendimento, organizziamo sia gli oggetti che le categorie in tassonomie utili e le mettiamo in relazione col linguaggio". La riproduzione di questa capacità nelle macchine "è il puzzle più difficile ed eccitante che devono affrontare gli scienziati e gli ingegneri della visione artificiale; i modelli su cui addestrare le macchine, per risolvere il puzzle, dovrebbero contenere centinaia, migliaia di parametri. Fei-Fei Li addestra il suo modello sulla base di 101 categorie, tra cui fragole, delfini, Snoopy, lampada, fenicottero, Buddha, moto. Le immagini sono riconosciute attraverso posizioni nello spazio, che costituiscono i punti di addestramento relativi alla forma di ogni singola categoria. Per migliorare l'accuratezza del riconoscimento, vengono utilizzati modelli bayesiani, in cui l'apprendimento si definisce come l'evoluzione dei dati pregressi sulla base dei nuovi dati, secondo un modello probabilistico.

Il teorema di Bayes offre un metodo per calcolare la probabilità che un evento sia stato causato da una causa specifica, considerando diverse cause possibili. Consente di aggiornare i dati iniziali (o probabilità a priori, prior) in base a nuove informazioni (dati oggettivi recenti) per ottenere dati nuovi e migliorati (probabilità a posteriori). La tecnica di Bayes, risalente agli anni attorno al 1740 ma non pubblicata in vita, viene recuperata dal collega Richard Price e poi ripresa nel lavoro autonomo del matematico francese Pierre-Simon Laplace.

Immaginiamo un sistema di visione artificiale come un giovane detective che sta imparando a riconoscere oggetti diversi. Inizialmente, questo "detective digitale" non ha molta esperienza e le sue ipotesi su cosa sia un oggetto (come una sedia, un tavolo o un cane) sono piuttosto vaghe. Queste ipotesi iniziali sono i suoi prior nel contesto del teorema di Bayes. Ora, il sistema inizia a "vedere" o analizzare migliaia di immagini, che sono le sue "prove". Ogni immagine che il sistema esamina porta nuove informazioni, come la forma, il colore e la texture degli oggetti. Utilizzando il teorema di Bayes, il nostro detective digitale combina le sue ipotesi iniziali con queste nuove informazioni per aggiornare costantemente la sua comprensione di cosa costituisca un certo oggetto. Per esempio, all'inizio, il sistema potrebbe non essere sicuro se un oggetto con quattro gambe sia una sedia o un tavolo. Ma man mano che vede più immagini e apprende dalle loro caratteristiche (le gambe, la presenza o assenza di un piano di appoggio ecc.), il sistema aggiorna le sue credenze. Quindi, se vede un'immagine di un oggetto con quattro gambe e un piano di appoggio, aumenterà la sua convinzione che sia un tavolo piuttosto che una sedia. Attraverso questo processo iterativo di aggiornamento delle credenze basato sulle prove (le immagini), il sistema diventa sempre più esperto nel riconoscere gli oggetti. Il sistema è il detective che risolve un caso sempre più complesso.

Il 2 novembre 2005 Amazon Mechanical Turk è pronto a essere lanciato. È il servizio che l'azienda di Jeff Bezos, al tempo ancora incentrata sull'e-commerce, rende disponibile in rete per coordinare e organizzare il lavoro di intelligenze umane al fine di rispondere a un problema, in cambio di piccole somme di denaro.

L'intelligenza collettiva lanciata da Amazon viene utilizzata per l'analisi dei dati, per ricerche e sondaggi, per il miglioramento dei processi aziendali, per la moderazione dei contenuti e per molti altri compiti. In questo mercato, chiunque ha la possibilità di attingere a un lavoro temporaneo e senza vincoli, in grado di risolvere i problemi che le macchine non sanno ancora affrontare in modo intelligente, o meglio, efficace.

Bezos conia l'espressione "intelligenza artificiale artificiale" per descrivere questo procedimento, come inversione dell'attività "normale", cioè la richiesta umana al computer di eseguire un compito che la macchina sa fare meglio dell'uomo. In questo caso, c'è "un compito facile per un essere umano ma straordinariamente difficile per il computer. Quindi, invece di chiamare un servizio informatico per eseguire la funzione, si chiama un essere umano".

Il servizio nasce dall'esigenza interna di Amazon di eliminare milioni di pagine doppie nella descrizione dei prodotti per i suoi clienti: inizialmente, non si riesce a farlo attraverso algoritmi e si decide di organizzare un sito dove gli utenti potevano essere pagati qualche centesimo per ogni pagina doppia identificata.
Amazon guadagna su questa quantità di lavoro, perché prende una quota del 10% sulle operazioni completate in modo soddisfacente. Non solo: si crea un circolo virtuoso per l'azienda, in cui i "lavoratori reinvestono i loro guadagni in acquisti sulla piattaforma Amazon.

Il lavoro non ha più bisogno di alcuna intermediazione, nell'estrema riduzione dei costi di transazione, dentro una piattaforma che è controllata da qualcuno. In questa piattaforma, il lavoro è un gioco improntato a uno scopo. Deve realizzare la funzione per cui è programmato, la razionalità alla quale risponde ora la macchina umana: risolvere un problema nel più breve tempo possibile, con la minore spesa possibile. Il Novecento è in rovina, seppellito da un mare di dati, che generano altri dati: produzione di merci a mezzo di merci, cioè di dati a mezzo di dati.

Fei-Fei si rivolge alla ricerca disperata di aiuto per classificare le immagini; il futuro della visione artificiale si basa non solo su una sempre più ampia disponibilità di dati, di immagini, che aumentano ormai a un ritmo vertiginoso su Internet, dato che gli utenti caricano sempre più foto. È essenziale la loro adeguata classificazione.
Con ImageNet gli algoritmi possono essere addestrati su un numero molto più ampio di immagini, apprendendo le distinzioni, sottili ma essenziali, tra categorie simili. Mentre la costruzione di ImageNet avanza, Fei-Fei accetta un'offerta da Bill Dally, al tempo preside di informatica a Stanford.

Fei-Fei utilizza i fondi della ricerca Amazon; giugno 2009: "Quindici milioni di immagini suddivise in ventiduemila categorie distinte, selezionate da un totale di quasi un miliardo e annotate da una squadra internazionale di oltre quarantottomila collaboratori provenienti da centosessantasette paesi. Possedeva la scala e la varietà che avevamo sognato per anni, pur mantenendo un livello stabile di precisione; ciascuna singola immagine non era stata soltanto etichettata manualmente, ma anche organizzata all'interno di una gerarchia e verificata tre volte

A ottobre 2011, Sabella, sezione ironica della fanzine che continua a curare, intitolata "Raccomandazioni per migliorare la qualità della vita", si conclude con questi consigli: Sostieni il movimento per limitare drasticamente il numero di immigrati in questo paese, legali e no. Che importa se ci sono studenti pigri e autoindulgenti che non sono in grado di specializzarsi e di lavorare nelle più importanti posizioni scientifiche e tecnologiche del paese? Che importa se vietare l'ingresso negli Stati Uniti alle persone migliori e più brillanti al mondo eroderà gradualmente la nostra economia e infine indebolirà gradualmente il nostro stesso tenore di vita? Quando questo accadrà, sarai comunque morto, e chi se ne frega di un futuro in cui comunque non sarai vivo?

Jensen, che è sempre amministratore delegato di NVIDIA e continua a vivere sentendo sul collo il fiato di Don Valentine di Sequoia, a quel punto ha già dato un po' di soldi a Fei-Fei per l'iniziativa AI4All, la non profit nata dall'idea della sua studentessa Olga Russakovsky per coinvolgere nell'intelligenza artificiale le comunità meno rappresentate. Quando nel 2023 Fei-Fei pubblica la sua autobiografia, che è anche una storia dell'intelligenza artificiale, insiste sulla scienza come stella polare che ha guidato le sue scelte, impedendole per esempio di dilapidare il sogno americano nella produzione di montagne di slide per McKinsey. Suo figlio, dopo aver letto il libro, le dice: "Mamma, ho trovato la mia stella polare: sono i videogiochi".

3. Alex, o dello scrivano

Geoffrey, nato nel 1947, un certo disorientamento, assieme alle scelte eccentriche dei suoi genitori. Suo padre, l'entomologo Howard Hinton, è un convinto stalinista. Sua madre, l'insegnante Margaret Clark, è la segretaria locale del partito laburista. Da bambino lo mandano in una scuola cattolica, dove tutti parlano di Dio, mentre a casa gli dicono che credere in Dio è assurdo. All'alba della guerra fredda, dopo aver sentito che tutte le cose buone vengono da Dio, il piccolo Geoffrey, sulla base dei precetti del padre, obietta con l'insegnante che la bontà del mondo viene da qualcun altro: i sovietici.

Passa il tempo a conversare con l'amico Imnan Harvey, secondo cui il cervello funziona come un ologramma. Il contenuto dell'ologramma, le informazioni che lo caratterizzano, viene distribuito, al contrario della struttura di un'immagine. Harvey argomenta con Geoffrey che la memoria umana funziona in modo simile all'ologramma: a contare non è il contenuto specifico di un neurone, ma la distribuzione della memoria tra i neuroni, che operano nella rappresentazione di diversi concetti. Questo senso dell'informazione come connessione fa accendere una scintilla in Geoffrey che, attraverso nuovi concetti, sente di potersi liberare della natura di predestinato e scrivere una sua storia, libera dalle aspettative altrui. La sua fissazione è capire il funzionamento del cervello: una domanda semplice, con una risposta sfuggente.

Accosta con speranza e trepidazione ai corsi, che ogni volta gli promettono una spiegazione del funzionamento del cervello, ma è preso da un'irrefrenabile delusione quando nessuno è in grado di farlo veramente.
Decide di mollare tutto e fare il falegname per un anno. Se non si riesce a capire il funzionamento del cervello, tanto vale costruire mensole e credenze.

Geoffrey incontra ben presto un falegname molto più bravo di lui, e perciò, sconfortato, abbandona anche quel tentativo. Si dedica invece come assistente a un progetto di psicologia sullo sviluppo del linguaggio infantile in relazione alle classi sociali.
Il superamento dei limiti avviene attraverso le reti neurali "profonde". La cosiddetta "profondità" è legata al numero di strati di elaborazione tra input e output della rete, formati dai neuroni artificiali. Alcuni strati della rete profonda sono nascosti e, con l'aumento del loro numero, consentono un apprendimento di rappresentazioni complesse di dati; i sogni delle reti neurali divengono paper, articoli, conferenze e nuove soluzioni. Un aiuto viene dallo sviluppo della fisica e dai "vetri di spin", materiali magnetici particolari in cui i momenti magnetici (o "spin") sono disposti in maniera casuale e disordinata, a differenza dei normali magneti dove sono allineati ordinatamente.

Questa disposizione casuale rende i vetri di spin sistemi complessi, per via delle loro interazioni. Il fisico italiano Giorgio Parisi, poi vincitore del premio Boltzmann e del premio Nobel, con i suoi studi innovativi già dalla fine degli anni settanta, apre la strada a una maggiore comprensione del problema; lo psicologo David Rumelhart dell'Università di San Diego, che fin dagli anni sessanta ha riflettuto sulla modalità di analisi dell'informazione da parte dei neuroni e sul miglioramento della strada, apparentemente senza uscita, del percettrone di Frank Rosenblatt, a introdurre una svolta decisiva negli anni ottanta, con l'algoritmo di retropropagazione (back-propagation).

Nella retropropagazione gli errori vengono portati all'indietro dalla fine della rete neurale fino all'inizio, consentendo un migliore aggiustamento dei pesi dei neuroni. La ripetizione del processo, numerose volte, consente di affinare i pesi e ridurre l'errore. Geoffrey collabora a San Diego con Rumelhart e con lo psicologo James McClelland, mentre anche Jensen si trasferisce in California per impegnarsi in parallelo nel lavoro sui semiconduttori, nel fast food di Denny's e negli studi a Stanford. Nella ricerca di una migliore comprensione dei processi neurali, la scuola di San Diego propone i modelli di elaborazione distribuita parallela, adatti per gestire compiti che richiedono il mantenimento simultaneo di molteplici pezzi di informazioni, ognuno dei quali influenza e viene influenzato dagli altri.

L'elaborazione delle informazioni avvenga tramite le interazioni di un grande numero di semplici elementi di elaborazione, chiamati unità, che inviano segnali eccitatori e inibitori ad altre unità. Attraverso questa struttura, i modelli cercano di "imitare" il cervello. Il problema è che i modelli sequenziali tradizionali di questa struttura cognitiva possono essere troppo lenti e non scalabili quando devono considerare un gran numero di vincoli, specialmente se questi sono imprecisi. Di contro, i modelli di elaborazione distribuita parallela offrono alternative ai modelli seriali, permettendo un processo di elaborazione più rapido ed efficiente di fronte a vincoli aggiuntivi, anche facendo uso dell'algoritmo di retropropagazione, per esempio nella simulazione di processi motori adatta al problema.

Il falegname misura e taglia il legno secondo un progetto preciso. Allo stesso modo, Geoffrey "addestra" la rete neurale, tagliando e adattando i pesi sinaptici attraverso la retropropagazione, che ottimizza i pesi per ridurre l'errore tra l'output previsto e quello effettivo. Il falegname unisce i pezzi di legno per costruire la struttura del mobile. In modo simile, Geoffrey collega vari strati di neuroni, o layers, per costruire la struttura della rete neurale, dove ogni strato ha una funzione specifica, come riconoscere bordi o texture in un'immagine. Il falegname leviga il legno per rimuovere le imperfezioni. Geoffrey affina la rete neurale regolando i parametri, come il tasso di apprendimento, per assicurarsi che la rete non sia né troppo adattata ai dati di esempio(overfitting) né troppo generica (underfitting).
Infine, il falegname applica vernice e finiture per proteggere il manufatto e migliorarne l'aspetto. Analogamente, Geoffrey affina la rete neurale per migliorarne le prestazioni su dati mai visti prima, assicurandosi che le soluzioni trovate siano generalizzabili e non legate solo ai dati dell'addestramento. Attraverso questo processo artigianale, Geoffrey è in effetti diventato un falegname. È stato a lungo un ragazzo di bottega ribelle e anticonformista. Ora è a capo della falegnameria.

Trasforma semplici pezzi di legno (dati grezzi e architetture di base) in mobili (modelli di apprendimento automatico). Questi mobili sono utili: funzionano. E sono anche in grado di apprendere e adattarsi, nei limiti determinati dalle tecniche e dalla capacità di calcolo. L'inverno dell'intelligenza artificiale si scioglie al cospetto dell'inverno di Toronto, per due principali ragioni. La prima sta nel sostegno canadese alla ricerca. L'avventura del Canada nell'intelligenza artificiale ha radici profonde, che risalgono all'anno in cui Jensen pulisce i bagni di Oneida, il 1973, quando la Western University ospita il primo workshop della Canadian Society for Computational Studies of Intelligence, ponendo le basi per una lunga tradizione di innovazione in questo campo. Nel 1982 viene fondato il CIFAR, Canadian Institute for Advanced Research. Sulla base del riconoscimento del lavoro svolto dai ricercatori

canadesi già negli anni settanta e ottanta, in particolare in campi come la visione artificiale e l'apprendimento automatico, nel 1983 il CIFAR lancia il suo primo programma, "Artificial Intelligence, Robotics and Society".

Nel 1987, all'inizio del secondo ciclo di vita del programma, Geoffrey si trasferisce all'Università di Toronto. "Sono venuto in Canada perché mi piace la società qui e perché hanno ottime risorse per la ricerca fondamentale. Non sono molti soldi ma li danno per quegli studi animati dalla curiosità, e non solo per le grandi applicazioni." Alla sua decisione aggiunge anche un colore politico: il fastidio per il ruolo centrale degli apparati militari degli Stati Uniti (e della DARPA) nel finanziamento dei programmi di intelligenza artificiale, con un ulteriore disprezzo per le politiche di Reagan, soprattutto da parte di sua moglie, la biologa molecolare socialista Rosalind Zalin.

Qualcuno deve esserci, in un paese, per fare ricerca. Chi? Questo è il secondo, fondamentale passaggio della storia canadese dell'intelligenza artificiale: la capacità di incanalare e utilizzare le energie di coloro per cui "non esiste passato e non esiste presente, ma solo l'avvenire", per dirla con Sombart. E cioè gli immigrati. Per la sociologia del capitalismo di inizio Novecento, lo studio dello straniero accompagna l'analisi di una stagione imprenditoriale "eroica", in cui l'impresa e il denaro sono i modi per sottrarsi ai vincoli sociali preesistenti, che "hanno cessato di essere una realtà", imponendo la creazione di una nuova realtà. Alla fine del Novecento acquistano sempre più rilievo i flussi globali dell'istruzione superiore e dell'imprenditorialità.

La letteratura ha dato grande rilievo al ruolo dell'immigrazione altamente qualificata per il successo della Silicon Valley ma, nella fila che si crea per giungere negli Stati Uniti, il Canada inizia a occupare un posto considerevole anche grazie alla sua legislazione, soprattutto a partire dagli anni settanta. Gli stessi programmi del CIFAR consentono di attrarre ricercatori stranieri esperti. Tra il 2001 e il 2022 il numero di persone con un permesso di studio in Canada aumenta da meno di 150.000 a oltre 800.000, una quota della popolazione ben più consistente rispetto alla media OCSE.

Ancora prima, il Canada beneficia in modo significativo dei flussi in entrata dovuti alla dissoluzione dell'Unione Sovietica, in particolare per quanto riguarda la comunità ebraica27, con la maggiore concentrazione proprio nell'area di Toronto. Gli studenti delle comunità ebraiche russe, mentre si adattano a un nuovo contesto, in genere "eccellono nelle materie che sono meno influenzate dalle abilità linguistiche (la matematica e le scienze); ad ottobre 2004, durante la conferenza "Visualization 2004" all'hotel Hyatt Regency di Austin, Texas, quattro docenti (Aaron Lefohn, Ian Buck, John Owens e Robert Strzodka) tengono un corso di una giornata sul tema GPGPU, che sta per General Purpose Computation on Graphics Processors. I ricercatori, attraverso i loro lavori accademici, stanno cercando di sistematizzare una tendenza che a Jensen non è sfuggita: l'uso della struttura parallela della GPU per accelerare compiti computazionali intensivi, con applicazioni che vanno molto al di là del mercato iniziale dei videogiochi. La potenza di calcolo di NVIDIA sta già trasformando il modo in cui si fa ricerca.

Nei primi anni a NVIDIA, Dally aiuta a sviluppare la nuova generazione di schede grafiche. Buck è la persona decisiva per la creazione di CUDA (Compute Unified Device Architecture), la piattaforma introdotta da NVIDIA nel 2006 per consentire a programmatori e sviluppatori di sfruttare la potenza di calcolo delle GPU per applicazioni non grafiche: cominciare da Quake per andare altrove. Cosa significa? CUDA sfrutta le capacità di elaborazione parallela delle GPU consentendo ai programmatori di definire funzioni che possono essere eseguite su diversi dati contemporaneamente. Questo approccio è particolarmente efficace in applicazioni che richiedono lo stesso calcolo su grandi set di dati, come l'elaborazione di immagini. In CUDA, quando una funzione viene eseguita, viene lanciata su un grande numero di thread paralleli. Ogni thread esegue la stessa operazione su diversi elementi di dati, sfruttando così l'architettura parallela della GPU. La vicenda di Dally e Buck mostra che i sogni accademici, come la GPGPU, nell'epoca di Jensen si possono realizzare in breve tempo, senza vincoli burocratici, perché dietro non c'è l'infrastruttura del settore pubblico coi suoi bandi e i suoi finanziamenti. C'è sempre Don Valentine che sussurra: "Se perdete i miei soldi, vi ammazzo".

Il corso del 2004 individua già sei grandi aree di applicazione per la GPGPU: 1) visualizzazione e analisi dei dati, utile in campi come la fisica, la biologia e l'ingegneria, dove la capacità di elaborare rapidamente grandi quantità di informazioni è cruciale; 2) simulazioni scientifiche, tra cui modelli di dinamica dei fluidi, simulazioni atmosferiche, simulazioni di particelle e modelli molecolari; 3) elaborazione di immagini e segnali, per esempio nella diagnostica medica, per risonanze magnetiche o tomografie computerizzate; 4) ricerca biomedica, per analizzare complesse interazioni molecolari, come nello studio delle proteine, per comprendere malattie come l'Alzheimer e il cancro; 5) crittografia e sicurezza informatica, per la ricerca di vulnerabilità nei sistemi e per la protezione dei dati; 6) astronomia e astrofisica, per l'analisi delle informazioni provenienti da telescopi e sonde spaziali.

NVIDIA utilizza GeForce, la GPU dedicata ai videogiochi, per coinvolgere la comunità di videogiocatori e costruire la prima base di utenti per gli sviluppatori. La scommessa di CUDA non è gratis per NVIDIA. Agli investitori sembra un'eccentrica perdita di tempo. Jensen ne è consapevole: "I profitti hanno subito un duro colpo. Per molti anni, la nostra capitalizzazione di mercato si è aggirata appena sopra il miliardo di dollari. Abbiamo sofferto per molti anni di scarse prestazioni. I nostri azionisti erano scettici nei confronti di CUDA e preferivano che ci concentrassimo sul miglioramento della redditività. Ma siamo andati avanti".
Nel 2009, NVIDIA lancia la prima GTC (GPU Technology Conference) per mettere insieme sviluppatori, ingegneri, ricercatori e informatici. La conferenza è tradizionalmente aperta da un discorso dell'uomo col giubbotto in pelle, che fin dal 2009 parla di intelligenza artificiale, guida autonoma, medicina, oltre che di videogiochi. Col tempo, le altre applicazioni sono arrivate, come una valanga.

A dicembre 2011, la GTC Asia svela che CUDA sarà parte dei corsi di programmazione in duecento università in Cina: solo con questo passaggio, ogni anno potranno esserci ventimila programmatori in più. Andrew Ng, informatico britannico, figlio di immigrati di Hong Kong, il quale ha pubblicato con due colleghi indiani nel 200939 un importante paper sull'uso delle GPU nel deep learning e che nel 2010 ha convinto Google ad avviare una divisione sulle reti neurali, Google Brain. Secondo una storia che Dally non smette mai di raccontare, Andrew Ng sta gestendo un grande progetto per riconoscere gatti su Internet attraverso 16.000 CPU.

La fascinazione per i gatti è un grande tema dell'arte e della letteratura che unisce i popoli, ben prima dell'era digitale. Lo storico dell'arte Lionello Puppi ha ripercorso le tracce del "mistero che si associa alla presenza del gatto", ricordando i versi di T.S. Eliot. Secondo il poeta, un gatto ha ben "tre nomi differenti": il primo è il nome comune, ordinario poiché dato dagli umani; il secondo è un nome unico e speciale per lui, che non condividerebbe con altri gatti e che può indicare il legame tra i gatti e la magia; il terzo nessuno può saperlo, ma "il gatto lo conosce", è il Nome in cui risiede quella "profonda meditazione" del gatto che affascina e disorienta gli umani. Il gruppo di Andrew Ng e Catanzaro riesce a passare da 16.000 CPU a 48 GPU, migliorando la performance sui gatti. Jensen capisce che si tratta di un ambito promettente, ma manca ancora un passaggio.

Nella stanza di Alex, nel suo letto, lavorano a pieno ritmo due GPU. Secondo Geoffrey, lo studente è un eccezionale programmatore. Sta nascendo AlexNet, un modello di rete neurale convoluzionale (CNN) profonda specializzato nel riconoscimento di immagini. La rete ha 60 milioni di parametri e 650.000 "neuroni". È composta da cinque strati convoluzionali, alcuni seguiti da strati di max-pooling, e tre strati completamente connessi, con un ultimo strato softmax a mille vie. Utilizza una struttura gerarchica di strati per processare l'informazione visiva, apprendendo caratteristiche dai dati in modo progressivo, dai tratti semplici ai dettagli più complessi. Ovviamente, dal punto di vista materiale, "strati" e "neuroni" non sono oggetti fisici ma unità di calcolo, per ricevere dati (input) e per dare dati (output), secondo i risultati che si vogliono ottenere, secondo operazioni matematiche e tecniche probabilistiche.

Nel 2012 un brillante studente di nome Alex Krizhevsky, all'Università di Toronto, ha addestrato il suo computer a riconoscere e classificare automaticamente gli oggetti. Lo ha fatto elaborando un milione di immagini su due processori NVIDIA. Addestrando la sua intelligenza artificiale, il modello ha impiegato circa una settimana; senza i nostri processori ci sarebbe voluto quasi un anno. I suoi risultati erano così accurati che vinse una competizione contro ricercatori che avevano dedicato la loro carriera ai sistemi di codifica manuale. Quel momento è stato il big bang dell'intelligenza artificiale moderna. I sistemi di intelligenza artificiale richiedono calcoli enormi. Per insegnare a un computer come riconoscere con precisione veicoli, per esempio, c'è bisogno di 100 milioni di immagini di automobili, camion, autobus, veicoli di emergenza ecc.
Senza GPU, sarebbe necessario addestrare un sistema a riconoscere quelle immagini per mesi. I sistemi all'avanguardia di oggi possono ridurre questo tempo a un giorno.

Baidu fa sul serio su questi temi: nel 2014 aprirà un laboratorio nella Silicon Valley e assumerà come chief scientist Andrew Ng52, il quale ruberà a NVIDIA Bryan Catanzaro. Sono le prime avvisaglie di una competizione che si gioca sempre di più sul talento e che, con l'ingresso di grandi capitali, cambia le regole del gioco delle conferenze di ricerca. Come raccontato da Cade Metz, la conferenza al lago Tahoe si trasforma in una sfida tra Baidu, Microsoft, Google, Facebook e una start-up chiamata DeepMind per acquistare DNN Research. È un'asta vertiginosa che viene bloccata da Geoffrey a 44 milioni. Non gli interessa alzare ancora il prezzo ma fare la scelta giusta, che in quel momento secondo lui è vendere a Google, l'azienda che sembra offrire le migliori condizioni per continuare l'attività di ricerca.

L’acquisizione di Google viene perfezionata a marzo 2013, e Geoffrey dichiara sul sito dell’Università di Toronto: “Sono molto eccitato per questa fantastica opportunità di mantenere la mia ricerca qui a Toronto e, allo stesso tempo, aiutare Google ad applicare i nuovi sviluppi dell’apprendimento profondo per la costruzione di sistemi che possano aiutare le persone”. In soli due anni, il numero di aziende che collaborano con NVIDIA sull’apprendimento profondo (comprando GPU e servizi) si sono moltiplicate di 35 volte, nella salute, nell’energia, nella finanza, nell’automotive, nella manifattura. Per Jensen parlare di big bang vuol dire: finalmente più soldi, per incassare la scommessa di CUDA. Dal 2017 la GTC sarà aperta dal video I am AI, realizzato nelle immagini e nella musica dall’intelligenza artificiale, e NVIDIA scriverà “I am AI” nei suoi uffici.

AlexNet entra in un’arena dove c’è già qualcosa e dove c’è già qualcuno. Qualcosa, di cui Baidu è il segnale: la competizione tra Stati Uniti e Cina. Qualcuno: le grandi aziende tecnologiche, che hanno risorse e incentivi per fare le loro mosse, secondo il famoso motto di Facebook, nel 2012 ancora in vigore, “Move fast and break things”. L’interesse di queste aziende per l’intelligenza artificiale non può essere disgiunto dal loro modello di business, incentrato sulla crescita di una base di utenti votati a passare sempre più tempo sui social media, per caricare sempre più contenuti e fruirne. Se i social media sono “mercanti dell’attenzione”, secondo la calzante definizione di Tim Wu, allora l’attenzione stessa può essere ottimizzata, e questo procedimento ha un valore inestimabile: traduzioni molto più efficaci, suggerimenti di video e immagini per gli utenti, riconoscimento di volti, e molto altro. Attraverso simili procedimenti, è possibile avere un maggiore coinvolgimento e quindi una maggiore presa pubblicitaria; i fondatori possono anche avere un sincero interesse per l’avanzamento scientifico dell’intelligenza artificiale o le sue prospettive filosofiche ma, ammesso che ciò sia vero, si può portare avanti solo grazie ai risultati economici.

Alex passa alcuni anni a Google, prima di dimettersi e andare a lavorare nel 2018 per una piccola start-up chiamata Dessa. In quell’occasione offre una sua definizione dell’intelligenza artificiale: “L’intelligenza artificiale è una sorta di obiettivo finale dell’informatica. L’informatica riguarda l’automazione delle cose, l’intelligenza artificiale riguarda l’automazione di tutto”. Una volta hanno chiesto a Geoffrey cosa si prova a essere chiamato “padrino” dell’intelligenza artificiale, un campo di studi che tante volte ha rifiutato i suoi approcci eccentrici. Ha risposto: beh, è soddisfacente. Ha anche voluto precisare che tutta questa storia è un utile diversivo, ma la verità è che lui voleva solo capire come funziona il cervello. E non ci è riuscito.

4. Morris, il mio eroe

Sotto il cielo dell’Arizona, il 6 dicembre 2022, si agitano i leader della filiera dei semiconduttori. Lo stesso giorno, la nomenklatura del Partito Comunista Cinese si raduna nella Grande Sala del Popolo per il funerale di stato di Jiang Zemin, il leader cinese che nel 2008, dopo il suo ritiro, ha scritto un articolo su quell’industria, discutendo anche la litografia ultravioletta estrema e le GPU, “su cui intensificare i nostri sforzi in ricerca e sviluppo”.

Sotto il cielo dell’Arizona sfilano il gigante dei macchinari dei Paesi Bassi, ASML, e il gigante dei clienti, Apple, rappresentata da Tim Cook in persona. Ci sono i campioni dei macchinari degli Stati Uniti come Applied Materials, KLA, Lam Research – fondata da David Lam –, tasselli fondamentali per i controlli sulle esportazioni della guerra tecnologica tra Washington e Pechino. C’è Lisa Su, la manager nata a Taiwan, imparentata con Jensen, e soprattutto capace di risollevare AMD, dove Jensen ha iniziato la sua carriera, dalle prospettive di bancarotta.

Cronologia della guerra tecnologica USA-Cina attraverso TSMC.
A novembre, la crescita delle ambizioni cinesi nella microelettronica porta a un intervento di Morris Chang. Cosa farebbe il fondatore di TSMC se i capitali cinesi puntassero sulla sua azienda? A suo avviso, “non c’è ragione di vietare gli investimenti dalla Cina”: chi vuole una quota di TSMC dovrà pagarla molti soldi e avere il favore degli azionisti. Il governo di Taiwan, azionista di TSMC, specifica a dicembre che potrebbe non approvare tutte le acquisizioni portate avanti da Tsinghua Unigroup di quote di aziende taiwanesi dei semiconduttori, come Siliconware Precision Industries (SPIL), ChipMOS e Powertech Technology, tutte attive nel segmento dell’assemblaggio, su cui la Cina vuole consolidare la sua posizione per competere sul prezzo. Taiwan non ha solo un meccanismo per autorizzare gli investimenti esteri nel proprio territorio, ma anche gli investimenti fuori dall’isola, proprio per monitorare le attività delle aziende di Taiwan nel loro più grande mercato: la Cina. La stessa TSMC ha dovuto adeguarsi ai vincoli politici imposti dal suo governo: in Cina non è possibile portare gli investimenti più avanzati, che devono restare a Taiwan.

2016. Morris riconosce che l'investimento in Cina presenta rischi politici, ma è ancora più rischioso non esserci, perché l'accesso al mercato cinese è fondamentale per la crescita di TSMC; si stima che dal 2010 al 2015 la quota globale cinese nella progettazione dei chip sia cresciuta dal 4% a oltre il 10%.

Il 2 dicembre, mentre Donald Trump si prepara a diventare presidente degli Stati Uniti, Barack Obama alla Casa Bianca firma un ordine esecutivo epocale per bloccare la vendita di una piccola azienda tedesca di macchinari per semiconduttori, Aixtron, al fondo cinese Fujian Grand Chip Investment Fund. Un affare da 670 milioni di euro per un'azienda di circa 800 dipendenti (di cui un centinaio negli Stati Uniti), che peraltro al tempo non gode di buona salute finanziaria, contiene un messaggio ben più profondo di questi numeri. Poche settimane prima della decisione di Obama, il governo tedesco ha fornito una prima approvazione della transazione, ma ben presto sono iniziati i colloqui con l'intelligence statunitense, in un implicito coordinamento per cui i tedeschi hanno riportato la chiusura dell'affare in un limbo, in attesa del pronunciamento degli Stati Uniti.

Utenti finali dei prodotti sviluppati grazie ai suoi macchinari: tra di essi, vi sono i sistemi di difesa Patriot. Tutto ciò avviene mentre la Cina è divenuta il primo mercato di Aixtron e al crescente potere di mercato si affianca il sempre più ambizioso piano per ottenere la proprietà intellettuale.

2017. A marzo, per la prima volta il valore di mercato di TSMC supera quello del rivale americano Intel. I maestri hanno problemi davanti alla corsa degli allievi. In primavera, TSMC lancia una previsione: nei primi cinque anni del decennio successivo, la domanda di supercalcolo e di chip avanzati per i data center diventerà un fattore di crescita sempre più consistente. Ciò richiederà capacità più avanzate, che si affiancheranno al mercato di riferimento per i volumi e i margini: gli smartphone. Per TSMC, l'intelligenza artificiale non è solo un mercato ma anche un fattore interno di accelerazione per giungere all'eccellenza manifatturiera all'interno di fabbriche che, nelle loro operazioni, prevedono un'integrazione sempre più stretta tra le macchine e gli uomini.

I primi processi di completa automazione sono iniziati nel 2000, e dieci anni dopo l'applicazione di processi di smart manufacturing si è diffusa, con l'introduzione di una piattaforma informatica integrata e di analisi di big data, per rispondere all'obiettivo di ridurre la tempistica di produzione (cycle time). TSMC ha ottimizzato i processi in tutte le sue fabbriche secondo un modello distintivo. Non solo per garantire che lo stesso iter possa essere replicato in tutte le fabbriche, ma anche che una fabbrica possa apprendere dall'altra, allineando i propri parametri alle performance della fabbrica che ottiene una maggiore ottimizzazione dei processi (fab-matching): un'innovazione notevole rispetto al modello di uniformità applicato dai rivali di Intel nelle loro fabbriche (copy exactly); l'11% dei ricavi dell'azienda nel 2017 proviene dalla Cina, in crescita dal 9% del 2016.

Il 1° dicembre, all'aeroporto internazionale di Vancouver, la direttrice finanziaria e figlia del fondatore di Huawei, Meng Wanzhou, viene arrestata per via della richiesta di estradizione degli Stati Uniti. Meng è accusata di aver commesso una frode allo scopo di aggirare le sanzioni contro l'Iran.

2019. Alla fine dell'anno, HiSilicon arriva al secondo posto tra i clienti di TSMC, dietro solo ad Apple, contribuendo al 14% del fatturato. Del resto, nel 2019 Huawei venderà più smartphone di Apple (240 milioni contro 197), finendo dietro a Samsung (298 milioni). Ma quello che conta, per TSMC, accade durante l'anno. Quando a gennaio il Dipartimento della Giustizia degli Stati Uniti pubblica formalmente il provvedimento contro Meng Wanzhou, si avviano due unità di crisi in parallelo. Una a Shenzhen, dove Huawei sta già attivamente lavorando per affrontare l'emergenza più dura della sua storia, valutando l'impatto che possono avere i provvedimenti del governo di Washington su tutti i rami dell'azienda.

L'altra unità di crisi ha sede presso il Morris Chang Building a Hsinchu: i manager di TSMC si interrogano sul caso Huawei e, in quanto società quotata, devono prepararsi alle domande degli investitori. È chiaro che, in questi mesi, le due unità di crisi tengano un fitto dialogo: Huawei, per salvaguardarsi dai provvedimenti di Washington, invita i suoi fornitori e partner, compresa TSMC, a spostare quanta più produzione possibile in Cina, e nel caso di TSMC proprio a Nanchino, per operare nel sistema della Repubblica Popolare, lontano dalle catene degli Stati Uniti. Sia la guerra commerciale tra Stati Uniti e Cina che le vendite deludenti di iPhone generano problemi per TSMC, ma nella prima parte dell'anno spiccano due eventi: in primo luogo, la fonderia cinese SMIC, anche con l'obiettivo di migliorare i propri processi per essere in grado di servire HiSilicon, aumenta gli investimenti in conto capitale del 20% circa; in secondo luogo, è proprio la domanda di Huawei, con una crescita del 45% in alcune parti della supply chain elettronica nel periodo tra gennaio e marzo 2019 rispetto all'anno precedente, a garantire la tenuta dell'elettronica di Taiwan e della stessa TSMC. Secondo gli analisti, a trascinare questi risultati sono il patriottismo dei consumatori cinesi e l'alta qualità ormai raggiunta dai prodotti di Huawei.

21 maggio, Huawei viene inserita nella Entity List del Bureau of Industry and Security. La decisione è accompagnata da parole destinate a diventare sempre più familiari tra i legali delle imprese tecnologiche (e ad arricchirli): “Il governo degli Stati Uniti ha stabilito che c’è ragione di credere che Huawei sia coinvolta in attività contrarie alla sicurezza nazionale o agli interessi di politica estera degli Stati Uniti.

Davanti all’azienda c’è “la più tragica ed eroica Lunga marcia nella storia della scienza e della tecnologia”: il disegno di una supply chain autosufficiente, capace di resistere allo schiaffo di Washington. È un memo leggendario, che viene esaltato dal fondatore Ren Zhengfei e dalla propaganda cinese.

2020. Nel secondo trimestre dell’anno, Huawei arriva in cima al podio: per la prima volta, in un mondo che combatte contro il coronavirus, supera Samsung negli smartphone. La sua riduzione delle vendite è minore di quella del gigante coreano. Washington ha deciso di fare leva sui numerosi segmenti dell’industria in cui ha una posizione privilegiata, dagli strumenti per la progettazione, con aziende come Synopsys e Cadence Systems, ai macchinari, con Applied Materials, KLA-Tencor, Lam Research, ai materiali e alla chimica, con Dow DuPont, 3M e Corning. Rispetto a questi segmenti, la Cina non ha ancora alternative praticabili, per non parlare di società in grado di saper fare lontanamente quello che ASML o TSMC sanno fare: il tempo scorre inesorabile, il sogno di Teresa He Tingbo è lontano.

TSMC affronta una pressione sempre più forte per stabilirsi negli Stati Uniti. Dapprima, il Pentagono aveva suggerito ai clienti di TSMC coinvolti in attrezzature a uso militare di utilizzare fabbriche presenti negli Stati Uniti, come quelle di Global Foundries, ma senza alcun risultato: il divario con TSMC era troppo alto. I clienti negli Stati Uniti, tuttavia, sono i più importanti per TSMC, responsabili di circa il 60% degli ordini. Questa realtà, insieme alla pressione di Washington, porta il 15 maggio all’annuncio di una fabbrica avanzata in Arizona, con investimenti stimati a 12 miliardi di dollari.

2021. Il 24 settembre, dopo un accordo raggiunto col Dipartimento della Giustizia degli Stati Uniti, Meng Wanzhou lascia il Canada e ritorna in Cina, accolta come una principessa, un’eroina nazionale. Il fatturato di Huawei scende del 28,5%, collocandosi sotto i 100 miliardi di dollari. I profitti netti crescono invece del 75,9%, soprattutto per effetto della vendita della divisione smartphone Honor a una serie di investitori governativi, stimata a circa 15 miliardi di dollari.
Gli investimenti in ricerca e sviluppo non si fermano, e ammontano al 22,4% del fatturato.

2023. Discorso dopo discorso, prende forma la dottrina di Morris, sovente accompagnata dalla citazione di un altro libro di grande successo: il volume di Graham Allison del 2017 sulla trappola di Tucidide che avvinghia Stati Uniti e Cina.

Morris osserva che gli Stati Uniti hanno lanciato una politica industriale sui semiconduttori per ridurre i progressi della Cina. È una politica che lui appoggia, ma allo stesso tempo ricorda che il concetto statunitense di friendshoring, ovvero la riorganizzazione della filiera industriale sulla base di amicizie mai chiaramente specificate e inimicizie chiare (la Cina), non include Taiwan. Anzi, Morris sottolinea che la segretaria al Commercio Gina Raimondo non perde occasione di rimarcare il rischio di Taiwan, il pericolo di Taiwan.

Ora va considerato un primato della sicurezza nazionale da cui derivare l’auspicata nuova definizione: “Permettere alle proprie aziende di fare profitti all’estero e consentire l’ingresso di prodotti e servizi stranieri nel proprio paese, senza danneggiare la propria sicurezza nazionale e i propri vantaggi competitivi tecnologici ed economici”.

Quando NVIDIA è ancora ai primi passi, Jensen cerca più volte e senza successo di contattare l’ufficio vendite negli Stati Uniti di TSMC, che in quel periodo mostra già un impressionante ritmo di crescita, capace di rivoluzionare l’industria con l’idea decisiva della separazione tra la produzione, compito perfetto per l’Asia, e la progettazione, perseguita da un ecosistema fabless sempre più ampio di aziende che vogliono servire nuovi mercati, come quello dei videogiochi, senza sobbarcarsi i costi della costruzione e della gestione delle fabbriche.

A gennaio 1998, quando NVIDIA consegna il milionesimo RIVA 128, Jensen è riuscito finalmente a parlare con Morris. La leggenda vuole che, dopo essere stato ignorato dagli uffici commerciali di TSMC negli Stati Uniti, Jensen abbia scritto direttamente a Morris per discutere urgentemente di affari. Secondo il racconto del fondatore di TSMC, a un certo punto gli ha effettivamente telefonato ma non si riesce a sentire nulla, per il chiasso negli uffici di NVIDIA.

Tutti sono radunati ad ascoltare la telefonata col manager, già un mito per ogni ingegnere elettronico. Jensen deve placare gli animi dei suoi dipendenti, e così inizia un'eccezionale relazione commerciale.

Morris spesso si reca in persona dai suoi clienti con un quadernetto nero, dove annota i mercati su cui vogliono investire, le loro prospettive, le loro richieste di wafer, per aggiustare la capacità produttiva di TSMC. La sua gioia più grande e il suo interesse personale sono sempre stati il successo e la crescita dei clienti, che sono il suo successo e la sua crescita.

Nel 2001, NVIDIA diviene l'azienda di semiconduttori ad avere raggiunto più in fretta un miliardo di dollari di ricavi, mentre le sue GPU prodotte da TSMC, con la serie GeForce, continuano ad accumulare riconoscimenti. La piattaforma CUDA si basa sull'attrazione di sviluppatori, che scrivono applicazioni per dimostrare i benefici delle GPU. Il processo è partito dalla base di giocatori delle GeForce, ed è stato costoso e complicato, colpendo la capitalizzazione di mercato di NVIDIA. Gli azionisti sono rimasti scettici su questa scommessa di lungo termine. Nel 2012 è arrivato lo spartiacque di AlexNet. CUDA viene adottato da ricercatori e scienziati, rendendo possibile quella che Jensen, sempre intento ad avvolgere le innovazioni nell'indispensabile patina del marketing, chiama "democratizzazione del supercalcolo".
Gli sviluppatori di CUDA, crescendo sempre di più come numero e come ambiti di applicazione, alimentano la "pila" (stack), ovvero i componenti software e hardware che costituiscono la piattaforma di NVIDIA: GPU progettate per elaborare enormi quantità di calcoli, particolarmente utili per il rendering grafico e il calcolo parallelo; librerie di accelerazione che gli sviluppatori utilizzano per migliorare le prestazioni, specialmente in compiti di calcolo intensivi; sistemi di hardware e software, tra cui il sistema operativo e l'infrastruttura necessaria per far funzionare le GPU; programmi software utilizzati dagli utenti finali e ottimizzati per sfruttare la potenza di calcolo delle GPU attraverso CUDA.

Ogni cosa può rientrare nell'accelerazione di NVIDIA. E succede. Secondo i dati ufficiali del 2023, ormai quattro milioni di sviluppatori lavorano a CUDA: NVIDIA è arrivata a due milioni in dodici anni, ma il numero è raddoppiato nei due anni e mezzo successivi. CUDA è stato scaricato più di 40 milioni di volte.

2014, NVIDIA produce ricavi di poco superiori ai 4 miliardi di dollari, mentre TSMC supera i 25 miliardi di dollari, e Intel si colloca oltre i 55 miliardi. Nel 2023 Intel realizza 54 miliardi di ricavi, ma TSMC ha superato il gigante americano con 69 miliardi. Anche NVIDIA è balzata a 60 miliardi ed è considerata il secondo cliente di TSMC dopo Apple.
Nel 2014, mentre Jensen celebra il suo eroe, NVIDIA rivendica anche il considerevole aumento dei ricavi nel segmento supercomputer e data center, pssato da poco più di 20 milioni nel 2010 a quasi 200 milioni in pochi anni. Dieci anni dopo, i ricavi annualizzati del segmento data center per NVIDIA si avvicinano ormai a 100 miliardi

5. Le fabbriche dell'orsetto Kumamon

Com'è fatta una fabbrica di TSMC? Le immagini tradizionali dei loro interni, con gli uomini che si alternano ai macchinari, sono state già sostituite da ambienti dove i tecnici si fanno strada per supervisionare le macchine in mezzo a robot che compiono velocemente i compiti richiesti.

"I tecnici lavorano tutti fuori dalla fabbrica. I corridoi all'interno della fabbrica sono vuoti, e i chip sono trasportati in automatico". Il cervello della fabbrica non sono gli operai bardati coi guanti, ma è l'Mcc (Manufacturing control center), dove gli ingegneri stanno seduti davanti a una fila di computer, con ampie finestre affacciate sul parco scientifico di Taichung. Fissano lo schermo e analizzano i dati che provengono dalle operazioni della fabbrica, per valutare che tutto vada secondo i piani. Perché nulla può scostarsi da quanto previsto, non possono esserci sbavature all'interno di un processo manifatturiero fatto di oltre mille passaggi.

In un circolo, gli ingegneri informatici e gli esperti di intelligenza artificiale di TSMC utilizzano, coi loro computer, la capacità di calcolo che TSMC stessa è in grado di offrire per produrne ancora di più nei progetti di aziende come NVIDIA, che per sua stessa natura si basa sull'automazione. In un percorso costante di precisione e ottimizzazione. Eppure, quelle fabbriche si trovano in luoghi fisici. Non c'è ottimizzazione senza energia e senza acqua. Tutto ciò porta alcune variabili nel cammino degli uomini e delle macchine in un mondo che, per parafrasare Lenin, ha bisogno "dei chip e dell'elettrificazione".

Si stima che nel 2025 a TSMC sarà dovuto l'8% dell'intero consumo di elettricità di Taiwan. C'è quindi una corsa con cui l'azienda di proprietà statale di Taiwan, Taipower, cerca di stare dietro alla continua espansione di TSMC, alla

sua sete di energia determinata dal mercato, che non può essere soddisfatta da fonti instabili. Taiwan si trova ad affrontare una pesante crisi idrica, esacerbata dagli eventi climatici estremi e dalle ondate di siccità, proprio in alcune aree come Kaohsiung, dove si sviluppano gli ecosistemi industriali di TSMC e degli altri campioni della tecnologia dell'isola, tra cui ASE, uno dei leader di test e packaging di semiconduttori, che a sua volta, davanti alle difficoltà nel reperire forza lavoro qualificata, punta sempre di più sull'automazione.

La sete di energia e di acqua delle industrie che rendono possibile la vita digitale genera così tensioni e conflitti con l'opinione pubblica che la stessa democrazia di Taiwan affronta con difficoltà.

Nessun sindacato: questo è il segreto del successo che Morris rivendica orgogliosamente in un'intervista del 2016, in cui già prevede un rilievo sempre maggiore del supercalcolo e dell'intelligenza artificiale per il futuro economico di TSMC. Anche in quell'intervista, Morris si concede una citazione letteraria, riprendendo il famoso titolo del libro del segretario di stato Dean Acheson: lui era "presente alla Creazione" dell'industria statunitense dei semiconduttori. Che cosa ha imparato in quel momento? Lui e gli altri manager dei semiconduttori consideravano l'industria automobilistica l'esempio negativo da non seguire.

"Avevamo già visto chiaramente che le dispute tra lavoro e management avevano fatto crollare l'industria automobilistica americana. Perciò, il settore dell'alta tecnologia in America decise di evitare i sindacati." È un tema che appassiona chiaramente Morris: "Google, Amazon, Facebook, Microsoft, nessuna di loro ha i sindacati, nemmeno Intel, nemmeno Texas Instruments". Il giornalista non può arginarlo, parla a ruota libera: "I presidenti e gli amministratori delegati delle aziende ad alta tecnologia sarebbero totalmente d'accordo su questo: una delle chiavi del loro successo è che non hanno sindacati. Perché un'azienda abbia successo, tutti devono lavorare insieme.

Le tensioni tra lavoratori e manager sono pessime. Magari portano benefici nel breve termine ai lavoratori, come salari un po' più alti o un po' meno ore di lavoro, ma nel lungo periodo sono un male per i lavoratori e per la società nel complesso".
Dicembre 2022, prende la parola il presidente Biden. Morris lo ascolta con attenzione. "Come vediamo qui a Phoenix, gli Stati Uniti sono una delle principali destinazioni per le aziende di tutto il mondo che vogliono fare investimenti, perché disponiamo di una forza lavoro di livello mondiale, altamente qualificata e impegnata: il lavoro sindacalizzato. Più di tremila lavoratori sindacalizzati, i più qualificati e i migliori al mondo, stanno contribuendo a costruire questa favola. Anche il secondo stabilimento sarà costruito con la manodopera iscritta ai sindacati. E stiamo lavorando con aziende, community college, scuole tecniche, università, per programmi di apprendistato e di formazione guidati dai sindacati. […] La classe media ha costruito il nostro paese e i sindacati hanno costruito la classe media."

Se la macchina si guasta all'una di notte, negli Stati Uniti verrà riparata il mattino successivo, ma a Taiwan verrà riparata alle due del mattino. Se a Taiwan un ingegnere riceve una chiamata mentre dorme, si sveglierà e inizierà a vestirsi. Sua moglie gli chiederà: 'Qual è il problema?'. Lui risponderà: 'Devo andare in fabbrica'. La moglie tornerà a dormire senza fiatare. E questa è la cultura del lavoro". Il nuovo modo di produzione asiatico, l'infrastruttura manifatturiera del mondo, che piaccia o meno è costituito anche da questi aneddoti e stereotipi. Stereotipi che per secoli gli occidentali e le potenze coloniali hanno indirizzato verso gli altri e che ora si ritorcono contro loro stessi.

Morris è l'élite che va a dire alla sua fabbrica delle élite che loro non hanno più le chiavi della civiltà delle macchine, proprio perché sono élite. La fame globale di efficienza richiede una vita efficiente per l'azienda. Ormai non c'è più alcun conflitto tra capitale e lavoro. Il conflitto non è nemmeno pensabile, se non come un errore del sistema operativo, che viene corretto dagli strumenti di fab-matching senza che nessuno debba muovere un dito. La rivoluzione è la produzione stessa, mentre i rapporti di forza non possono più essere rivoluzionati.

Morris procede "con l'arma dei bassi salari", eppure, allo stesso tempo, modera la fame della rendita: "Quello che il governo non deve mai fare è abbassare le tasse. I tagli alle tasse sono sbagliati". Nella "repubblica di Samsung", la Corea del Sud, i proprietari del grande avversario di TSMC corrompono il presidente dello stato, finiscono in prigione, escono dalla prigione, pagano una gigantesca tassa di successione alla morte del patriarca, e poi la fiscalità generale è nuovamente convertita in crediti fiscali annunciati dal nuovo presidente dello stato nell'Università di Sungkyunkwan, che è gestita da Samsung. I crediti fiscali, nella zona più popolata della Corea, il Gyeonggi, garantiranno gli investimenti di Samsung (che fa la parte del leone, con più della metà), SK Hynix e una pletora di imprese minori.

Gli investimenti sono 471 miliardi di dollari fino al 2047 e tre milioni di posti di lavoro, per garantire a Seul un 10% della filiera complessiva globale dei semiconduttori e aumentare la propria quota su alcuni materiali fondamentali dal 30% al 50%. Certo, c'è la democrazia in Corea! La gente va a votare. Devono pur divertirsi, e infatti si divertono: se guardate le immagini della sera elettorale in Corea, i politici sono rappresentati come cartoni animati divertentissimi, con una grafica tridimensionale impagabile, nelle numerose variazioni della creatività e vivacità artistica di quel popolo, con la continua espansione globale della sua musica pop. Poi comanda Samsung, anche oltre il 20% del PIL coreano che da essa dipende. Ma non è che comanda come un politico che si candida al Parlamento europeo e dice "voglio comandare, ora cambia tutto". No, nel caso di Samsung il comando è una realtà effettiva: pagare la tassa di successione, prendere il credito fiscale, costruire. Oggi, domani e dopodomani.

"What I cannot create, I do not understand."
È la frase che Richard Feynman lascia sulla sua lavagna nera prima di morire. È un dilemma di quello che Max Weber chiamava Geistige Arbeit, lavoro dello spirito: come posso agire su ciò che non so fare? Come posso fare politica senza fare scienza? E d'altra parte: come può essere scienza quello che non so ricreare, che non so costruire in modo sistematico?

La politica, persino la politica che ha un potere effettivo (gli Stati Uniti) non ha più gli strumenti per capire e così vede quello che accade con approssimazione: la tecnica si è separata dal suo percorso, e nella sua rincorsa deve andare per strappi. Come lo strappo della sicurezza nazionale.

Di tecnologia, cosa ne capiscono a Washington? Si sono fissati di mettere qualche soldo pubblico per costruire fabbriche di semiconduttori, pensando di ricavarne un dividendo politico, ma non hanno chiaro nemmeno come funzioni quella filiera, né hanno mai pagato un dividendo agli azionisti. E l'ultima volta che si sono intrufolati nel meccanismo del mercato negli anni ottanta, ricorda Jensen, quando hanno detto che lo facevano per salvare la litografia degli Stati Uniti, è andata a finire che le imprese dell'America, le nostre imprese, sono fallite e sono rimasti in piedi solo gli olandesi.

Tra i documenti che Jensen riguarda, ci sono quelli relativi alla possibilità di vendere ai clienti cinesi la GPU A800, in sostituzione della A100 che il governo degli Stati Uniti non vuole che si venda. Diremo che la A800 va incontro alle richieste del governo degli Stati Uniti e che non può essere programmata per superare i limiti che il governo considera preoccupanti, pensa Jensen. Ma tanto, secondo lui, non è finita qui. È un circolo che si rafforza: l'incentivo per la Cina di provare a fare quello che noi sappiamo fare è troppo grande, e la prepotenza della sicurezza nazionale rovina le dinamiche di mercato, facendo danni a tutti e rischiando di rallentare NVIDIA, che ha l'accelerazione nel suo DNA. Allo stesso tempo, Jensen lo ammette, non possiamo fare finta di vivere in un mondo senza vincoli politici. Possiamo costruire altri mondi virtuali, ma questo è il mondo in cui operiamo e costruiamo.

 # Parte seconda. L'intelligenza del carciofo
1. Cynar

Il 22 ottobre 1973 è la giornata inaugurale dell'Istituto della Fondazione Dalle Molle per gli studi semantici e cognitivi, nella Villa Heleneum di Lugano-Castagnola. È il primo atto dell'investimento di Dalle Molle, per indagare "le strutture profonde comuni" a tutte le lingue. Dalle Molle è un pioniere dell'idea di informatica incentrata sull'uomo: "Il progresso scientifico in generale ed il progresso nell'informatica in particolare non dovrebbero asservire l'uomo ma al contrario essere al suo servizio". Il primo Istituto, poi spostato a Ginevra nel 1976, è solo l'inizio dell'attività pionieristica di Dalle Molle, che nel 1987 avvia a Lugano l'IDSIA, Istituto Dalle Molle di Studi sull'Intelligenza Artificiale.

Michie non poteva pensare che, proprio in quello stesso periodo, l'inventore del Cynar l'avrebbe contattato per dirgli che, oltre al mini-istituto dedicato al carciofo a Polignano a Mare, aveva intenzione di finanziare un istituto sull'intelligenza artificiale. Dalle Molle si reca a chiacchierare a Lugano coi ricercatori, vestito coi suoi abiti di seta, con il papillon fatto a mano. Le prime attività sono legate all'interpretazione del linguaggio naturale e alla traduzione automatica, ma dopo pochi anni, quando Dalle Molle è ancora in vita, l'istituto si sposta sulla ricerca legata al machine learning.

Nel 1997, l'IDSIA è tra le principali istituzioni pubbliche e private al mondo sull'intelligenza artificiale secondo "Business Week": in mezzo a Stanford, Microsoft Research, MIT, Carnegie Mellon, Bell Labs, spunta IDSIA. Luigi Einaudi, "apostolo della libertà", ha scritto in una lettera che Dalle Molle conserva con cura: "Mi compiaccio che un imprenditore versato in una di quelle industrie alimentari, le quali hanno tutto da perdere da protezionismi e

vincolismi, si associ a coloro i quali colla penna e coll'opera difendono la causa della libertà economica contro le 'signorie' d'ogni specie, statali e private. Ed auguro a lei quel successo che a noi non è arriso".

L'informatico tedesco Jürgen Schmidhuber, dopo aver studiato a Monaco e aver rifiutato un'offerta di post-doc al Caltech, si trasferisce a Lugano per assumere nel 1995 la direzione dell'istituto fondato da Angelo Dalle Molle. La sua biografia accademica ufficiale inizia con la frase: "Fin da circa quindici anni di età, l'obiettivo principale del professor Jürgen Schmidhuber è stato costruire un'intelligenza artificiale in grado di migliorarsi da sola, che fosse più intelligente di lui, e poi andare in pensione".

Schmidhuber ha avuto un ruolo significativo non solo per l'affermazione dell'IDSIA, ma anche per alcune ricerche specifiche. Soprattutto quelle legate alle reti LSTM (Long Short-Term Memory). Si tratta di un tipo avanzato di rete neurale progettato per tenere traccia delle informazioni, in modo da lavorare meglio con dati in sequenza, come le rappresentazioni del linguaggio o le serie temporali. La struttura di queste reti consente a esse di mantenere o scartare le informazioni. Le reti LSTM sono diventate strumenti fondamentali per numerose applicazioni, tra cui la traduzione, il riconoscimento vocale, la generazione di testo, l'analisi previsionale.

Gli studi di Schmidhuber hanno introdotto le reti LSTM, e questo non solo ha avuto uno sviluppo commerciale, ma è stato riconosciuto dal "campo avverso". Nel 2021 NVIDIA inserisce Schmidhuber tra i "Da Vinci della nostra epoca" presenti alla sua GTC, insieme alla trinità del premio Turing. Quando vuoi attaccare mercati da centinaia, migliaia di miliardi di dollari, quando vuoi accelerare la crescita dei tuoi ricavi, non è un problema inserire qualche Da Vinci in più nelle tue slide.

Al culmine di quest'edificio, c'è colui che Schmidhuber definisce il fondatore dell'informatica: Gottfried Wilhelm von Leibniz. Il pensatore nato a Lipsia il 1° luglio 1646 è celebrato come genio universale. La sua universalità non sta solo in una vorace curiosità che rifugge i confini disciplinari, figlia sia di un concetto di filosofia che non ha un'astratta separazione con la scienza, sia di un interesse per gli aspetti meccanici e ingegneristici dei saperi e delle tecniche. La potenza del suo pensiero sta anche nella continua capacità di varcare i confini del tempo e dello spazio, di eroderli e riscriverli.

Leibniz ha progettato la prima macchina in grado di eseguire le quattro principali operazioni aritmetiche (1673) e la prima con una memoria interna. Leibniz si accosta alle arti della memoria all'interno del suo progetto del calcolo universale, che si specchia nell'ambizione di una clavis universalis, "quel metodo o quella scienza generalissima" con cui è possibile "decifrare l'alfabeto del mondo; riuscire a leggere, nel gran libro della natura, i segni impressi dalla mente divina; scoprire la piena corrispondenza tra le forme originarie e la catena delle umane ragioni; costruire una lingua perfetta capace di eliminare gli equivoci e di svelare le essenze mettendo l'uomo a contatto non con i segni, ma con le cose.

La scienza del "vero" è la capacità di determinare con chiarezza e precisione la materia del contendere, di qualunque contesa, affinché ne sia possibile il calcolo. Non si potrà facilmente terminare le controversie né imporre il silenzio ai settari, se non ricondurremo i ragionamenti complicati a calcoli semplici, i vocaboli di vago e incerto significato a caratteri determinati. Bisogna fare cioè in modo che ogni paralogismo non sia altro che un errore di calcolo e che un sofisma, espresso in questo nuovo tipo di scrittura, non sia altro in realtà che un solecismo o un barbarismo, facilmente refutabile in base alle stesse leggi di questa grammatica filosofica.

Ciò fatto, quando nasceranno controversie, non vi sarà bisogno di disputare tra due filosofi più che tra due contabili: basterà infatti prendere in mano la penna, sedersi davanti all'abaco e (preso con sé, volendo, un amico) dirsi a vicenda: calcoliamo. È il celebre "calculemus": il trionfo di una ragione determinata, basata sulla possibilità di tradurre i problemi in schemi, in unità di informazione che possono essere calcolate e riprodotte. Un passaggio che ha una grande potenza non solo filosofica, ma anche politica.

Il calcolo è la liberazione fondamentale dell'uomo, non solo dalla "fatica di cercare nuovamente quel che è già stato trovato", ma anche dalle pene che lo assalgono, dalle fatiche, dai drammi che infligge a sé e ai suoi simili. Come Leibniz stesso scrive in uno di quegli incredibili manoscritti rispuntati secoli dopo a Hannover e sconcertanti per la loro attualità: "Indignum enim est excellentium virorum horas servili calculandi labore perire". È indegno per gli uomini eccellenti perdere tempo nella fatica servile del calcolare.

La macchina può svolgere quelle funzioni. Creiamo macchine che sappiano calcolare, perché siano strumento della nostra liberazione, per dare tempo alla nostra vita creativa, alla vita della mente. Dall'altro lato, l'uomo si libera

attraverso il calcolo stesso. Ma per ottenere la liberazione ultima, occorre la piena padronanza dell'arte caratteristica, la sua piena potenza scientifica. Leibniz, tuttavia, limita le pretese effettive dell'arte, o, per essere più precisi, le lega in modo decisivo alla disponibilità di dati e alla loro trattazione da parte dell'ingegno.

Ciò consente che l'arte divenga propriamente una scienza, la scienza universale, che consentirà di sapere veramente, cioè di sapere tutto: Affinché però nessuno pensi che io vanto o auspico cose impossibili, bisogna sapere che con quest'arte si possono ottenere, con un'adeguata applicazione, soltanto le cose che, con tutto l'ingegno, si possono ricavare dai dati, ossia che sono determinate dai dati proprio come nei problemi della geometria. Che invero sono di fatto e dipendono dalla sorte o dal caso, nella misura in cui è manifesto che non sono pertinenti all'arte della scoperta. Non disponiamo in modo "automatico" di una clavis universalis. La chiave non è universale perché non può applicarsi a ogni cosa, ma solo a quello che si può "ricavare dai dati".

1973, mentre Jensen Huang pulisce i bagni di Oneida, in Nuova Zelanda nasce Shane Legg. Il piccolo neozelandese riceve un regalo per il suo decimo compleanno: un personal computer. Si tratta del Dick Smith VZ200 prodotto da VTech, un'azienda di Hong Kong, col processore Z80A della Zilog fondata dall'italiano Federico Faggin. Il VZ200, conosciuto altrove come Laser 200, acquista una certa popolarità in Australia e Nuova Zelanda, nonostante la sua improbabile tastiera in gomma. Shane lo usa per giocare a Space Invaders, come tutti, ma non solo. Inizia a programmare.

Nell'anno decisivo, il 1999, mentre i videogiocatori sparano su Quake in attesa del Millennium bug, Raymond Kurzweil rivela al mondo le sue previsioni. Nato nel 1948 a New York, nel Queens, mentre Jensen pulisce i bagni di Oneida Kurzweil ha già finito di studiare al MIT con Marvin Minsky, che ha conosciuto quando aveva quattordici anni. Kurzweil non è un accademico, ma un inventore-imprenditore. Negli anni settanta e ottanta fonda aziende pionieristiche, tra cui Kurzweil Computer Products, specializzata in tecnologie di riconoscimento ottico dei caratteri, e Kurzweil Music Systems, un'impresa rivoluzionaria nel campo degli strumenti musicali elettronici.

Nell'inverno dell'intelligenza artificiale, Kurzweil formula le sue previsioni sull'intersezione tra le capacità cognitive umane e la potenza di calcolo, e raggiunge una certa notorietà grazie al suo libro del 1990, The Age of Intelligence Machines. La consacrazione come futurista avviene nel fatale 1999, con la pubblicazione di The Age of Spiritual Machines e le sue previsioni sul 2029. Secondo Kurzweil, in quell'anno la crescita esponenziale della capacità di calcolo consentirà la creazione di intelligenze artificiali con abilità paragonabili a quelle del cervello umano.

Kurzweil rende popolare la legge dei ritorni acceleranti: sostiene che il progresso tecnologico, specialmente in informatica, avanzi in modo esponenziale e non lineare. Entro il 2029 avremo computer con potenze di calcolo simili a quelle del cervello umano, capaci di eseguire compiti che richiedono comprensione emotiva e cognitiva. Tra le frasi raccolte da Kurzweil, c'è una citazione enigmatica di John von Neumann del 1949: "Sembrerebbe che abbiamo raggiunto i limiti di quello che è possibile fare con la tecnologia del computer, sebbene uno debba fare attenzione con queste affermazioni, visto che tendono a suonare piuttosto stupide nel giro di cinque anni".

Shane è tornato all'accademia per cominciare il dottorato. Il ragazzo della Nuova Zelanda è arrivato a Lugano, all'Istituto Dalle Molle per l'Intelligenza Artificiale, fondato coi soldi del Cynar. Una parola dopo l'altra, una foglia di carciofo dopo l'altra, fino a giungere al suo cuore. All'Istituto Dalle Molle, Shane Legg lavora con Schmidhuber, e con l'informatico tedesco Marcus Hutter. La solida formazione matematica di Shane lo porta a un confronto con uno degli eroi della disciplina, il compagno di passeggiate di Einstein a Princeton: Kurt Gödel.

1959 Minds, Machines and Gödel, pubblicato nel 1961, che inizia con la frase: "Mi sembra che il teorema di Gödel provi che il meccanicismo è falso, e cioè che le menti non possono essere spiegate come macchine". In estrema sintesi, Lucas usa il teorema dell'incompletezza di Gödel per dimostrare che la mente umana non è una macchina di Turing, cioè un computer. Secondo il teorema di Gödel, in ogni sistema formale sufficientemente potente per includere una porzione rilevante della matematica, esistono proposizioni che, pur essendo vere, non possono essere dimostrate all'interno del sistema stesso. Gödel dimostra l'esistenza di affermazioni matematiche che sono intrinsecamente "indimostrabili". Ogni sistema formale possiede limiti intrinseci insuperabili all'interno del sistema stesso. Lucas applica questo concetto all'intelligenza artificiale: una macchina, che opera secondo un insieme fisso di regole algoritmiche, avrà sempre limiti intrinseci e non potrà mai eguagliare la flessibilità e la capacità di comprendere paradossi e contraddizioni propri dell'intelligenza umana, nonché alcune sue caratteristiche specifiche, come la creatività.

Nel volume del 2006 che introduce a un pubblico vasto l'intelligenza artificiale generale, Schmidhuber propone, sulla base di varie ricerche precedenti, il concetto di "macchina di Gödel" per sostituire la più comune "macchina di Turing". Su questa base, e sugli studi di Marcus Hutter, procede a smontare l'articolo di David Chalmers sulla filosofia della singolarità. Le macchine di Gödel, invece di seguire un algoritmo fisso e predefinito, possono teoricamente riscrivere e migliorare i propri algoritmi. Le macchine che apprendono sono, portate all'estremo, macchine che auto-apprendono. Non solo riescono a eseguire compiti o risolvere problemi in base a un set di istruzioni predefinite, ma anche a cambiare quelle istruzioni se scoprono che esiste un modo migliore per raggiungere i loro obiettivi: la massimizzazione di una specifica funzione di utilità.
Per quanto riguarda la classica obiezione sulla creatività, Schmidhuber sostiene la possibilità di matematizzare "la scienza e l'arte" con l'uso delle "funzioni di valore" che stanno dietro la creatività. È inutile che i filosofi tirino Gödel per la giacchetta. Non c'è un resto. Non c'è qualcosa di ulteriore.

Un altro modello teorico importante per il ragionamento di Shane sull'intelligenza artificiale generale è AIXI63, frutto dell'elaborazione di Marcus Hutter, che sarà suo professore e relatore della sua tesi all'Istituto Dalle Molle. Shane scrive: "L'AIXI sta all'intelligenza artificiale generale reale come le macchine di Turing stanno ai computer reali. Lo scopo di queste costruzioni teoriche è agire come modelli matematici sintetici che possono essere studiati e investigati teoricamente.

Fin dagli inizi delle sue attività all'Istituto Dalle Molle con Hutter, Shane entra a pieno titolo in questi dibattiti. Scrive diversi articoli e contributi scientifici sul rapporto tra il teorema dell'incompletezza di Gödel e gli obiettivi dell'intelligenza artificiale. Le sue ricerche culminano nella tesi di dottorato del 2008, "Machine Super Intelligence", che si apre con la domanda: "Che cos'è l'intelligenza?".

Shane propone una definizione di intelligenza che misura la capacità di un agente di raggiungere obiettivi in una vasta gamma di contesti. Il rapporto fra i tre elementi (agente, obiettivo, ambiente) può essere il tema unificante da applicare sia a entità biologiche che artificiali, e su cui misurare la generalità (l'applicazione, da parte dell'agente, degli obiettivi a una pluralità di ambienti). La discussione si estende poi alle questioni etiche e di sicurezza legate allo sviluppo di macchine superintelligenti, sottolineando la necessità di sviluppare un'intelligenza artificiale amichevole per garantire la sicurezza umana.

Il 7 dicembre 2009 scrive: "La mia previsione per gli ultimi dieci anni è stata per una intelligenza artificiale generale di livello più o meno umano nel 2025 (nonostante preveda anche che gli scettici negheranno che ciò accadrà quando accadrà!). Quest'anno ho provato a ideare qualcosa di più preciso. Così ho scoperto che mentre la mia moda è intorno al 2025, il mio valore atteso è in realtà un po' più alto, 2028". A fine anno, Shane aggiunge una previsione operativa per quello che succederà nel decennio successivo, gli anni dieci: Altri gruppi avvieranno progetti di intelligenza artificiale generale, in particolare dal 2015 in poi. Questi gruppi diventeranno sempre più mainstream, seri e ben finanziati. Tutto ciò sarà guidato da computer più veloci, migliori algoritmi di apprendimento automatico e una migliore comprensione dell'architettura del cervello. Alcuni di questi gruppi produrranno piccole intelligenze artificiali generali che impareranno a fare alcune cose interessanti, ma non saranno neanche lontanamente vicine al livello dell'intelligenza umana. Tuttavia, prepareranno la strada a questo. La preoccupazione per i pericoli dell'intelligenza artificiale diventerà meno marginale, ma non diventerà mainstream.

Una delle previsioni più interessanti di Shane è la più sbagliata. A suo avviso, alla crescita di capacità di calcolo e all'evoluzione tecnologica dell'industria dei semiconduttori non corrisponderà una domanda adeguata da parte dei consumatori. In particolare, ci sarà "il collasso del mercato delle schede grafiche". La previsione è inusuale e Shane stesso dopo qualche giorno deve pubblicare una correzione e dire che quel crollo non avverrà nel giro del decennio. In realtà peggiora la situazione, perché attribuisce un possibile collasso "all'aggressività di aziende come Intel, che diviene aggressiva e costruisce GPU all'avanguardia all'interno dei suoi chip di CPU, rendendo così ridondanti le GPU". Insomma, Shane nel 2009 formula previsioni sullo sviluppo dell'intelligenza artificiale ma non ne ha capito il motore, mentre Bryan Catanzaro e Andrew Ng hanno già scritto i loro paper sulle GPU e stanno avviando una grande trasformazione basata sull'irrefrenabile interesse umano per i gatti.

2. Tre leggi muovono il mondo

Tre leggi muovono il mondo del calcolo e accompagnano l'evoluzione tecnologica, nel XX e XXI secolo: la legge di Moore, la legge dei ritorni acceleranti, la legge di Huang.
La legge di Moore è così chiamata in onore del chimico Gordon Moore, il co-fondatore di Intel morto nel 2023, che in un articolo del 1965, di tre anni precedente alla fondazione dell'azienda, pubblica un grafico che copre il periodo

dal 1959 al 1975 per illustrare il raddoppio annuale dei numeri di transistor nei circuiti integrati, e il conseguente aumento esponenziale della potenza di calcolo. Questa legge empirica inizialmente osserva quello che è avvenuto in precedenza nel mercato dei semiconduttori e ne prevede l'evoluzione, per un periodo limitato di dieci anni. Nel 1975, Moore aggiorna la sua previsione: il raddoppio avviene circa ogni due anni.

La resistenza della legge di Moore ben oltre la revisione iniziale mette in luce la sua potenza, grazie alla quale Intel ha introdotto il concetto di "siliconomia" (economia del silicio) per mostrare l'impatto dell'industria dei semiconduttori sul PIL mondiale, stimato dal 1995 al 2015 di 3000 miliardi di dollari diretti e di 11.000 miliardi di dollari indiretti. In questa vicenda, Moore emerge come figura di straordinaria modestia, perché non ha mai affermato di aver inventato la legge che muove il mondo. L'espressione "legge di Moore" emerge in un momento indefinito all'inizio degli anni settanta e viene introdotta da Carver Mead, professore al Caltech dove tiene anche un corso con Feynman e Hopfield all'inizio degli anni ottanta5: è amico e collaboratore di Moore, con cui conversa a lungo prima delle sette del mattino, quando gli altri non sono ancora arrivati in ufficio. Mead ha spiegato in numerose occasioni che Moore coglie il processo dal punto di vista economico (prodotti complessi più piccoli, più valore, costi più bassi), mentre lui si concentra sulla parte scientifica, la garanzia che la fisica consenta quell'evoluzione.

L'azienda simbolo della legge di Moore, ovviamente, è Intel, e il prodotto per eccellenza è il personal computer. In The Age of Spiritual Machines, Ray Kurzweil allarga la prospettiva della legge di Moore. A suo avviso, la crescita esponenziale della capacità di calcolo non inizia con quella legge, ma risale all'inizio del XX secolo. Tutto il lavoro di Kurzweil è incentrato sulle tendenze esponenziali nella tecnologia, che si applicano anche alla capacità di calcolo. Secondo Kurzweil, dispositivi come l'Analytical Engine del 1900, il tabulatore Hollerith del 1908, il calcolatore Monroe del 1911, il tabulatore IBM del 1919 e il National Ellis 3000 del 1928 sono esempi di questa crescita esponenziale iniziale. La velocità dei computer, misurata in calcoli al secondo per mille dollari, raddoppia ogni tre anni tra il 1910 e il 1950, ogni due anni tra il 1950 e il 1966, e poi ogni anno: c'è una "crescita della crescita esponenziale", non un rallentamento.

Tale crescita si applica anche ad altre tecnologie e altri mercati, come le telecomunicazioni e le biotecnologie, e quindi è più vasta e ambiziosa della legge di Moore. Nel nostro caso, la capacità di calcolo supererà gli esseri umani: le macchine oltrepasseranno la quantità di dati che può essere processata dal cervello, e saranno sempre più economiche (consumeranno meno energia) per svolgere questi compiti. Il cambiamento quantitativo sarà un cambiamento qualitativo, perché il superamento sarà così significativo da lasciare la specie umana disarmata in termini intellettuali: a un certo punto, sarà incapace di comprendere le capacità delle macchine.

La legge dei ritorni accelerati, così intesa, implica che questo evento, noto come "singolarità", avverrà nel ventunesimo secolo, con proiezioni che lo collocano attorno al 2040, quando 1000 dollari di potenza computazionale potrebbero suppergiù equivalere a un milione di cervelli umani. La singolarità è un momento decisivo, in cui si porrà la questione dell'integrazione e della fusione dell'uomo con la macchina.

La misura indicata da Shane, comune per i supercomputer, calcola le operazioni al secondo: 1018 FLOPS è la cosiddetta exascale, cioè un miliardo di miliardi di operazioni al secondo. Nella sintesi di Scott Atchley, exascale vuol dire gli abitanti della terra, otto miliardi di persone, che lavorano contemporaneamente sullo stesso problema con una calcolatrice, potendo fare almeno un'addizione o una moltiplicazione al secondo: così, in quattro anni, potrebbero fare ciò che il supercomputer che ha raggiunto la soglia exascale, Frontier, fa ogni secondo. Frontier si trova presso l'Oak Ridge National Laboratory in Tennessee, e ha raggiunto la soglia nel 2022, due anni dopo la previsione di Shane. D'altra parte, un sistema di calcolo distribuito a fini scientifici, Folding@Home, ha raggiunto la soglia nel 2020, esattamente l'anno della sua previsione.

Una variante della legge dei ritorni acceleranti è espressa nel 2021 da Sam Altman, co-fondatore e amministratore delegato di OpenAI, che parla di una "legge di Moore per tutto". Come abbiamo visto, la legge di Moore, oltre alla sua modestia iniziale, ha un obiettivo specifico e incide su un mercato preciso in un dato periodo storico. Già l'allargamento della legge implica un suo travisamento, o quanto meno un cambiamento profondo, ma comunque l'allineamento di incentivi continua, la profezia che si autoavvera procede. Il cambio di fase di Kurzweil e Altman allarga invece l'orizzonte a un futuro tecnologico generale, in cui i computer faranno tutto e "espanderanno il nostro concetto di 'tutto'". E questo tutto, in termini di bisogni e di merci, sarà esponenzialmente più accessibile. "Immaginate un mondo in cui, per decenni, tutto – alloggio, istruzione, cibo, vestiti ecc. – ogni due anni costa la metà." La "legge di Moore per tutto" non viene da una "teoria del tutto", perché come abbiamo detto non siamo in quel campo, ma da un "prodotto del tutto", l'intelligenza artificiale.

È un piano di distribuzione, inclusività, attuazione. “Un grande futuro non è complicato: ci serve la tecnologia per creare più ricchezza, e la politica per distribuirla in modo giusto. Ogni cosa necessaria sarà abbordabile, e ognuno avrà abbastanza soldi per permettersela.” Avremo la libertà di un lavoro creativo, fornita dall’abbondanza: in termini weberiani, ogni lavoro sarà un lavoro dello spirito. L’elemento politico deve rispondere a questo processo, perché “i cambiamenti che stanno arrivando non possono essere fermati”. La scienza non sta su palafitte, come nell’immagine di Popper, ma su GPU. Per controllare e falsificare dobbiamo avere GPU, altrimenti non possiamo prendere in considerazione e calcolare un numero ragionevole di casi. Se vuoi la scienza, devi avere i supercomputer: non esiste un’altra strada, non esiste un altro modo di procedere. Quello che le GPU stanno facendo, sullo sfondo del chiacchiericcio sul crypto, è decisivo: reinventare le batterie a ioni di litio, mappare il nucleo della terra, comprendere e simulare il tempo, capire la struttura dell’HIV. Per fare tutto questo, con modelli che avranno una taglia sempre più grande, con supercomputer sempre più grandi, bisogna accelerare. Mostrando la differenza tra la GPU del 2013, chiamata Fermi, e quella del 2018, chiamata Volta, Jensen dice che l’accelerazione compiuta in questo periplo italiano è stata di venticinque volte.

Venticinque volte in cinque anni. La legge di Moore è dieci volte in cinque anni. La legge di Moore è la legge miracolosa. Ha abilitato praticamente ogni industria. Il progresso della scienza, il progresso della società: dieci volte ogni cinque anni. Le nostre GPU hanno accelerato le simulazioni di dinamica molecolare per venticinque volte negli ultimi cinque anni. C’è una nuova legge. È una legge potenziata, una superlegge. C’è una nuova legge. Adesso esiste un nuovo mondo, il mondo delle GPU, che traina l’apprendimento profondo e l’intelligenza artificiale. Jensen parla al suo uditorio di ormai noti scienziati e ricercatori dell’informatica e dell’intelligenza artificiale.

L’addestramento per AlexNet, a fine 2012, ha impiegato sei giorni in due GTX580 di NVIDIA, mentre poco più di cinque anni dopo, all’inizio del 2018, l’ultimo prodotto di NVIDIA, DGX-2, può realizzare lo stesso processo in diciotto minuti: Grazie, legge di Moore. Hai fatto il tuo tempo. Ora, sostiene Jensen, “il mondo vuole una GPU gigantesca. Non una GPU grande. Una GPU gigantesca”. Per averla, ha bisogno di una nuova legge, una superlegge.

La rivista dell’Institute of Electrical and Electronics Engineers, “IEEE Spectrum”, titola: “Spostati, legge di Moore, fai spazio alla legge di Huang”. Nel 2020, il “Wall Street Journal” ribadisce il concetto: la legge di Huang è la nuova legge di Moore. In senso tecnico, la ragione per cui Dally ritiene che la legge di Moore sia superata è la sua relazione con la scalabilità di Dennard (Dennard scaling), denominata anche legge di Dennard: in estrema sintesi, la costanza di consumo energetico a fronte della crescita dei transistor.

Secondo Dally, non può esserci scalabilità delle prestazioni dei processori se non si considera la scalabilità della potenza energetica, oltre a quella dei transistor. I bisogni, economici e sociali, che richiedono un aumento costante del calcolo, devono essere serviti da un nuovo paradigma, che ovviamente è quello del calcolo parallelo, espresso dalle GPU. Dally è criticabile in quando attribuisce a Moore una parte degli argomenti di Dennard, ma è innegabile che vi sia una parte dello storico articolo di Moore, denominata proprio “Il problema del calore”, che affronta la questione energetica come parte dell’evoluzione complessiva: “Sarà possibile rimuovere il calore generato da decine di migliaia di componenti in un singolo chip di silicio? […] Ridurre le dimensioni su una struttura integrata rende possibile azionare la struttura a velocità più elevata per la stessa potenza per unità di superficie”. E la legge di Dennard cessa di avere efficacia già all’inizio del XXI secolo, attorno al 2004-2005: questi limiti sono riconosciuti da un’ampia letteratura, e anche dai ricercatori di Intel. Il destino di leggi del genere, d’altra parte, è essere superate. La controversia in cui si inserisce Dally riguarda quindi la capacità della legge di Dennard di trascinare con sé il “miracolo”, la legge di Moore

3. Il gioco delle perle dei gesuiti

Nel 1973, in una visita scientifica a Boston, Poggio incontra Marvin Minsky e David Marr nel laboratorio di intelligenza artificiale del MIT, e coglie l’occasione per invitare quest’ultimo ai seminari scientifici che si tengono a Erice, in Sicilia. Marr, nato nel 1945, ha già scritto tre importanti paper sul funzionamento del cervello, dedicati al cervelletto, alla neocorteccia e all’ippocampo. Poggio e Marr cominciano a lavorare insieme, e condividono la passione del volo. Tra gli studenti di Marr al MIT c’è il fisico israeliano Shimon Ullman. Nel 1980 Poggio si trasferisce al MIT per continuare le ricerche sul cervello e la visione artificiale, ma nello stesso anno Marr muore di leucemia, a soli trentacinque anni.

Nella sua nuova carriera negli Stati Uniti, Poggio trova un’opportunità professionale per rimanere nell’ambito universitario arrotondando i guadagni. È un’azienda chiamata Thinking Machines, fondata nel 1983 da uno studente

di Marvin Minsky, Danny Hillis, che aveva fatto la tesi di dottorato sul calcolo parallelo, insieme a una dottoranda di pianificazione urbana, Sheryl Handler. L'azienda inizia la sua attività grazie a un consistente contratto con apparati di difesa degli Stati Uniti. Il motto aziendale è "Costruire un computer che sia orgoglioso di noi", ma intanto devono esserne orgogliosi i generali, per ritenere i suoi calcoli paralleli utili per obiettivi militari specifici, dall'identificazione dei nemici al trattamento dei dati satellitari.

Il più grande orgoglio di Poggio sono i suoi allievi. Si sofferma spesso sul successo imprenditoriale di due di loro, che hanno già caratterizzato la nostra epoca dell'intelligenza artificiale: Amnon Shashua e Demis Hassabis. Shashua è uno studente israeliano che arriva al MIT nei primi anni novanta. Ha iniziato l'università dopo il servizio militare nelle forze di difesa israeliane, durante la prima guerra del Libano. Studia matematica e informatica a Tel Aviv e all'Istituto Weizmann, laureandosi con l'allievo di Marr, Shimon Ullman, prima di trasferirsi negli Stati Uniti, dove Poggio diviene il suo mentore. Shashua combina in modo distintivo gli studi sulla visione artificiale e il riconoscimento degli oggetti, già al centro della sua tesi di dottorato, con l'attività imprenditoriale. Nel 1999, mentre Jensen lancia la GPU, Shashua fonda a Gerusalemme Mobileye, azienda dedicata all'assistenza della guida e alla guida autonoma, con la realizzazione sia di prodotti di elettronica che di algoritmi di visione artificiale. Nel 2014 Mobileye si quota a New York e nel 2017 viene acquisita da Intel per la cifra record in Israele di 15,3 miliardi di dollari.

L'intreccio tra le due "industrie delle industrie" (quella dell'automobile e quella dei semiconduttori) è già una realtà della nostra epoca, in termini tecnologici, economici, politici: il contenuto elettronico delle auto continua ad aumentare, e con esso aumentano gli elementi di controllo autonomo. In questo contesto, i numeri di Mobileye a partire dall'acquisizione hanno rappresentato spesso uno degli elementi più positivi di Intel, mentre altre aree dell'azienda hanno perso terreno verso Samsung e TSMC.
Avere "l'intelligenza in tasca". L'impatto può essere rivoluzionario, perché "non si tratta solo di fare una domanda, ma di avere una conversazione con Albert Einstein, con un filosofo, con qualsiasi tipo di esperto ti venga in mente. Sarà tutto nelle tue tasche. Ciò sblocca aspetti che oggi sono persino difficili da immaginare. E la frontiera linguistica apre la porta all'intelligenza generale. Cinque anni fa, se un esperto di intelligenza artificiale avesse parlato di intelligenza generale, sarebbe stato trattato con scetticismo. Oggi penso che sia dietro l'angolo". L'interesse per l'intelligenza artificiale applicata al linguaggio, con l'ascesa di un enorme mercato già disponibile in attesa del mantenimento delle promesse della guida autonoma, ha perfino generato un riavvicinamento con Jensen.

Il 5 aprile 2022, Poggio presenta con orgoglio l'altro allievo che, dopo la pausa della pandemia, è tornato a trovarlo di persona al centro: Demis Hassabis. Lavora con un grande creatore di videogiochi, Peter Molineux, e il risultato della loro collaborazione è Theme Park del 1994, dedicato alla gestione di parchi giochi: vende oltre quindici milioni di copie e raggiunge un'enorme popolarità in Giappone. Da un lato, i videogiochi sono per Hassabis la porta d'ingresso all'intelligenza artificiale, perché sono l'arena in cui negli anni novanta questa scienza incerta viene sperimentata e fatta avanzare. Demis alterna alle sedute sfiancanti di programmazione le letture delle saghe di fantascienza, da Asimov a Iain Banks, a un approfondimento teorico: proprio mentre lavora a Theme Park, legge Gödel, Escher, Bach di Hofstadter. Dall'altro lato, attraverso i videogiochi Hassabis fa esperienza del fallimento: la casa di produzione che fonda poco più che ventenne è un buco nell'acqua. Il suo sistema avanzato per lui è un modo di esplorare l'intelligenza artificiale, ma non convince gli utenti. Senza prodotti adeguati, senza mercato, il suo grande disegno non può procedere. Non si tratta solo di giocare, ma di creare. Il giocatore assoluto è colui il quale sa creare un gioco.

Tra i primi investitori esterni, nel 2011, ci sono il Founders Fund di Peter Thiel e Horizons Ventures, il venture capital che fa capo al magnate di Hong Kong Li Ka-shing; in seguito una piccola quota viene acquisita da Elon Musk. Cruciale è l'appoggio di Thiel, dopo molti infruttuosi tentativi che DeepMind compie per ottenere capitali in Gran Bretagna. La comunità finanziaria londinese non si mostra interessata a progetti troppo avveniristici, senza chiarezza sui ritorni economici: questo segna una distanza incolmabile con la Silicon Valley. Nella City ci sono i soldi, certo, ma non c'è nessuna reale capacità di mobilitarsi per grandi scommesse tecnologiche.

Thiel e Demis si incontrano nell'estate 2010, a margine della conferenza del Singularity Summit a San Francisco, dove anche Shane ha la possibilità di illustrare le sue teorie sull'intelligenza e di rivendicare la paternità dell'etichetta AGI. Demis sa di non avere tempo: deve impiegare un minuto per convincere il primo finanziatore esterno di Facebook a investire nella loro azienda e per mesi pensa a quale sia l'approccio migliore. Demis ha letto a nove anni Il Signore degli Anelli, testo sacro per Thiel, ma non cerca con lui quel tipo di connessione. Sceglie di puntare tutto sugli scacchi. Parla con Thiel della natura del gioco, dell'equilibrio delle mosse, del rapporto tra il cavallo e l'alfiere. Pochi mesi dopo, Thiel autorizza personalmente l'investimento di Founders Fund per 1,4 milioni di sterline, e con

la sua reputazione e la sua rete garantisce risorse sufficienti per la fase di start-up dell'azienda. È una delle prime volte che Thiel fa un investimento fuori dagli Stati Uniti, ma non capisce cosa ci faccia a Londra un'azienda come DeepMind.

Londra dà un vantaggio competitivo nell'acquisizione dei talenti, nell'accesso a università britanniche ed europee, a persone col dottorato in fisica e matematica che cercano un'alternativa all'approdo classico per chi vuole monetizzare i propri studi, l'industria finanziaria, né continuare la carriera accademica. Dopo essersi assicurata la start-up del trio di AlexNet, è Google a emergere vincitrice anche in questa nuova partita: nel 2014 acquisisce DeepMind per circa 400 milioni di sterline. All'interno dell'impero Google, DeepMind inizialmente conserva la sua autonomia. E Google paga. Le perdite di DeepMind sono coperte dalle linee finanziarie di Google. Nel 2018 la perdita è di 470 milioni di sterline, nel 2019 di 477 milioni. Nel 2020, per la prima volta c'è un profitto di 43,8 milioni, dovuto anche a un fatturato che sale nello stesso anno a 826 milioni – dai 265 milioni del 2019 –, a fronte di un limitato aumento delle spese a 717 a 780 milioni. DeepMind resta in una traiettoria di crescita nel 2021, con un fatturato che supera il miliardo di sterline (1365 milioni) e costi più consistenti (1254 milioni). Pertanto, il profitto è di poco superiore ai 100 milioni.

DeepMind contribuisce al miglioramento dell'esperienza dei video di YouTube, alla riduzione di circa il 30% del consumo energetico dei data center, alla qualità della sintesi vocale realizzata da Google Assistant con il prodotto WaveNet. Sono prodotti concreti, che indicano una strada di scalabilità per l'intelligenza artificiale, tanto ristretta quanto applicata. Per esempio, alla riduzione del consumo energetico nelle strutture produttive ad alta intensità e a una migliore gestione delle risorse idriche. DeepMind applica le sue tecniche di ottimizzazione, i suoi programmi, a processi gestionali che fanno parte dei costi aziendali di un grande conglomerato tecnologico per renderlo più efficiente. Allo stesso tempo impara da Google, che ha in mano, oltre alle risorse per assicurarsi la capacità di calcolo, ciò che va calcolato e appreso: un'enorme base di dati, un laboratorio in cui sperimentare. E Google, con le sue risorse, può sostenere una struttura di ricerca molto più ampia di quella di DeepMind. Può comprare il prezioso tempo del trio di Toronto.

Fin dalla sua fondazione, DeepMind ha allenato le sue intelligenze artificiali come agenti in diversi ambienti di gioco, in cui si intrecciano le vite dei fondatori e dei dipendenti, da Demis a Shane, fino a Oriol Vinyals, il co-autore di Ilya Sutskever che raggiunge DeepMind da Google. Si va da classici come Space Invaders e Breakout fino a complessi giochi di strategia in tempo reale, come StarCraft II della Blizzard. AlphaFold è il progetto di DeepMind legato a un notevole problema scientifico che ossessiona da tempo Demis, fin dalle discussioni al pub con gli amici di Cambridge, negli anni in cui ascolta i dischi dei Prodigy: il cosiddetto protein folding, la determinazione della forma di una proteina a partire dalle catene di amminoacidi che la costituiscono.

AlphaFold costruisce il più ampio e accurato database, liberamente disponibile, delle più di duecento milioni di proteine conosciute. Secondo Demis, più di un milione di ricercatori ha utilizzato le strutture previste da AlphaFold. DeepMind ha lanciato uno spin-off, Isomorphic Labs, per accelerare la realizzazione di nuovi farmaci attraverso l'intelligenza artificiale. Il matrimonio tra biologia e intelligenza artificiale è di chiaro interesse per Demis, il quale ritiene che il modo fondamentale per descrivere l'universo sia "l'informazione", che "la comprensione della fisica in termini di teoria dell'informazione" sia il modo migliore per comprenderlo. Se oggi possiamo dire che "la macchina si guasta, cancella, ma non può dimenticare" e che "forse l'uomo è rimpiazzabile in tutto, non certo nella sua angoscia", anche questo è messo in discussione dall'ipotesi ultima del neuroscienziato Demis. Nelle sue vite parallele: deve essere sia il ricercatore dell'oblio e dell'immaginazione sia l'imprenditore della vita come informazione.

La "soluzione" dell'intelligenza artificiale, pertanto, è la "soluzione" della vita stessa: deve applicarsi alla biologia, deve farsi conoscenza della vita, delle forme che la rendono riproducibile, come le altre unità di informazione. La vita è programmata, la vita è programmabile. Demis rivendica il suo credo: al mantra della Silicon Valley "Move fast and break things", usato da Facebook fino al 2014 e reso celebre dal film The Social Network del 2010, occorre contrapporre il metodo scientifico. Demis sembra dire: sì, la Silicon Valley mi ha dato i soldi per portare avanti DeepMind, ma quel mondo non mi appartiene, il mio mondo è quello della scienza, questo è il riconoscimento che cerco, e questo deve essere il percorso dell'intelligenza artificiale; l'avventura di Demis e Shane sarà interessata dalla curiosa e sorprendente ascesa di un altro attore, nel suo cortile di casa: OpenAI.

Negli anni novanta, il settore dei videogiochi strategici in tempo reale vede l'ingresso di un'azienda fondata da tre laureati dell'Università della California di Los Angeles: la Blizzard. Il suo successo accelera in breve tempo. Nel 1994 inizia la saga di Warcraft, nel 1995 esce Warcraft 2. Dopo l'uscita del videogioco Diablo nel 1997, la nuova

scommessa della Blizzard è superare le ambientazioni fantasy dei suoi primi successi per portare i videogiochi di strategia in una galassia lontana nel tempo e nello spazio. Il nuovo videogioco, StarCraft, è il più venduto del 1998.

La forza della Blizzard non sta solo nel fascino delle storie, nella grafica delle ambientazioni e nella giocabilità, ma in una piattaforma per giocare online, lanciata inizialmente a fine 1996: Battle.net. La possibilità di giocare online ha risultati dirompenti in Corea del Sud, dove genera un cambiamento sociale in grado di convertire ogni scettico sul ruolo epocale dei videogiochi nella storia recente dell'umanità. Il paese vede in quegli anni una proliferazione degli Internet café, chiamati PC Bang (letteralmente "sala PC"), migliaia di locali dove i clienti possono avere una connessione a Internet relativamente veloce e a poco prezzo, ed essere in grado quindi di giocare.

In Spagna, c'è Oriol Vinyals, che all'uscita di StarCraft ha quindici anni. Parlando di quei tempi, Vinyals ha ricordato un trilemma: non si poteva contemporaneamente studiare, stare con una ragazza e giocare a StarCraft. Era possibile, con fatica, portare avanti due di queste attività, giammai tutte e tre. Gli esseri umani hanno un tempo limitato. Il fascino di StarCraft è stato discusso a diversi livelli. Oltre al segreto dell'esperienza del gioco (quella che in gergo si chiama "giocabilità"), si fa spesso riferimento all'interesse suscitato dalle tre civiltà della saga. È possibile impersonare gli Zerg, uno sciame di insetti alieni veloci e letali, basati su Alien, oppure i Protoss, antichi ed evoluti abitanti della galassia con poteri psionici, o i Terrestri (Terran), litigiosi cowboy pseudoeredi dell'umanità. Ogni civiltà ha pregi e difetti che giocatori, attori e spettatori del loro scontro, all'inizio del XXVI secolo, devono imparare a conoscere mentre cercano di battere l'intelligenza artificiale e, in seguito, gli altri giocatori umani che sfidano online.

 # 4. Le grandi trasformazioni

Satya Nadella nasce nel 1967 in India, a Hyderabad. Da bambino, lo osservano due poster nella camera da letto dei genitori: uno ritrae l'inconfondibile barba di Karl Marx, passione politica del padre burocrate di simpatie socialiste; l'altro, voluto dalla madre studiosa di sanscrito, ritrae la dea della prosperità e della bellezza, Lakshmi. Davanti a questi due modelli, Nadella cerca un'identità personale. La trova nel poster di un campione di cricket. E fa sul serio. A lungo, Nadella intende fare del cricket la sua professione. Ma c'è qualcosa che suo padre gli insegna. Appartiene al gruppo di burocrati che cerca di costruire lo stato indiano.

Quando si sveglia nel cuore della notte, il figlio lo trova mentre legge un grosso libro sul letto, circondato da carte, da documenti. È la vita di un "costruttore di istituzioni" secondo cui "il vero impatto di una persona può essere valutato solo dopo che quella persona ha lasciato il suo incarico attuale". "La produzione per mezzo della macchina in una società commerciale implica in realtà una trasformazione che può essere paragonata a quella della sostanza naturale e umana della società, in merci. La conclusione per quanto macabra è inevitabile; niente di meno potrà bastare allo scopo: ovviamente lo sconvolgimento causato da questi strumenti spezzerà i rapporti dell'uomo e minaccerà di annientamento il suo ambiente naturale."

Questo è il concetto della società di mercato e di macchine, dove un aspetto è essenziale all'altro. Il concetto per Polanyi descrive già l'imminente pericolo, la catastrofe. Certo, quei "rapporti dell'uomo" che Polanyi cerca di definire attraverso le sue ricerche antropologiche sembrano in certo modo corrispondere al mito di una buona società preindustriale, capace di governarsi perché stretta nei suoi confini, che erano anche i confini di una vita breve e di malattie.

L'economia dell'uomo, di regola, è immersa nei suoi rapporti sociali. Ciò sottolinea l'importanza delle passioni umane "dirette verso fini non economici", o comunque non riducibili alla logica del guadagno. Motivazioni e inclinazioni che resistono alla società di mercato e di macchine, che cerca di rispondere al problema di governare crescenti masse di poveri, mettendoli all'opera per il funzionamento di macchine tecniche e sociali. Nella sua lunga esplorazione degli autori che rispondono ai dilemmi della società industriale, Polanyi indica una guida, forse l'unica strada che ritiene davvero promettente: il movimento di Robert Owen. Per Owen, al contrario del luddismo, il percorso non è cercare di arrestare la macchina. Non è possibile alcun ritorno all'esistenza rurale. Si tratta di "scoprire una forma di esistenza che renda l'uomo padrone della macchina".

Owen "enfaticamente sosteneva di non essere nemico della macchina". L'uomo non è subordinato alla macchina, perché la sua natura sociale viene salvaguardata dall'organizzazione attraverso la cooperazione. Una "religione dell'industria" che pensa "l'uomo come un tutto", senza alcuna separazione tra economia e politica, con la continua affermazione della loro connessione. Saper leggere la propria epoca, del resto, è cercare costantemente la "conoscenza delle connessioni, la visione della totalità".

Nel 1958, verso la fine della sua vita, Polanyi traccia un bilancio. Afferma che la rivoluzione industriale è stata uno "spartiacque della storia dell'umanità" perché ha rappresentato il legame e l'accelerazione di "tre forze, la tecnologia, l'organizzazione economica e la scienza". Questo è il "vortice sociale, il quale continua ancora ad avvolgere con un impeto irresistibile milioni e milioni di persone". L'insieme di queste forze aumenta la "velocità" di tutto. Questa corsa vertiginosa può avvenire, nota Polanyi, "tanto attraverso il mercato, quanto attraverso la pianificazione".

L'Occidente ha un compito storico, divenuto ormai una "necessità per la sopravvivenza": "disciplinare le proprie creature" diffuse per tutta la terra. Polanyi non vede separazione tra i problemi della guerra fredda: l'età nucleare con le minacce di un conflitto atomico, le rivoluzioni in Asia che rendono instabile l'area più popolosa del pianeta, l'emergere dei problemi ambientali su cui già si sofferma nella Grande trasformazione. Sulla scissione dell'atomo Polanyi ritorna in vari scritti, paragonando il suo impatto alle varie reazioni alla rivoluzione industriale: ormai "la macchina è andata al di là dell'immaginazione di scrittori in cerca di ispirazione.

Amartya Sen, che ha poi svolto fondamentali ricerche economiche sul problema della carestia in India, da ragazzo negli anni quaranta, alla scuola di Santiniketan, già si interrogava – anche nel confronto filosofico con Adam Smith, che l'ha impegnato per tutta la vita – sugli effetti dell'imperialismo britannico, mentre in India si iniziava a respirare l'aria dell'indipendenza. Sen ricorda che anche per l'India, come del resto per la Cina, è sempre esistito un filone che ha rivendicato i grandi risultati del passato in matematica, letteratura, musica, medicina, astronomia e molte altre discipline, nonché la potenza commerciale indiana prima dell'era coloniale. Eppure, Sen e i suoi compagni di classe, impegnati in accesi dibattiti, non possono sottovalutare l'arretramento indiano – ancora, in un percorso parallelo a quello cinese – rispetto all'accelerazione della storia europea, del suo primato tecnico, della sua brama di conquista. Sen e i suoi compagni non leggono Polanyi ma studiano gli scritti di Marx che, nella rivendicazione più generale della potenza e della capacità del capitalismo, considera l'importanza del governo coloniale per lo sviluppo dell'India.

La tecnologia è "incarnazione materiale della libertà, creatrice di vita e della sua abbondanza", consente di diffondere la ricchezza. L'uomo di Polanyi sta ancora "all'alba di una civiltà tecnologica" e quindi è collocato in un interregno. La tecnologia mostra l'esistenza "precaria", irrisolta della società dove la "pace dipende da chi schiaccia i bottoni", dove avviene un confronto costante tra la decisione del potere politico e la continua opportunità generata dal potere tecnologico. Questo "pericolo razionale" richiede sempre da parte della società un "potere sufficiente". In questo ragionamento su potere e tecnologia negli ultimi scritti di Polanyi si può cogliere l'evoluzione di un concetto cruciale per La grande trasformazione, che Nadella di certo non ha dimenticato: il doppio movimento. Con esso, Polanyi identifica la forza sociale e politica, con le sue conseguenze non intenzionali, che fa da contraltare all'espansione continua del mercato.

Con un bottone automatico non può esserci alcun doppio movimento, poiché la creazione di vita e abbondanza della tecnologia procede da sola. E, ammesso invece che il bottone esista, chi lo preme sa come funziona? Riesce a capirne i meccanismi, o gli sono totalmente estranei, lontani dalla vista, lontani dalla comprensione? "What I cannot build, I cannot understand.".

In questo processo pluridecennale, le scelte degli studenti cinesi divengono un metronomo dell'integrazione tra Pechino e Washington, e del suo dilemma. Il numero di studenti giunti dalla Repubblica Popolare Cinese agli Stati Uniti dagli anni settanta è stimato a oltre tre milioni. In questo secolo, triplicano nel giro di un decennio, dai 130.000 dell'anno accademico 2009-2010 ai 370.000 del 2019-2020, per poi scendere sotto i 290.000 nel 2022-2023. Secondo il Dipartimento del Commercio, contribuiscono nel 2018 per quasi 15 miliardi di dollari all'economia degli Stati Uniti. Il 2020 rappresenta uno spartiacque, per via delle politiche restrittive sotto l'amministrazione Trump e di un deterioramento delle relazioni che non sembra più recuperabile. In altri contesti, gli studenti cinesi continuano a crescere: per esempio, nell'immancabile Università di Toronto sono passati da 6000 nel 2013-2014 a 15.700 nel 2022-2023. Ma esiste un altro bacino di studenti con un'ampiezza e un impatto comparabile: si tratta, ovviamente, di quelli indiani, che nel 2022-2023 raggiungono il record di 268.923, dopo una crescita costante di tre anni, e sono ormai un quarto di tutti gli studenti stranieri negli Stati Uniti.

C'è un video del 1993, divenuto virale, in cui Nadella, specializzatosi in informatica in Wisconsin ed entrato in Microsoft l'anno precedente, spiega il funzionamento di Excel, una delle applicazioni di maggiore successo dell'azienda fondata da Bill Gates e Paul Allen. Mentre i tre di Denny's fondano NVIDIA, lui spiega Excel agli sviluppatori. L'informatico e manager indiano comincia dal basso a farsi strada nell'azienda che domina l'economia digitale di quella stagione, con la diffusione del personal computer e dei suoi software.

Durante il XXI secolo, nella sua accelerazione, Microsoft non sembra più avere un ruolo dominante. È diventata "l'equivalente tecnologico di un'azienda automobilistica di Detroit, che porta nella catena di montaggio modelli più appariscenti del solito prodotto mentre i suoi concorrenti sconvolgono il mondo". Nel 2000 Microsoft è la più grande azienda al mondo per capitalizzazione, con oltre 500 miliardi di dollari; nel 2012 scende sotto i 250 miliardi, mentre nello stesso periodo altri giganti hanno occupato la scena, a partire da Apple, Google e Facebook.

I ragazzi che un tempo hanno schernito e superato IBM ormai hanno perso il tocco. Microsoft è la nuova IBM. Un gigante arenato, paralizzato, che non sa più agire ma solo reagire. L'emblema di questo declino è l'acquisto nel 2013 della divisione telefonica di Nokia. L'esperimento fallito che nel 2016 è già costato 8 miliardi di dollari e la perdita di migliaia di posti di lavoro. Ma i necrologi sull'azienda si rivelano prematuri. Nel 2011, Satya Nadella è nominato presidente di un'area di Microsoft al tempo sottovalutata, Server and Tools Business (STB), e procede alla sua riorganizzazione. Questa è la sua "grande trasformazione": capire che la fornitura di servizi attraverso il cloud è la principale area di crescita per il vecchio gigante. Il cloud computing, secondo Nadella, è un'opportunità della stessa proporzione del personal computer: "Renderà possibile per le aziende e per le nazioni di ogni dimensione sfruttare al meglio le ultime tecnologie per migliorare la produttività e le vite delle persone".

Il cloud, secondo la stessa definizione di Microsoft, "non è un singolo supercomputer, ma una rete globale interconnessa di milioni di computer in data center in tutto il mondo che lavorano insieme per archiviare e gestire i dati, eseguire applicazioni, fornire contenuti e servizi". Nel 2021, Microsoft afferma di avere oltre duecento data center in trentaquattro paesi, connessi da 165.000 miglia di fibra ottica, terrestre e sottomarina. L'anno precedente l'azienda ha anche lanciato il prototipo di un data center sottomarino.

A giugno 2008 Amazon ha già 180.000 sviluppatori che costruiscono applicazioni e servizi per la sua piattaforma cloud. La Microsoft in quel momento subisce ancora la crisi del suo decennio perduto: il rallentamento delle vendite dei personal computer, l'ascesa degli smartphone in cui non ha vantaggi competitivi, l'uscita di scena di Bill Gates, ormai concentrato su attività filantropiche, la grande recessione. Ma in questa tempesta perfetta, Nadella inizia a fare le sue mosse, a partire dalla costruzione di un motore di ricerca, Bing, all'inizio ridicolizzato perché chiaramente incapace di competere con l'invincibile Google, ma che rappresenta una delle premesse tecnologiche dei servizi di cloud per le imprese. Nadella si concentra sulla parte cloud in termini di riorganizzazione aziendale e di risorse umane, con la valorizzazione delle competenze esistenti e il reclutamento di nuovi manager, anche da Amazon Web Services.

Nadella allontana Microsoft dall'ossessione sugli standard proprietari nella lotta all'open source, decidendo di abbracciare una programmazione di questo tipo anche in relazione allo storico rivale, Linux. Ma Microsoft non si limita a una scelta culturale verso la comunità open source, perché ha i soldi per comprare pezzi di ecosistema. Sotto la leadership di Nadella, dopo l'ingresso nel social networking professionale con l'acquisizione di LinkedIn nel 2016 per circa 26 miliardi di dollari, nel 2018 Microsoft compra per 7,5 miliardi GitHub, una delle più popolari piattaforme per lo sviluppo di software, utilizzata al tempo da quasi trenta milioni di persone. Ciò porta alle critiche di diversi sviluppatori, poco convinti dalle promesse di "libertà, apertura e innovazione" di un'azienda che ha come fine fare più soldi coi suoi servizi, ma Microsoft continua a raggiungere i suoi obiettivi in modo sempre più coerente.

Il cambio di passo impresso da Nadella si riflette, anno dopo anno, nella capitalizzazione. Nel 2019, Microsoft diventa la terza azienda, dopo Apple e Amazon, a superare i 1000 miliardi di dollari di valore di borsa, ma Nadella non vuole che il traguardo sia celebrato: non bisogna guardare a queste stime contingenti e instabili, non si deve guardare al passato ma solo alle prospettive del futuro. Un futuro alimentato sempre di più dalla crescita e dalle performance dei data center. Le stime sul totale dei servizi cloud sono spesso divergenti, e per Microsoft c'è una differenza significativa tra indicare i soli ricavi di Azure oppure mettere insieme i servizi di Office 365 e LinkedIn.

Secondo i dati di Gartner relativi al solo mercato di servizi di infrastrutture del 2022, pari a oltre 120 miliardi di dollari, Amazon Web Services pesa per il 40%, mentre Microsoft per il 21,5%. Lontano dai due giganti, con circa 9 miliardi di ricavi, ci sono Alibaba e Google Cloud, mentre con una quota del 4,4% al quinto posto spunta Huawei; immancabili i videogiochi, ma ripensandoli attraverso la centralità del cloud. Mentre Nokia coi ricavi dalla vendita della divisione telefonica a Microsoft compra Alcatel-Lucent, e con essa quello che resta dei leggendari Bell Labs, la Microsoft di Satya Nadella affronta la nuova stagione disponendo di un avamposto di ricerca realizzato nello scorso secolo.

Nel 1998, nel pieno dell'ottimismo degli anni novanta, viene avviato Microsoft Research Asia, il laboratorio alla cui direzione viene chiamato l'informatico nato a Taipei Kai-Fu Lee, esperto di riconoscimento vocale, che dichiara: "Nel

lungo termine, vogliamo permettere ai computer di vedere, ascoltare, parlare e imparare". Kai-Fu Lee lascerà Microsoft per gestire le attività di Google in Cina dal 2005 al 2009, e in seguito – quando il mercato cinese diviene di accesso sempre più difficile per le aziende di software statunitensi – per avviare un fondo con cui finanziare le imprese cinesi. È lui a dare l'accattivante definizione di AlphaGo come "momento Sputnik" dell'intelligenza artificiale in Cina. Attraverso Microsoft Research Asia e altri investimenti, l'azienda di Redmond contribuisce in modo significativo a costruire l'ecosistema di intelligenza artificiale in Cina.

Nadella riconosce il dilemma che anche la sua azienda deve affrontare: da un lato, difende il concetto di ricerca aperta ed è convito che bloccare la collaborazione con la Cina sarebbe un fatto negativo, ma dall'altro ammette che "ogni tecnologia può essere uno strumento o un'arma". La storia di Microsoft Research Asia, come quella di ogni laboratorio alla frontiera tra Stati Uniti e Cina, è destinata alla diffidenza e all'incertezza. Uno dei ricercatori di Google, l'ucraino Illia Polosukhin, grande appassionato di fantascienza, discute dei dilemmi del film con alcuni colleghi impegnati nelle sfide dell'intelligenza artificiale e ragiona con loro sul parallelo tra la rappresentazione del film di un linguaggio non lineare, olistico, e il potenziale per un nuovo approccio all'elaborazione del linguaggio umano. L'analogia porta allo sviluppo del concetto di "auto-attenzione" (self-attention) nel modello Transformer.

È una "grande trasformazione" che supera alcuni problemi delle reti neurali ricorrenti e delle reti neurali convoluzionali, per migliorare l'elaborazione del linguaggio. Il paper che illustra questo meccanismo, presentato alla conferenza Neurips del 2017, "Attention Is All You Need", citato più di 100.000 volte, è un lavoro collettivo che ha coinvolto otto ricercatori di Google Brain, Google Research e dell'Università di Toronto, guidati dall'indiano Ashish Vaswani. Nel modello Transformer, l'auto-attenzione è il meccanismo che permette all'intelligenza artificiale di elaborare un'intera frase contemporaneamente, al contrario del processamento sequenziale, in cui ogni parola viene scandita e tradotta a sua volta. Analizzando tutte le parti di una frase simultaneamente, il modello Transformer ha una migliore comprensione del contesto, simile a quella con cui la dottoressa Banks decodifica l'insieme del linguaggio alieno in Arrival.

Transformer dell'intelligenza artificiale si adattano alla complessità del linguaggio naturale, passando da un modello di elaborazione sequenziale a un modello più dinamico e versatile. Questa capacità li rende molto efficaci nel catturare relazioni complesse e di lungo termine, non solo nell'ambito della traduzione automatica (l'interesse immediato di Google), ma anche della generazione di testi e di video.
Si tratta, nel 2017, di fornire un migliore risultato nella traduzione da una lingua all'altra, impiegando meno tempo per l'addestramento. I risultati del Transformer sono ottenuti con l'impiego di otto GPU NVIDIA per 3,5 giorni, in un set di dati di circa un miliardo di parole. Come sempre, in questi modelli c'è lo zampino di Jensen. Puoi tradurre una porzione di testo da una lingua all'altra, puoi generare testi sempre più ampi, ma dovrai sempre mettere in ordine tre lettere: G, P, U. "Cosa creerete? Qualunque cosa sia, correte verso di lei come abbiamo fatto noi. Correte, non camminate. Ricordate, o state correndo per mangiare o state correndo per non essere mangiati. A volte non si riesce a capire la differenza. Comunque vada, correte!"

Nella corsa delle schede grafiche, Jensen continua a combattere a lungo con ATI. All'inizio del secolo, nell'esplosione delle società degli anni novanta, sono rimasti soltanto ATI e NVIDIA come veri duellanti. ATI è stata fondata negli anni ottanta da studenti dell'Università di Toronto, cinesi di Hong Kong immigrati in Canada. Nel 2006, la corsa delle schede grafiche cambia perché un gigante dell'industria dei semiconduttori, nonché la prima azienda per cui ha lavorato Jensen prima di passare a LSI Logic, AMD, acquista ATI. Quali sono le conseguenze? Da un lato, Jensen non dovrà più combattere con i cino-canadesi che l'hanno a lungo impensierito, ma dall'altro lato, mentre cerca spazio rispetto a Intel e comincia l'investimento di lungo termine in CUDA che brucerà un sacco di risorse, ha davanti un avversario con le spalle larghe, che con l'acquisizione di ATI è determinato a fare dei videogiochi un'area importante dei propri ricavi. Ma le cose per AMD non vanno secondo le previsioni. Lo storico co-fondatore e amministratore delegato, Walter Jeremiah "Jerry" Sanders III, ha reso celebre l'espressione "Real men have fabs" ("I veri uomini hanno le fabbriche"). Sanders, veterano di Fairchild Semiconductor, reagiva così alla crescita esplosiva dell'ecosistema fabless degli anni novanta: la rivoluzione resa possibile da Morris Chang, che ha portato alla nascita di aziende come NVIDIA, non era una cosa seria.

La frase di Sanders senz'altro rivela il sessismo che ha spesso albergato nell'ingegneria elettronica. Non si può dimenticare, però, che la rivoluzione della progettazione è giunta dai fondamentali studi di Carver Mead e di Lynn Conway. Quest'ultima nel 1968 lavora con successo per IBM e informa l'azienda che sta affrontando una transizione di genere: per tale ragione, una delle menti più geniali della storia dell'industria si ritrova sbrigativamente licenziata, mentre la DARPA saprà accoglierla poco dopo. Oltre al sessismo, c'è un problema di modello di business: il mercato

mostra che Morris Chang ha visto e costruito il futuro, mentre Jerry Sanders è rimasto aggrappato al passato senza riuscire a salvarlo.

Il rovescio della medaglia della rapsodia sulla mascolinità delle fabbriche è che AMD, con ATI sul groppone e incapace di competere sulla frontiera tecnologica in modo profittevole, ha dovuto letteralmente abbandonare le sue fabbriche, con lo spin-off di quest'attività, grazie ai soldi dei "veri uomini" degli Emirati Arabi Uniti, in particolare il fondo sovrano Mubadala, voluto da Mohammed bin Zayed nelle sue ambizioni di tecnologia militare sempre più sofisticata. Quando ciò è avvenuto, nella brutalità del ciclo dei semiconduttori amplificata dalla crisi del 2008, l'allora amministratore delegato di AMD Dirk Meyer si è sentito obbligato a dichiarare: "Ci sentiamo ancora piuttosto mascolini ad AMD". Nonostante questa baldanza, i conti non sono migliorati. L'azienda ha rischiato la bancarotta, per poi essere salvata a partire dal 2014 da un nuovo amministratore delegato, che per la legge del contrappasso ovviamente è una donna: Lisa Su, nata a Taiwan, peraltro parente dello stesso Jensen.

Così AMD, grazie all'integrazione delle tecnologie di ATI e a una migliore gestione manageriale, ha continuato a competere con NVIDIA, pur dovendo senz'altro recuperare terreno nell'intelligenza artificiale che Jensen ha mostrato di dominare. Secondo la divisione resa celebre da Isaiah Berlin e mutuata da Archiloco, "la volpe sa molte cose, ma il riccio ne sa una grande", nella storia delle idee i ricci vedono il mondo attraverso una prospettiva avvolgente, una chiave interpretativa fondamentale, mentre la volpe si arricchisce attraverso diverse esperienze, mai riducibili a unità. Secondo questo gioco, NVIDIA è un riccio. La sua grande idea è la GPU, quello che gli altri hanno presagito, hanno sperimentato, ma che NVIDIA ha inventato in un mercato di applicazione su vasta scala. E non basta. Il riccio vuole portare la GPU su tutti i mercati possibili, vuole allargare sempre il proprio ambito d'azione. Tutto deve sottostare al grande mezzo con cui crea il grande fine, essere indispensabile per tutti i mondi virtuali. I mondi si chiudono a riccio in parallelo, in una sinfonia di GPU.

Nel 2014, quando per la prima volta saranno venduti oltre un miliardo di smartphone, per la precisione 1,24 miliardi, con una crescita di quasi il 30% sull'anno precedente, NVIDIA annuncia che abbandonerà i prodotti per smartphone e tablet per concentrarsi sugli amati videogiochi e sull'automotive. Jensen sembra avere perso il tocco magico: la centralità degli smartphone è evidente, e le sue affermazioni in parte coprono la difficoltà di competere con altri attori, tra cui Qualcomm e Mediatek, un'azienda di Taiwan che cresce in modo sostenuto anche grazie ai prezzi economici per gli smartphone asiatici. Il fallimento di NVIDIA è ancora più chiaro nel 2015, quando annuncia al mercato l'interruzione delle attività di Icera. Jensen ha perso. La sconfitta ha un importante strascico legale, perché NVIDIA nel 2016 accusa Qualcomm di un abuso di posizione dominante volto ad abbattere Icera. Secondo questa tesi, accolta nel 2019 dalla Commissione europea, Qualcomm ha offerto i suoi prodotti a un bassissimo prezzo a due aziende cinesi destinate a diventare molto più note, ZTE e Huawei, proprio per eliminare la minaccia di Icera.

Per Jensen la difficoltà di competere con Mediatek e Qualcomm porta alla decisione di lasciar perdere il mercato mobile. Quella sconfitta, nella storia riscritta da Jensen, diviene il sacrificio che ha forgiato NVIDIA una volta per tutte, nella rivendicazione della sua missione iniziale con l'apertura definitiva sull'intelligenza artificiale: la missione di NVIDIA è costruire computer in grado di risolvere problemi che i normali computer non possono risolvere. Dovremmo dedicarci a realizzare la nostra visione e a dare un contributo unico. La nostra ritirata strategica ha dato i suoi frutti. Lasciando il mercato della telefonia, abbiamo aperto le nostre menti per inventare un mercato nuovo. Abbiamo immaginato di creare un nuovo tipo di computer per computer robotici con processori di rete neurale, architetture di sicurezza che eseguono algoritmi di intelligenza artificiale. A quel tempo, questo era un mercato da zero miliardi di dollari.

Non c'è venditore migliore del ragazzo di Denny's, il quale avvisa gli ambiziosi studenti di Taiwan che ritirarsi non è facile per le persone brillanti, ma il vero successo sta nella capacità di rinunciare a qualcosa. Il riccio, alla fine della lezione, non chiede loro solo cosa sapranno inventare, dopo che la sua generazione ha inventato tutto, ma li mette davanti alla responsabilità di scegliere a cosa rinunciare. I mercati cominciano a considerare il nuovo ruolo giocato da SAMR (State Administration for Market Regulation), l'agenzia di livello ministeriale che opera sotto il Consiglio di stato della Repubblica Popolare Cinese, formata nel 2018 per razionalizzare un'attività che per Pechino – e non solo – ha conosciuto una significativa escalation politica. A marzo 2018 si verificano due eventi nelle retrovie della "guerra commerciale" che l'amministrazione Trump sta confusamente combattendo contro la Cina, dopo l'interessata difesa della globalizzazione economica da parte di Xi Jinping a Davos nel 2017.

Il primo è il blocco con ordine esecutivo del presidente Trump della transazione dei record, l'acquisizione di Qualcomm (al centro di diverse dispute legali, non solo con NVIDIA ma anche con Apple) da parte di Broadcom.

Cosa avviene? Un'azienda basata al momento a Singapore, ma con una significativa identità americana, ha tentato di acquisire un'azienda americana per 117 miliardi di dollari, senza poterlo fare per ragioni di sicurezza nazionale. La ragione profonda di queste decisioni, secondo il giudizio del CFIUS (Committee on Foreign Investment in the United States), è la paura che l'avanzamento cinese nel 5G divenga incontenibile. Il secondo evento è, appunto, la costituzione formale dell'agenzia cinese competente per decisioni relative all'antitrust, alla competizione, alla struttura dei mercati. L'agenzia, nel momento in cui gli Stati Uniti iniziano a erigere il muro della sicurezza nazionale, presidia lo spazio cinese e detta le regole di chi vuole stare all'interno di quel "piccolo cortile" che è la principale potenza manifatturiera mondiale.

Il messaggio, destinato a essere amplificato sempre di più, è il seguente: vendere in Cina non è gratis. Il potere di mercato di Pechino è il più importante fattore in mano alla Cina, anche nell'elettronica. Attualmente la SAMR è competente su operazioni di fusioni e acquisizioni di aziende con ricavi di almeno 117 milioni di dollari in Cina. Un potere enorme. Nell'estate del 2018 Qualcomm, non più acquisita da Broadcom, deve abbandonare l'acquisto da 44 miliardi dell'azienda dei semiconduttori dei Paesi Bassi erede delle attività Philips, Nxp, perché la SAMR non autorizza l'operazione. Entrambe le società hanno una vastissima esposizione al mercato cinese, nel caso di Qualcomm nel 2022 ancora superiore al 60% delle vendite, anche per via della struttura dell'assemblaggio dei componenti elettronici.

Anche prima dell'avvento della SAMR, le autorità cinesi hanno agito per anni contro Qualcomm sulle pratiche anticompetitive, prima che l'azienda raggiungesse un accordo nel 2015 pagando circa un miliardo di dollari di multa e portando a un abbassamento dei prezzi delle licenze sui brevetti in Cina, anche per aiutare aziende come Xiaomi e, ovviamente, Huawei. Questo è un altro lato del ritiro strategico di NVIDIA dal mobile: visti i due fattori dirimenti, cioè il ruolo centrale della Cina nell'assemblaggio di componenti elettronici e l'ambizione tradotta brevemente in potere di mercato delle aziende cinesi di smartphone e telecomunicazioni, l'azienda di Jensen, se avesse scommesso di più sul mobile, sarebbe stata più esposta ai conflitti tra Washington e Pechino.

L'acquisizione fondamentale di NVIDIA emerge nel 2019, un anno che per l'azienda si preannuncia difficile. Il titolo ha già subito una forte correzione per lo scoppio della bolla crypto, l'altro tema – oltre agli smartphone – che fa tremare le certezze di Jensen. Il 27 gennaio 2019, NVIDIA comunica ricavi ben sotto le aspettative, legati anche "al deterioramento delle condizioni macroeconomiche, soprattutto in Cina". Poche settimane dopo, l'11 marzo 2019, viene annunciata l'acquisizione di Mellanox. Da dove viene quest'azienda e perché è così importante?

Mellanox viene fondata nell'anno decisivo della GPU, il 1999, con sede a Sunnyvale, California, e a Yokneam, Israele, un'area vicina a poli universitari come il Technion e l'Università di Haifa, dove si trova un vibrante ecosistema di start-up. Il Jensen di Mellanox è Eyal Waldman, co-fondatore, presidente e amministratore delegato dell'azienda. Classe 1960, Waldman ha servito nella brigata Golani delle forze di difesa israeliane. Durante la guerra del Libano, deve restare in servizio per alcuni mesi prima di poter frequentare l'università, pertanto si iscrive al Technion, a un corso dove trova subito posto, ingegneria chimica, per poi spostarsi a informatica senza avere mai toccato un computer prima. Dopo i primi studi, lavora come ingegnere per uno dei principali fornitori israeliani di tecnologia militare, Elbit, e dopo essersi specializzato si trasferisce a Intel negli Stati Uniti.

La nascita di Mellanox porta Waldman a tornare a vivere in Israele. Tra gli investitori inziali emerge l'immancabile bollino di qualità di Sequoia Capital, ma c'è anche la stessa Intel, che peraltro conosce benissimo l'ecosistema israeliano, dove ha iniziato a operare nel 1974, ai tempi di Jensen a Oneida, Kentucky. Mellanox si specializza in prodotti e soluzioni di interconnessione di rete, essenziali per facilitare una trasmissione efficiente dei dati tra server, sistemi di archiviazione e infrastrutture di comunicazione all'interno dei data center. All'inizio del secolo dei data center, Mellanox attiva una gamma di prodotti con schede di interfaccia di rete (NIC o adattatori di rete), switch, router, cavi e software. I prodotti operano su due protocolli principali di interconnessione di rete: Ethernet e InfiniBand. Il primo è il protocollo più utilizzato per le interconnessioni di rete a livello globale, mentre InfiniBand rappresenta la vera novità di Mellanox, con capacità notevoli in termini di larghezza di banda e riduzione della latenza. Con InfiniBand, Mellanox fornisce soluzioni personalizzate per i vari clienti. Chi sono? I più importanti supercomputer al mondo, da Wuxi in Cina a Los Alamos negli Stati Uniti. I leader del cloud e delle telecomunicazioni usano Mellanox, per poi essere affiancati, nella seconda metà degli anni dieci, da start-up poco conosciute ma affamate di velocità di interconnessione come ByteDance e OpenAI.

2016, Waldman è abbastanza sicuro delle potenzialità di Mellanox nella nicchia dell'interconnessione e dichiara che si è lasciato ormai alle spalle il datore di lavoro e investitore iniziale, Intel. Per il gigante dei semiconduttori, i data center sono un segmento fondamentale, ma Mellanox è ormai senz'altro superiore nelle soluzioni di

interconnessione, in grado di migliorare le prestazioni con un'accelerazione superiore a quella della legge di Moore, verso la legge di Huang.

A inizio 2019, mentre NVIDIA annuncia i suoi risultati deludenti, Mellanox fa sapere di aver superato per la prima volta il miliardo di ricavi nell'anno precedente. Nell'asta per Mellanox sono coinvolte Intel, Xilinx, Broadcom. Pochi menzionano Jensen, che nel giro di poche settimane sorprende tutti grazie a una stretta collaborazione con Waldman, cresciuta nel corso di anni di soluzioni per i supercomputer e i data center, un'interconnessione di GPU dopo l'altra. Sempre che i cinesi siano d'accordo.

Nel 2012, quando viene annunciata la creazione di Mellanox Federal Systems, una divisione con sede in Virginia legata a tutte le agenzie governative degli Stati Uniti per gli integratori di sistemi per il governo federale, i clienti comprendono già il Dipartimento dell'Energia, la NASA, l'Esercito, la Marina, l'Aeronautica, l'FBI, diverse agenzie di intelligence, Northrop Grumman, Lockheed Martin, General Dynamics, Boeing, Raytheon. I prodotti di Mellanox ricevono le più elevate certificazioni di sicurezza dagli enti governativi degli Stati Uniti, come rivendicato dalla stessa azienda nelle comunicazioni al mercato. Quando gli hacker russi del gruppo Fancy Bear colpiscono i fornitori statunitensi della difesa con una campagna di phishing, cercano anche di entrare nell'account Gmail di un dipendente di Mellanox Federal Systems.

Waldman ha ricordato l'esperienza di essere stretto tra i due mercati e tra i due attori politici: da un lato gli Stati Uniti, in pubblico e in via confidenziale, facevano già pressione per evitare le vendite più sensibili (per esempio, a utenti militari) nel mercato cinese, e dall'altro lato i cinesi gli chiedevano come mai continuasse a rispondere a interessi politici statunitensi, essendo un'azienda israeliana votata al profitto. Waldman gestiva la situazione dicendo che il mercato degli Stati Uniti restava il più grande, e quindi – al di là della politica – doveva guardare anzitutto al suo principale cliente.

L'inizio del 2019 non è un momento come gli altri, perché il Dipartimento di Giustizia degli Stati Uniti annuncia a gennaio le accuse di frode finanziaria verso Meng Wanzhou, direttrice finanziaria di Huawei e figlia del fondatore dell'azienda, che si trova agli arresti domiciliari a Vancouver. È nel contesto in cui si accumulano queste tensioni, questi colpi e contraccolpi, che NVIDIA invia alla SAMR i materiali per approvare l'acquisizione di Mellanox il 24 aprile 2019. L'autorizzazione dell'agenzia cinese giunge dopo circa un anno e un duro negoziato dove la SAMR ottiene la protezione dei clienti cinesi, come appunto Alibaba e Huawei, anche attraverso alcune condizioni rimaste confidenziali.

La verità è che nessuno ha veramente capito la posta in gioco. Il campionato di chi ha capito meno di tutti è vinto dai regolatori europei, anch'essi chiamati ad autorizzare la transazione. Come abbiamo visto, nuove aziende ambiziose, tra cui ByteDance e OpenAI, sono andate da Mellanox negli anni precedenti per i suoi prodotti. In Europa aziende simili non esistono, perché in questa parte del mondo non succede niente di nuovo. Qualche data center viene realizzato da aziende tradizionali che investono in tecnologia (in Italia, per esempio l'Eni) e la Commissione europea promuove programmi sui supercomputer al servizio delle capacità scientifiche europee: tutto ciò è pieno di GPU di NVIDIA, la supply chain è in prevalenza americana e asiatica. Operare in questo contesto non aiuta gli attori europei a rispondere a due domande fondamentali: come sarà fatto il mercato dei "mulini satanici" nel breve-medio termine, e che effetti avrà l'acquisizione di Mellanox da parte di NVIDIA? Se non comprendi come è fatto quel mercato, se non ne conosci la struttura, allora non puoi capire le sue dinamiche competitive, tra cui c'è anche il vantaggio che NVIDIA, grazie a Mellanox, acquista nel fare il prezzo dei suoi servizi per i clienti.

Il divario tra chi costruisce il mondo e chi prova a valutarlo, in senso weberiano tra la scienza come professione e la politica come professione, non è mai stato così ampio. La Commissione europea, nel caso Mellanox, scrive che "NVIDIA è un attore molto piccolo". Ci tiene a precisare che "al momento, le vendite dei server DGX di NVIDIA sono piccole". Ciò non significa che i regolatori dovessero opporsi all'acquisizione di Mellanox. Significa solo che parlano a vanvera. Nella presentazione dei risultati del secondo trimestre 2023, il 23 agosto dello stesso anno, Jensen afferma: "Una nuova era del calcolo è cominciata. Le aziende in tutto il mondo stanno passando al calcolo accelerato e all'intelligenza artificiale generativa. Le GPU di NVIDIA connesse dalle nostre tecnologie di rete Mellanox che eseguono il nostro software sull'intelligenza artificiale CUDA costituiscono l'infrastruttura di calcolo dell'intelligenza artificiale generativa". Nel solo secondo trimestre, i ricavi sui data center di NVIDIA superano i 10 miliardi di dollari, +141% dal trimestre precedente, +171% dall'anno precedente. Risultati destinati a essere superati ancora dai due successivi aggiornamenti, il 21 novembre 2023 e il 21 febbraio 2024.

Nello stesso periodo, in Israele, si chiedono: com'è possibile che 93 + 7 faccia 1000? Com'è possibile che la capitalizzazione di NVIDIA, scesa appunto a 93 miliardi in quel momento difficile di inizio 2019, e il prezzo pagato per Mellanox abbiano condotto Jensen nel "club dei mille miliardi" (trillion dollar club)? Certo, gran parte del successo è dovuto all'esplosione dell'intelligenza artificiale, che Jensen ha agognato e preparato a lungo. Ma è vero che, proprio in quel momento, è avvenuta l'integrazione di Mellanox in NVIDIA, fornendo un eccezionale vantaggio competitivo. Anche questo passaggio ha richiesto un sacrificio: quello dello stesso Waldman, che ha deciso di lasciare l'azienda perché sa di non sapere e di non volere essere il numero due di nessuno. Allo stesso tempo, NVIDIA ha costruito una vera identità israeliana, aumentando i dipendenti che ora operano sotto la divisione NVIDIA Networking, coinvolgendo gli ingegneri israeliani nei progetti di frontiera e portando a diretto riporto di Jensen alcuni manager di Mellanox, tra cui il co-fondatore Michael Kagan, ora capo della tecnologia di NVIDIA. La giornata di Colette Kress, direttrice finanziaria di NVIDIA, inizia con gli incontri sincronizzati su Israele.

Nella sua impossibile pensione, popolata da costanti investimenti e progetti, Waldman continua a intervenire pubblicamente e a ripetere: perché dobbiamo ammazzarci a vicenda? Ricorda che per anni, d'accordo col governo israeliano, ha venduto un sacco di materiale a Saudi Aramco, anche prima dei cosiddetti "accordi di Abramo": gli arabi pagano, arricchiamoci a vicenda e basta. Il 7 ottobre 2023 Waldman si trova in vacanza in Indonesia, isolato da tutti. È l'anniversario dei controlli sulle esportazioni del 7 ottobre 2022. NVIDIA attende nuovi ostacoli di Washington alle vendite in Cina, per adattare i suoi prodotti a nuovi requisiti tecnici. Waldman deve tornare all'improvviso in Israele, perché nessuno riesce più a rintracciare sua figlia Danielle, che insieme al ragazzo Noam Shai, ex dipendente di NVIDIA, si trovava al festival musicale Supernova, a pochi chilometri dal confine con Gaza. Il fondatore di Mellanox insegue disperatamente il segnale di emergenza dell'iPhone della figlia, e con il suo seguito si trova davanti lo scenario di guerra seguito all'attacco dei terroristi: le auto crivellate dai colpi, i cadaveri della figlia e del suo ragazzo.

5. Il pappagallo stocastico e il momento iPhone

"Muoiono gli imperi, ma i teoremi di Euclide conservano eterna giovinezza". Nel suo saggio del 1913 L'uomo matematico, Robert Musil raffronta la matematica a un'economia del pensiero che consente di eseguire in poco tempo processi di cui non si verrebbe facilmente a capo. Il sogno del calcolo di Leibniz sta procedendo, generando vantaggi alla portata di tutti. Infatti, nota Musil, per risolvere i problemi non occorre ormai "farsi un viaggio dal signor Newton a Londra o dal signor Leibniz a Hannover": per chiunque, basta impostare il problema e "girare la manovella o qualcosa di simile". Almeno per quanto riguarda i quesiti che possono essere automatizzati, facilmente calcolati. "E anche nei casi naturalmente mille volte più numerosi di quesiti non risolvibili a macchina si può definire la matematica un apparato ideale dello spirito con lo scopo e il risultato di pensare preliminarmente in via di principio tutti i casi possibili."

Tutti i casi possono essere pensati. E di essi, grazie a essi, ci può essere un'impalcatura, un'"organizzazione spirituale". Gli umani, grazie alla matematica, non procedono più a tentoni. "Tutta la nostra civiltà è sorta col suo aiuto, non conosciamo altro mezzo." Eppure, in quest'esaltazione della civiltà matematica, Musil considera già l'altro lato della medaglia, "l'altro e autentico volto di questa scienza", un volto che "non è teso a uno scopo, ma invece diseconomico e appassionato". La civiltà sorta con l'aiuto della matematica è intimamente razionale, quindi rivolta a scopi ben precisi, sempre più efficace nel loro perseguimento, e per tali scopi si muove su fondamenti matematici.

Le figure della razionalità, compreso l'ingegnere, hanno però bisogno di una porzione ridotta della conoscenza matematica, quella che basta per affrontare i casi determinati che riguardano i loro compiti. Nella ricerca matematica, invece, c'è sempre un'eccedenza rispetto allo scopo, fino allo "spreco di ardimento della pura ragione, uno dei pochi che esistano oggi". In questo senso, il fondamento della razionalità rivolta a uno scopo si muove su tutt'altro piano e "abbraccia talune delle avventure più divertenti e intense dell'esistenza umana". C'è una ricerca costante, per salvare la potenza della matematica di non farsi trascinare da una catena di cause, costruendo un'esattezza che abbracci anche la creatività, le emozioni. Tutto questo potrà essere detto con un linguaggio dell'esattezza, senza essere abbandonato al turbinio dei casi e ai loro castelli in aria: "Non bisogna disarmare abbandonando l'esattezza a ingegneri e scienziati, ma avventurarsi nell'esperimento, proprio in analogia con la matematica, di dar forma a una paradossale combinazione di esattezza e indeterminatezza".

Andrej Karpathy, nato a Bratislava nel 1986, ama la matematica ma non la terra in cui cresce. Appena i genitori gli prospettano la possibilità di andarsene, è raggiante. Arriva a quindici anni in Canada, dove studia informatica nel posto giusto al momento giusto: l'Università di Toronto. Geoffrey Hinton gli parla delle reti neurali come se fossero una cosa viva, citando sempre il cervello e le sue analogie. Dopo il big bang di AlexNet, si sposta a San Francisco per

il dottorato all'Università di Stanford, dove studia con Fei-Fei Li. Karpathy non è solo un brillante studente e ricercatore della nuova disciplina che va affermandosi. È anche un docente e divulgatore eccezionale, che con le sue spiegazioni fa crescere l'attenzione per l'intelligenza artificiale a Stanford. Come nel 1979 migliaia di appassionati si sono recati in libreria per comprare Gödel, Escher, Bach, per leggere di temi apparentemente oscuri e sconnessi, centinaia di studenti affollano le classi di Fei-Fei Li e Andrej Karpathy: 150 nel 2015, 330 nel 2016, 750 nel 2017. I video di Karpathy generano sempre più visualizzazioni su YouTube e contribuiscono a far crescere la comunità interessata all'intelligenza artificiale.

Ilya Sutskever nasce nel 1986 nell'allora Unione Sovietica e a cinque anni emigra in Israele con i genitori e il fratello Noam. Vive a Gerusalemme fino all'età di sedici anni e, viste le sue doti precoci e la sua curiosità per le materie scientifiche, ha la possibilità di frequentare alcuni corsi alla Open University of Israel. Quando i Sutskever si trasferiscono in Canada, Ilya ha già sviluppato solide capacità di programmazione13 e viene ammesso all'Università di Toronto per studiare matematica. La fine della guerra fredda non è solo la riduzione dei contratti per i supercomputer di Danny Hillis e quindi la necessità di trovare nuove strade per praticare il calcolo parallelo, per aumentare la capacità di calcolo, per attaccare nuovi mercati: Jensen si inserisce in questo spazio decisivo, nell'intreccio tra la diffusione dei personal computer e le aspettative dei videogiocatori. Ma questi processi sono resi possibili dal fattore umano, dalla "grande trasformazione" della geografia dei talenti, dove processi già in corso divengono sempre più diffusi e inarrestabili, come una valanga che dalla scienza passa all'industria.

Il "momento unipolare" degli Stati Uniti, seppure di breve durata, ha una consistenza chiara nelle scelte di persone che decidono di lasciare la loro casa e mettersi alla prova, una parola dopo l'altra, con un'altra lingua e un altro mondo: di varcare l'Atlantico alla ricerca dell'America. Certo, l'ascesa dell'Asia orientale e della Cina in particolare cambiano per sempre il centro di gravità della manifattura, ma quello delle scelte dei talenti non si sposta. Il percorso di coloro che abbandonano la propria terra per il Canada non è comprensibile senza considerare l'ambizione di qualcuno che si inserisce in tutti questi intrecci. Un ragazzo arrivato a Montreal dal Sudafrica, poco prima di compiere diciotto anni: Elon Musk.

Sono figure che non possiedono solo ampie risorse da poter scommettere in progetti di lunghissimo termine, ma vedono nell'intelligenza artificiale un oggetto di studio stimolante, un modo per lasciare la loro impronta nella storia. Oltre a Peter Thiel, è chiaro che Elon Musk risponde alla perfezione a questo identikit. Com'è noto, Musk investe all'inizio del secolo la fortuna ottenuta da PayPal soprattutto su due imprese: SpaceX, per rivoluzionare l'economia spaziale e rendere l'umanità una specie multiplanetaria, e Tesla, per rivoluzionare la mobilità in senso sostenibile e trasformare l'industria automobilistica. Mentre Musk persegue queste due linee parallele di cambiamento radicale – e di indubbio successo, visto che SpaceX cambia, come mai nessun'altra impresa nella storia, il rapporto pubblico-privato nello spazio, e che Tesla supera i pregiudizi dell'industria automobilistica raggiungendo un'ampia capacità di mercato oltre a un'elevatissima capitalizzazione di borsa – porta avanti anche altri progetti, che rientrano, in modo più o meno laterale, nella sua idea di innovazione. Alcuni di essi, come HyperLoop, Neuralink o Solarcity, non hanno avuto lo stesso successo, ma ciò che conta è la possibilità e volontà di Musk di effettuare numerose scommesse più o meno futuristiche.

Nel 2014, oltre a Internet, Musk identifica quattro aree principali su cui agire: le prime due sono, appunto, la transizione alla produzione e al consumo sostenibile di energia e l'estensione della vita umana ad altri pianeti. Le altre due aree sono la lettura e la scrittura del codice genetico e l'intelligenza artificiale. Musk conclude quell'intervento dicendo: "Spero che l'intelligenza artificiale sia buona con noi". In quel periodo, è già avvenuto un famoso incontro tra Demis e Musk del 2012, di cui esistono numerose versioni. La conversazione tra i due avviene nella sede di SpaceX, che sta portando avanti la collaborazione con la NASA e iniziando i test sulla riusabilità, la grande idea ingegneristica e commerciale di Musk in cui all'inizio non crede nessuno. Eppure, l'imprenditore nato in Sudafrica, nella sua conversazione col maestro dei giochi di DeepMind, vede già l'approdo finale nei componenti che li circondano: Marte, il sistema solare, la colonizzazione degli altri pianeti.

Andare altrove è la strada per preservare la "coscienza", il dono che gli uomini portano con sé, dai rischi esistenziali in questa nostra fragile Terra, soprattutto dai rischi che creiamo da soli. Tuttavia, Demis osserva che in questo cammino "l'umanità" non sarà sola né può pretendere di esserlo: al suo fianco avanzerà l'intelligenza artificiale, e questo può da un lato accrescere enormemente la possibilità di riuscita del progetto di Musk e dall'altro aumentare quegli stessi rischi esistenziali. Dipende da come sarà progettata l'intelligenza artificiale, dalle potenzialità che saprà realizzare, nonché dalla nostra capacità di monitorare questo processo. Implicito, nel ragionamento di Demis, è che la migliore posizione in questa partita sia dentro DeepMind.

Tesla, società quotata dal 2010, si muove nell'affollato mercato automobilistico, dove la Model S ha un effetto significativo. Nel 2011, Musk non si preoccupa per quell'azienda cinese di auto elettriche e batterie in cui ha investito Warren Buffett, BYD (Build Your Dreams): non trattiene le risate quando una conduttrice televisiva la cita come concorrente di Tesla, che nel 2013 ripaga con anticipo il prestito ricevuto dal governo statunitense nel 2009. SpaceX, con i test sulla riusabilità del vettore e con i progetti sulle costellazioni di satelliti in orbita bassa, si appresta a costruire in modo massiccio un mercato che non è mai esistito prima, surclassando le aziende dell'ecosistema spaziale rimaste per decenni attaccate alle mammelle delle commesse pubbliche.

Google è vicina ad acquisire Tesla, che da ultimo Musk non vuole vendere, perché col successo della Model S vede lo sviluppo di quel mercato. Musk e Page non concordano sul futuro dell'intelligenza artificiale. Per il primo, la scommessa di DeepMind serve a garantire il controllo umano delle macchine, perché solo l'umanità è dotata della luce della "coscienza" che si deve espandere per l'universo. Per il secondo, Musk è uno "specista" che non vuole accettare che l'umanità potrebbe essere solo uno stadio dell'evoluzione, che sarà superato da qualcos'altro – auspicabilmente da una sua creazione. In mezzo a questo contrasto, dopo il big bang dell'intelligenza artificiale con AlexNet, avviene la trattativa di Google per l'acquisto di DeepMind.

In seguito, nel discutere del ruolo delle aziende, Charlie Rose fa a Page una domanda che non sospetta essere imbarazzante, perché legata alla questione di DeepMind: "Una volta hai detto, se non sbaglio, che invece di lasciare i tuoi soldi per una causa generica, preferiresti darli semplicemente a Elon Musk, perché hai fiducia che possa cambiare il futuro". Page specifica che Musk vuole "andare su Marte e difendere l'umanità". Il percorso di SpaceX mostra che si può essere un'azienda e avere allo stesso tempo un fine filantropico. Ne consegue, secondo Page, che il dibattito sulla "cattiveria" delle aziende è sbagliato, perché le grandi aziende del nostro tempo, potendo finanziare le grandi sfide esistenziali, sono in grado di produrre il maggior "bene" possibile.

La risposta di Page cela, in quel momento, il suo aspro contrasto con Musk sull'intelligenza artificiale, che colpisce anche la loro amicizia. Nel 2014, Musk diviene, dopo Stephen Hawking, la voce più influente tra coloro che sottolineano i pericoli dell'intelligenza artificiale. Il fisico di fama mondiale, succeduto da sir Lighthill alla cattedra di matematica a Cambridge, dopo aver ricordato l'evidente contributo della tecnologia per affrontare la sua stessa disabilità, confessa alla BBC i suoi timori: "Le forme primitive di intelligenza artificiale che abbiamo già si sono dimostrate molto utili. Ma penso che lo sviluppo di una piena intelligenza artificiale possa portare alla fine della razza umana. Una volta che gli umani svilupperanno quest'intelligenza artificiale, decollerà da sola e si ridisegnerà a un tasso sempre più elevato. Gli umani, che sono limitati da una lenta evoluzione biologica, non potranno competere e saranno sorpassati".

Da par suo, Elon Musk interviene in un dibattito del sito Edge.org con un commento – poi cancellato, ma mai smentito – dove afferma: "Il ritmo del progresso nell'intelligenza artificiale (e non mi riferisco all'intelligenza artificiale ristretta) è incredibilmente veloce. A meno che non si abbia un'esposizione diretta a gruppi come DeepMind, non si ha idea di quanto velocemente stia crescendo, a un ritmo vicino a quello esponenziale. Il rischio che accada qualcosa di seriamente pericoloso può essere stimato a cinque anni, dieci al massimo". Nella formula che Musk utilizza al MIT, l'intelligenza artificiale è "l'evocazione del demone". Qualunque cosa sia, rischia di non finire bene.

DeepMind interviene per cambiare la narrazione, nel 2017, con la creazione di una nuova unità di ricerca, DeepMind Ethics & Society, volta ad "aiutare i tecnologi a mettere l'etica in pratica, e ad aiutare la società ad anticipare e dirigere l'impatto dell'intelligenza artificiale in modo che vada a beneficio di tutti". L'unità vuole fornire ricerche disponibili a tutti, nello spirito delle altre attività di DeepMind, incentrate sulla diffusione e la pubblicazione, mentre il board etico, secondo Mustafa Suleyman, è incentrato sull'intelligenza artificiale generale, "che è sempre stata di lungo termine, oltre dieci, venti, trent'anni, mentre costruiamo sistemi che sono sempre più autonomi, davvero capaci di svolgere funzioni realmente umane

Secondo il racconto del "New York Times", il famigerato board etico di DeepMind si è effettivamente riunito solo una volta, il 14 agosto 2015, in una stanza con vista sulla fabbrica di SpaceX, che ha ormai lanciato ufficialmente Starlink e intensifica i suoi test per la riusabilità, tra un fallimento e l'altro. Poche settimane prima, Suleyman interviene a Londra a una conferenza di ricercatori di machine learning e afferma tra gli applausi dell'uditorio: "Qualsiasi discorso su una macchina superintelligente che aspira a tutta la conoscenza del mondo e poi prende le proprie decisioni è assurdo. Ci sono ingegneri in questa stanza che sanno quanto sia difficile inserire input in questi sistemi"

L'intelligenza artificiale è una sfida di lungo termine, ma anche un enorme problema organizzativo di breve termine. È facile parlare di Bell Labs e spalmare qua e là risorse per far giocare Kurzweil e Demis; nel mentre, però, bisogna rispondere al mercato e affrontare sfide sociali e politiche, come le accuse di discriminazioni che vengono da alcuni dipendenti e il tema del rapporto col governo degli Stati Uniti sulla sicurezza nazionale.

Tutti questi aspetti vengono accentuati dalla mossa di Musk, che il suo carattere rende scontata. Nel 2015, fonda i "suoi" Bell Labs sull'intelligenza artificiale: OpenAI. Nel momento della fondazione, a fine 2015, sono un'istituzione di ricerca non profit che vuole "costruire valore per tutti invece che per gli azionisti". Ad affiancare Musk nell'impresa è anzitutto Sam Altman, imprenditore e investitore al tempo trentenne, con una notevole esperienza nel principale acceleratore di start-up, Y Combinator, che ha ricevuto il supporto della leggendaria Sequoia Capital. OpenAI annuncia di aver ottenuto capitali per un miliardo di dollari da parte di Musk, Altman, il nuovo direttore delle tecnologie Greg Brockman, Jessica Livingston di Y Combinator, YC Research, i "soliti" Reid Hoffman e Peter Thiel, e anche aziende come Amazon Web Services e Infosys. Nel nucleo dei fondatori, Musk e Altman sono riusciti ad assestare la vera botta nei confronti di Google, rubando al gigante Ilya Sutskever, che diventa direttore delle ricerche di OpenAI. Pochi mesi dopo, nell'estate del 2016, l'intelligenza artificiale è sulla bocca di tutti grazie al successo di AlphaGo.

Jensen è pronto ad approfittarne, finalmente. Eppure, il suo investimento di lungo termine nell'hardware per l'intelligenza artificiale sembra essere messo in discussione da Intel, che vuole presidiare il suo storico dominio nel mercato dei data center e annuncia un nuovo prodotto con prestazioni migliori di quelle di NVIDIA; a Ian Buck è riservato il ruolo di contestare tecnicamente le affermazioni dei concorrenti, pochi giorni prima della conferenza degli sviluppatori di Intel a San Francisco, ad agosto 2016, Jensen si presenta di persona negli uffici di OpenAI per consegnare direttamente a Elon Musk il primo supercomputer DGX-1. I tre (i due umani e il computer siglato NVIDIA) sono immortalati in una fotografia. NVIDIA e OpenAI, secondo Jensen, sono uniti nell'obiettivo di "democratizzare l'intelligenza artificiale". Uno studente nato a Shanghai, Linxi "Jim" Fan, dottorando di Fei-Fei a Stanford, è il primo stagista della storia di OpenAI: in quell'occasione incontra Jensen, che lo assumerà come ricercatore sulla robotica e i giochi. Anche lui mette la firma nel DGX-1, dove Jensen ha scritto: "A Elon e al team di OpenAI. Per il futuro del calcolo e dell'umanità vi presento il primo DGX-1".

Nell'estate 2016, suscita una certa ilarità il fatto che l'addestramento dei modelli, per cui serve il DGX-1 di NVIDIA consegnato a OpenAI, avvenga attraverso l'utilizzo dei messaggi di Reddit, il sito che raccoglie opinioni degli utenti su qualunque cosa. Le idiozie su argomenti a piacere buttate giù su un forum dovrebbero migliorare la comprensione probabilistica delle conversazioni nella produzione di testo. Come no. Ma Andrej Karpathy e Ilya Sutskever non stanno scherzando. La fotografia del 2016 riassume al meglio il breve regno di Musk in OpenAI. Anche in seguito, ha sempre sostenuto che il cuore del progetto è stato "la democratizzazione del potere dell'intelligenza artificiale", "ridurre la probabilità che l'intelligenza artificiale venga monopolizzata". OpenAI rappresenta la risposta necessaria alla "fortissima concentrazione di potere sull'intelligenza artificiale, soprattutto a Google-DeepMind". Quando fa queste affermazioni, a fine 2018, Musk ha già lasciato OpenAI: a febbraio è stato annunciato che continuerà a "donare e a consigliare l'organizzazione" ma non è vero. Ancora una volta, l'investimento – psicologico e pratico – di Musk sull'intelligenza artificiale si lega al percorso delle sue aziende.

Nel 2018, SpaceX ha ormai raggiunto il suo obiettivo storico, la riusabilità del lanciatore, e Tesla ha inaugurato il suo primo veicolo di massa, la Model 3. Musk ha cambiato idea rispetto all'indipendenza di un'organizzazione di ricerca sull'intelligenza artificiale. Il motivo è semplice: ora lui dispone di una piattaforma, Tesla, in cui le tecnologie possono essere sperimentate su vasta scala, attraverso dati e prodotti con una dimensione commerciale per l'enorme mercato automobilistico; non ha più senso tenere le due dimensioni separate. Il futuro di OpenAI è TeslaAI, ovvero Autopilot: un sistema di assistenza alla guida che indica l'orizzonte sempre prossimo della guida autonoma.

In America è davvero possibile costruire qualcosa, la manifattura avanzata non è destinata esclusivamente e incontrovertibilmente al dominio asiatico. Certo, la produttività delle fabbriche americane non è pari a quella delle fabbriche cinesi, ed è improbabile che Musk accetti la sindacalizzazione dei suoi dipendenti: non a caso, prima nei paesi scandinavi e poi altrove, combatte e combatterà contro i resti dell'equilibrio tra capitale e lavoro. Musk, in questo senso, si muove nel solco dell'ingegnere per eccellenza della nostra epoca, Morris Chang, perché è ormai strutturalmente impossibile che il lavoro abbia la forza di opporsi al capitale, ma ciò non toglie la capacità di Musk di dimostrare effettivamente che in America si può essere "costruttori". Questa presenza concreta, fatta di spazi e fornitori, di prototipi e di prodotti che devono rispondere a precisi standard di sicurezza per stare sul mercato, può fornire un vantaggio significativo per l'applicazione dell'intelligenza artificiale alla robotica, che non è solo la

possibile ondata successiva ai grandi modelli linguistici, ma è anche il sogno più classico dell'intelligenza artificiale, percorso senza successo da Marvin Minsky e molti altri pionieri.

OpenAI ha dovuto comunque compiere il suo "sacrificio", l'abbandono della robotica nel 2021. Per portare le simulazioni in un ambiente adeguato, c'è bisogno di lavorare a stretto contatto con una fabbrica. Ed è quello che Tesla potrà sempre fare, in modo più efficace rispetto ad altri attori, mentre l'allargamento del mercato dei veicoli elettrici potrà garantire risorse di ricerca e sviluppo da reinvestire in modo funzionale. Tesla come vincitrice di questo processo per cui non viene più considerata un'azienda automobilistica, ma "la più grande armata commerciale di robot che raccolgono una quantità monumentale di dati dalle nostre strade per l'elaborazione e l'apprendimento automatico".

Non è un caso che Musk ripeta che Tesla vada considerata non solo un'azienda automobilistica, ma un leader della robotica e dell'intelligenza artificiale. Lo dice sia per coprire le difficoltà nel mercato automobilistico, dove svanisce progressivamente l'effetto sorpresa di Tesla ed emerge sempre più la concorrenza cinese, sia perché vede veramente uno sbocco più ampio. Le strade di Musk e di OpenAI si separano per due ragioni. In primo luogo, il concetto di "democratizzazione dell'intelligenza artificiale" cambia in base a chi lo impugna, quindi Musk, a seconda della posizione relativa e della forza di Tesla, diviene Davide oppure Golia. In secondo luogo, l'idea dell'impresa di ricerca non profit deve comunque confrontarsi con una delle forze essenziali per la crescita delle potenzialità dell'intelligenza artificiale, ovvero la crescita continua della capacità di calcolo, ovvero la necessità di pagare Jensen e i suoi emuli: senza un aiuto fondamentale nella corsa dell'hardware, che richiede risorse superiori a "donazioni" pur molto cospicue, non è possibile competere. Per questo OpenAI, senza più Musk, non può che bussare alla porta di uno degli altri giganti: Amazon, che pur essendo tra i donatori iniziali perde quest'occasione storica nel 2018, e Microsoft, con cui si concretizza nel 2019 l'accordo decisivo che cambia la natura stessa di OpenAI, da impresa non profit a impresa che deve continuare a raccontarsi come non profit ma è costretta, ovviamente, a garantire ritorni dell'investimento a Microsoft, che rende possibile il suo sviluppo infrastrutturale.

L'accelerazione di Ilya Sutskever è evidente da un confronto tra la sua tesi di master del 2007 e la tesi di dottorato del 2013. La prima si apre con l'indicazione di quattro compiti: la conversione di una nota scritta a mano digitalizzata in un'immagine coerente; l'individuazione dell'identità e delle posizioni delle persone in una fotografia; la determinazione delle parole in una registrazione audio; la determinazione del contenuto di una mail come spam. Questi compiti, scrive Sutskever, sono svolti quasi senza sforzo dagli esseri umani mentre, con le tecniche tradizionali di programmazione non possono essere facilmente risolti. C'è un enorme divario tra la facilità dell'esecuzione umana e le complicate modalità di addestramento affinché i programmi possano giungere a risultati simili, ma comunque insoddisfacenti. Basandosi soprattutto sul lavoro di Hinton, e in particolare sulle macchine di Boltzmann, nonché sulla letteratura esistente, Sutskever cerca di introdurre alcuni modelli probabilistici con algoritmi efficienti di apprendimento per la soluzione dei compiti specifici, verso un piano di lavoro più ambizioso che conduca i compiti in una sequenza temporale, e quindi renda possibile per il modello "avere una sorta di memoria più di lungo termine, simile a quella del modello Long Short-Term Memory, o connessioni che cambiano durante le operazioni della rete e imparano ad archiviare importanti informazioni contestuali per un ampio numero di passaggi".

La tesi di dottorato del 2013 è diversa perché Sutskever stesso è diventato letteratura. Oltre a ribadire i suoi ringraziamenti ai genitori, "che hanno affrontato due immigrazioni per me e mio fratello", inserisce all'inizio della tesi un sommario di riferimenti a venti lavori che ha già pubblicato, tra cui il paper di AlexNet del 2012. La tesi si concentra sul miglioramento delle reti neurali ricorrenti, la classe di reti neurali artificiali progettate per riconoscere modelli in sequenze di dati, come serie temporali, testi o altre tipologie di dati sequenziali. Le reti neurali ricorrenti hanno connessioni cicliche che permettono l'elaborazione di input precedenti: questa struttura permette alle informazioni di essere "trattenute" per un certo periodo di tempo.

Il lavoro mostra "che le reti neurali ricorrenti sono ben più facili da addestrare di quanto ritenuto in precedenza", stimolando così gli studi successivi anche grazie alla costante introduzione di GPU più potenti, che rendono l'addestramento più veloce e praticabile anche per reti complesse e di grandi dimensioni. Nel 2015, Sutskever viene invitato da Sam Altman all'incontro con Elon Musk, Greg Brockman e altri investitori che rappresenta l'inizio di OpenAI. Accetta di abbandonare Google, attratto dal bonus iniziale di quasi due milioni di dollari e dall'idea di una società di ricerca non profit dove gli scienziati lavorano fuori dall'ombrello di una grande azienda coi suoi interessi – al contrario di DeepMind – e allo stesso tempo sono liberi di interagire con gli ingegneri per far crescere il loro campo in modo esponenziale. "I ricercatori sono addestrati a pensare in piccolo," afferma quando ritorna su quella scelta, lasciando intendere la necessità di pensare in grande. Il decennio dell'accelerazione dell'intelligenza

artificiale è anche “una fuga di cervelli senza precedenti dall’accademia all’industria”. Certo, OpenAI è un’industria sui generis, almeno all’inizio. Ma competere è nella sua natura. Dovrà produrre e diffondere conoscenza, sfidando DeepMind per la corona dei Bell Labs dell’intelligenza artificiale, e allo stesso tempo dovrà lanciare prodotti in grado di mostrare al mondo le capacità pratiche degli scienziati.

Per avanzare una “missione” sull’intelligenza artificiale con risultati superiori a quelli accademici sono necessari allo stesso tempo un gruppo talentuoso e ambizioso (e OpenAI riesce ad attrarre notevoli competenze), infrastrutture e capacità computazionale. Quest’ultimo punto ha a che fare ovviamente con le GPU, con la consegna simbolica dei prodotti da parte di Jensen alla presenza di Elon Musk e con molte altre consegne fuori dai riflettori. Per alimentare questo processo, i ricercatori e gli ingegneri che vogliono diventare imprenditori apprendono una lezione semplice: servono sempre più soldi. E se Google può permettersi di finanziare alcune soluzioni e alcuni prodotti che può usare internamente per il miglioramento delle sue fonti di profitto, lo stesso gioco laterale non può essere praticato da OpenAI, che ha bisogno di clienti per i suoi prodotti.

All’inizio del 2017, nell’apprendimento supervisionato, “il successo è garantito, se si hanno una rete abbastanza ampia, un dataset per l’addestramento abbastanza grande, e abbastanza soldi per le GPU”. L’accelerazione generalizzata della ricerca, la crescita della comunità che scrive paper e li diffonde attraverso arXiv, l’uscita dal cono d’ombra delle conferenze specializzate (sempre più partecipate) e le risorse messe a disposizione dalle aziende si intrecciano con le svolte concettuali, come i Transformer. Giugno 2018, OpenAI lancia il primo GPT (Generative Pre-trained Transformer), una serie di modelli che compiono progressi nell’elaborazione del linguaggio naturale. Si tratta di modelli sempre più grandi, con maggiori parametri, un insieme di dati di apprendimento più vasto, e tecniche sempre più accurate per giungere all’obiettivo, che è la generazione di testo sulla base di una specifica richiesta.

GPT-1 ha 117 milioni di parametri e viene addestrato utilizzando il “BooksCorpus”, un dataset consistente in testi estratti da libri, che gli fornisce una varietà di stili e argomenti di scrittura. GPT-2, introdotto nel novembre 2019, amplia il numero di parametri a 1,5 miliardi e viene addestrato su un dataset molto più ampio, che include una parte di Common Crawl, un archivio web di dati pubblici. GPT-3, lanciato nell’estate 2020, ha 175 miliardi di parametri. Oltre a un insieme di dati ancora più vasto, che include Common Crawl, libri, Wikipedia e altri testi, usa tecniche di addestramento innovative, come il cosiddetto apprendimento per rinforzo da umano, sviluppato da OpenAI. Ognuno di questi passaggi ha richiesto un numero sempre più ampio di GPU. Il 19 maggio 2020, Microsoft annuncia di aver sviluppato per OpenAI un supercomputer con 285.000 processori tradizionali (CPU cores), 10.000 GPU e le interconnessioni di Mellanox-NVIDIA. La sede fisica di questi “mulini satanici” non viene diffusa. L’invisibilità, oltre che per proteggersi dalle critiche sullo sfruttamento ambientale, serve per ragioni competitive e di sicurezza. Solo in seguito Microsoft svelerà che l’infrastruttura per l’intelligenza artificiale, inserita in Azure, fa parte dei suoi investimenti a West Des Moines, in Iowa, dove in mezzo ai campi di grano c’è un impressionante cluster di data center in continua crescita, con sempre maggiore sete di acqua ed elettricità per il suo funzionamento. I posti di lavoro (operai, elettricisti, tecnici) sono concentrati nella costruzione dei data center, non nella loro gestione.

9 novembre 2018, Ilya tiene la conferenza “Progress towards the OpenAI Mission”. La missione è ormai cambiata rispetto all’annuncio iniziale del 2015. Elon Musk non fa più parte del progetto e Altman sta discutendo la cooperazione con Microsoft. La missione, ricorda Ilya riferendosi allo statuto di OpenAI, è “assicurare che l’intelligenza artificiale generale, definita come sistemi autonomi in grado di superare gli umani nella maggior parte dei lavori con un valore economico, vada a beneficio di tutta l’umanità”. Sul crinale tra la ricerca e l’impresa, questa definizione è interessante perché lega immediatamente il risultato fondamentale, l’idea dell’intelligenza artificiale generale, al lavoro e al valore economico. In questo modo, OpenAI cerca di tradurre la sua ambizione scientifica, ingegneristica e teorica in un prodotto, un mega-prodotto, perché il suo lavoro, superando il lavoro umano, fornisce un valore economico tendente all’infinito.

Negli ultimi sei anni la quantità di capacità computazionale utilizzata dai più grandi esperimenti di reti neurali è aumentata di un fattore di 300.000. Questo progresso, alimentato dal calcolo parallelo, è molto di più veloce di quello a cui eravamo abituati con la legge di Moore. Con l’intelligenza artificiale si possono fare cose fantastiche, cose incredibili: automatizzare la salute, renderla mille volte più economica e mille volte migliore, curare molte malattie, risolvere il riscaldamento globale”. La modalità con cui una “cosa fantastica” diviene “mille volte più economica” non viene spiegata. Del resto, “l’intelligenza artificiale generale non è un termine scientifico. Identifica una soglia, un punto di riferimento. È l’idea, è il punto in cui l’intelligenza artificiale è così intelligente che, se una

persona può svolgere alcuni compiti, anche l'intelligenza artificiale può farlo. A quel punto, puoi dire di avere l'intelligenza artificiale generale".

Il 30 novembre 2022, a uscire dalla caverna dell'apprendimento delle macchine è il prodotto di OpenAI chiamato ChatGPT, introdotto da poche righe di presentazione: "Abbiamo addestrato un modello chiamato ChatGPT che interagisce attraverso conversazioni. Il formato del dialogo consente a ChatGPT di rispondere a domande di follow up, ammettere i propri errori, contestare premesse errate e rifiutare richieste inappropriate". Il modello è reso disponibile gratuitamente, mentre a febbraio 2023 viene introdotta una versione a pagamento. ChatGPT ottiene una crescita di utenti impressionante dal momento del lancio. Secondo Greg Brockman, in appena cinque giorni si supera il milione di utenti, mentre il prodotto precedente, GPT-3, aveva avuto bisogno di due anni. Dopo soli due mesi, ChatGPT è la prima applicazione della storia a raggiungere cento milioni di utenti e le sue capacità sono al centro del discorso pubblico a inizio 2023, con una diffusione e un'influenza culturale che nessun altro prodotto dell'intelligenza artificiale è riuscito a conseguire in precedenza.

Secondo una felice definizione di Andrej Karpathy, i grandi modelli linguistici sono "due file". Uno è il file dei parametri e l'altro è il file dell'esecuzione dei parametri. Il modello statistico viene formato raccogliendo e analizzando grandi quantità di dati, che vengono utilizzate per una fase di addestramento resa possibile dall'uso di cluster di GPU per diversi giorni. Il modello funziona prevedendo la parola successiva in una frase, in modo da rispondere agli input ricevuti in modo efficace. Come ha ricordato anche Mira Murati, la previsione della parola successiva ha radici nell'esperimento del matematico russo Markov all'inizio del Novecento con le prime ventimila lettere dell'Eugenio Onegin di Puškin. Markov analizza la frequenza di vocali e consonanti per mostrare come anche in un testo letterario esistano percorsi statistici descrivibili in termini probabilistici. Queste cosiddette "catene di Markov" sono processi stocastici (dal greco stochastikós, congetturale), ovvero rientrano tra strumenti, procedimenti, teorie, modelli che descrivono situazioni che variano in base alla probabilità.

Ai tempi di Markov, ricorda Murati, sarebbe stato impossibile ottenere una mappatura dettagliata di tutte le lettere e le rispettive frequenze in relazione al resto del testo in combinazioni di due e tre lettere. "Oggi le macchine rispondono a queste domande in un istante, motivo per cui vediamo così tante applicazioni interfacciarsi attraverso il linguaggio conversazionale. Invece di prevedere la lettera successiva, GPT-3 prevede quale parola viene dopo rivedendo il testo che la precede." Un modello linguistico, per essere più efficace e utile per una varietà di compiti, viene affinato attraverso istruzioni umane. Come può avvenire, concretamente? OpenAI, anche grazie alle risorse ottenute da Microsoft, ha potuto pagare una serie di aziende che hanno lavorato a cottimo per affinare i risultati del modello, sia dal punto di vista del miglioramento della qualità dei testi, sia dal punto di vista della rimozione di alcuni contenuti sensibili e problematici. Per esempio, OpenAI tra il 2021 e il 2022 si è affidata a un'azienda, Sama, che ha assunto lavoratori kenyani a meno di 2 euro l'ora per svolgere questi compiti. Si tratta, in certo modo, di una permanenza dell'intelligenza artificiale artificiale, secondo il paradigma del Turco meccanico di Amazon.

Bill Dally di NVIDIA, quando deve spiegare il funzionamento di ChatGPT nelle conferenze in cui racconta lo sviluppo dell'hardware, parla di un processo di "distillazione". L'uso di questo termine deriva dalle origini dell'informatico: Vengo dalle colline del Tennessee orientale, ho grande familiarità con la distillazione. Quello che facciamo è prendere un sacco di mais, inserirlo in un punto e da un'altra parte far uscire il whisky. È la stessa cosa. Prendiamo un sacco di dati, un miliardo di miliardi di parole o anche di più, suddivisi in token, che sono circa due per ogni parola, e addestriamo un modello generale. Sapete, è come mandare vostro figlio a prendere una laurea umanistica: in un certo senso, conosce un po' di tutto. Il modello è stato addestrato su un enorme corpus di tutto ciò che si può assorbire da Internet. Ci sono cose che forse preferiremmo non sapere, perché potrebbero portare a risposte sgradevoli. Per dare valore a questi modelli, allora, li ottimizziamo con dati speciali per applicazioni particolari, per la programmazione, la consulenza medica, la formazione, la scrittura, quello che vi pare, e poi il modello diventa migliore in compiti particolari. Cosa stiamo facendo? Stiamo distillando i dati, e infine eseguiamo qualcosa, un'inferenza, una query, e otteniamo qualcos'altro.

Altri non hanno in mente il whisky. Una definizione influente e significativa per modelli linguistici come i prodotti di OpenAI è stata proposta nel 2021 da alcune ricercatrici, tra cui la linguista Emily Bender e l'allieva di Fei-Fei Li ed ex informatica di Google Timnit Gebru. È il concetto di "pappagallo stocastico", che si inserisce in una critica articolata rivolta ai modelli. La principale preoccupazione espressa dalle ricercatrici riguarda la loro tendenza a riprodurre e amplificare i pregiudizi e le visioni del mondo presenti nei dati di addestramento. Il riflesso del mondo da parte dei modelli è quindi viziato dall'origine. Questo può portare a risultati distorti e pericolosi, soprattutto quando i modelli vengono utilizzati in applicazioni che influenzano la vita reale delle persone.

I modelli sono appunto “pappagalli stocastici”: mettono una parola dopo l’altra, per effetto della loro arte probabilistica. Anche se le risposte generate sono fluenti e coerenti, esse sono prive di una vera comprensione o intenzione comunicativa. I modelli, appunto, si limitano a riprodurre modelli di linguaggio basati sulla probabilità, derivati da enormi set di dati di addestramento, senza alcuna consapevolezza del contesto. Diventa essenziale, allora, comprendere la profonda impotenza di questo lato “ambientale” della critica all’intelligenza artificiale, che da ultimo risulta caricaturale.

Del resto, è già possibile scrivere un dialogo con ChatGPT tra Jensen e Greta Thunberg in cui il ragazzo di Oneida, Kentucky, spiega all’attivista che la corsa delle GPU è in realtà una corsa dell’efficienza energetica, che NVIDIA, l’azienda verde sempre più green, ha già ridotto le emissioni di gas serra, usato le rinnovabili per i data center, tagliato i consumi del raffreddamento. Tutto viene documentato attraverso parametri precisi. L’attivista può presentarsi all’incontro armata del rapporto dell’Agenzia Internazionale dell’Energia: ci sono più di ottomila data center al mondo, di cui il 33% negli Stati Uniti, il 16% in Europa, il 10% in Cina; il consumo di elettricità è significativo, col caso limite dell’Irlanda in cui è stimato che un terzo della domanda nel 2026 verrà dai data center. Greta potrà dire: per colpa tua, Jensen, l’intelligenza artificiale consumerà nel 2026 il decuplo dell’elettricità del 2023. Appena Greta dice parole come ambiente, rinnovabili, energia, il generatore di testo di Jensen cita una miriade di aziende e centri di ricerca che si affidano a NVIDIA per rendere il mondo più verde con l’azienda che è nata e cresciuta col colore verde. La sua missione è proprio sostituire i “mulini satanici”, coi loro processori inadeguati, con sistemi di raffreddamento inadeguati, con infrastrutture più efficienti e sostenibili: le sue. Greta gli risponde che non vuole sentire il blablabla, lui ribatte che si tratta di progetti concreti, che migliorano le vite delle persone e, per evitare le accuse di “specismo”, anche dei rinoceronti e dei pangolini in pericolo, letteralmente salvati dalle GPU, come può essere dimostrato more geometrico. Del resto, immagazzinare e trasmettere energia, come scoprire nuove soluzioni – fino alla fusione nucleare –, è sempre un calcolo, e bisogna fare presto, bisogna accelerare proprio perché il pianeta non aspetta. “L’alba dell’intelligenza artificiale è arrivata e il calcolo accelerato è il calcolo sostenibile.” Se l’attivista chiede urgenza, Jensen risponde che l’accelerazione è urgenza. La partita è già finita prima ancora di cominciare, anche se nei prossimi mesi e anni vedremo un’infinita letteratura sull’elettricità e i data center, alimentata dallo spauracchio di un mondo trasformato in “mulino satanico”.

Microsoft, l’equivalente tecnologico dei dinosauri automobilistici di Detroit, ha già sfiorato i 2000 miliardi di capitalizzazione a giugno 2021 e giungerà a 3000 miliardi a gennaio 2024. NVIDIA, la piccola creatura dei videogiochi, all’inizio del 2016 (l’anno in cui Jensen si presenta da Musk negli uffici di OpenAI col DGX-1) vale in borsa meno di 20 miliardi, a maggio 2023 supera i 1000 miliardi di capitalizzazione e a gennaio 2024 raggiunge 1500 miliardi. Quando addestriamo una grande rete neurale a prevedere con precisione la parola successiva in molti testi presi da Internet, ciò che stiamo facendo è imparare un modello del mondo. In superficie, sembra che stiamo semplicemente imparando le correlazioni statistiche nel testo. Ma in realtà, basta conoscere le correlazioni statistiche nel testo per comprimerle davvero bene. Ciò che la rete neurale apprende è una certa rappresentazione del processo che ha prodotto il testo. E questo testo è in realtà una proiezione del mondo. La pretesa che i modelli linguistici siano “proiezioni del mondo” li pone a un livello radicalmente diverso rispetto a quello del “pappagallo stocastico”.

Nei suoi studi, Boltzmann ha introdotto l’interpretazione statistica dell’entropia, ponendo le basi per una comprensione del disordine e dell’informazione che è importante anche per gli algoritmi di intelligenza artificiale. L’entropia incrociata, inoltre, viene usata nelle reti neurali per l’ottimizzazione. Negli anni ottanta, Geoffrey e i suoi colleghi omaggiano il fisico con la Macchina di Boltzmann, una rete neurale probabilistica che utilizza un approccio statistico per apprendere e fare inferenze dai dati.

Arm, che NVIDIA si impegna a pagare complessivamente 40 miliardi di dollari, è un gioiello nato a Cambridge, nel Regno Unito, e divenuto leader della progettazione di architetture di microprocessori ceduti per licenza ad altre aziende. Per dare un’idea della pervasività e dell’accelerazione del suo ecosistema, si pensi che i partner di Arm hanno impiegato ventitré anni, dal 1991 al 2014, per produrre i primi cinquanta miliardi di chip con la sua architettura, poi altri tre anni, dal 2014 al 2017, per i successivi cinquanta miliardi, e poi quattro anni e mezzo per aggiungerne altri cento miliardi125. Con l’acquisizione, NVIDIA passerebbe dal ruolo di partner di Arm, che paga per ottenere le licenze, a padrone delle licenze pagate dai suoi concorrenti.

L’operazione si colloca al crocevia di tutti i problemi politici possibili: la posizione anticoncorrenziale di NVIDIA, ben chiara ai regolatori europei e statunitensi (dove alla Federal Trade Commission si è appena insediata Lina Khan, esponente di una rinnovata attenzione contro i monopoli tecnologici); l’orgoglio nazionale britannico, che pur non controllando Arm attraverso i capitali può bloccare l’operazione per ragioni di sicurezza nazionale; la complicata

relazione con la divisione cinese di Arm, perdipiù nel vortice della guerra tecnologica tra Stati Uniti e Cina. Non è pane per il laboratorio scientifico di NVIDIA, ma questo è il mondo in cui NVIDIA deve vivere. Il 7 febbraio 2022 Jensen firma la sua resa, ormai evidente: l'acquisizione non potrà avvenire per "significativi problemi regolatori".

Ruolo del marketing come fattore abilitante dell'innovazione. Se gli studiosi di OpenAI nel 2018 scrivono sul loro blog che hanno combinato i Transformer e il preaddestramento non supervisionato utilizzando alcune GPU, bene, qualcuno lo leggerà, ma la verità è che non importa. Quando invece un risultato viene condiviso da milioni di persone che non hanno mai sentito nominare Geoffrey Hinton, che non ascolteranno mai la teoria del romanzo di Ilya Sutskever, che non sapranno mai ciò che il professor Sabella ha insegnato a Fei-Fei Li, e ignorano chi sia quel cinese coi capelli bianchi che indossa un giubbotto in pelle, allora è avvenuto qualcosa di grande. Jensen trae la sua conclusione, aggiungendo a tutti i suoi slogan una nuova declinazione: "È il momento iPhone dell'intelligenza artificiale".

NVIDIA ha parlato di industrie da 100.000 miliardi di dollari di ricavi nel lungo termine (circa il 2030, ma la data non viene specificata), sostanzialmente pari al PIL mondiale del 2022, in cui vede per sé un'opportunità appunto dell'1%: vuol dire 1000 miliardi, suddivisi in 100 miliardi di giochi, 300 miliardi di automotive, 300 miliardi di chip e sistemi, 300 miliardi di enterprise software, divisi a loro volta tra Omniverse e intelligenza artificiale. Quando dice "momento iPhone", Jensen sta dicendo che alcuni di questi numeri apparentemente assurdi diverranno realtà. Infatti, di questi 1000 miliardi complessivi, nel 2023 NVIDIA ha confermato che 600 miliardi "sono" intelligenza artificiale, resa possibile dalle sue "accelerazioni".

Washington, non potendo consentire a Pechino di dominare l'industria dei semiconduttori, ha attivato i suoi strumenti di guerra economica: blocco degli investimenti esteri, controllo delle esportazioni, sussidi pubblici. Come in un battito di ciglia, il movimento di migliaia di persone dai treni e dai dormitori a Taiwan, di decine di milioni di persone a Shenzhen, il sottile rumore di robot nelle fabbriche, e poi le telefonate dei fornitori per gli ordini, le file fuori da migliaia e migliaia di negozi: tutto è avvenuto allo stesso tempo, tutto ha fatto parte dello stesso momento. Non è un'epoca: è un momento. Jensen l'ha vissuto come uno spettatore, perché il momento iPhone è stato degli altri. Ed è finito. Doveva necessariamente finire: a livello economico, perché non è possibile vendere smartphone per decenni con la stessa curva di crescita; a livello politico, perché non è possibile tenere insieme in un triangolo Stati Uniti, Cina e Taiwan per arricchirsi a vicenda.

 # Parte terza. L'orologio e lo specchio
1. La principessa e la progettista

Deng Xiaoping è l'eroe di Ren Zhengfei, fondatore nel 1987 di Huawei. Dopo l'arresto di sua figlia Meng Wanzhou, direttrice finanziaria dell'azienda, all'aeroporto di Vancouver, e l'inizio formale della guerra economica degli Stati Uniti, l'imprenditore, dapprima noto per la sua riservatezza, durante il 2019 rilascia una serie di interviste ai media occidentali, tra cui "The New York Times", "The Economist", "Sky News". Con questi interventi Ren tenta l'impossibile, perché convincere il governo degli Stati Uniti che non vi sia una connessione tra il Partito Comunista e la sua azienda è impossibile. Anche se ripete che i suoi unici obblighi sono l'obbedienza alla legge cinese e il pagamento delle imposte, non può essere creduto. Usa addirittura Deng contro Xi Jinping, rivendicando la sua "diversa interpretazione" quando gli chiedono come l'idea dell'indipendenza di Huawei possa tenere davanti alla teoria di Xi Jinping del primato onnicomprensivo del Partito. "Se il Partito sa gestire tutte le organizzazioni economiche, allora non c'è bisogno di sviluppare imprese private. Basteranno i comitati di partito per gestire tutto, e non ci sarà bisogno di nient'altro. Tuttavia, l'esperienza cinese negli ultimi decenni ha dimostrato che questo modello non funziona. Ecco perché Deng Xiaoping aveva proposto le riforme e l'apertura."

Nel 2011, NVIDIA annuncia che, grazie alle sue GPU, l'istituto cinese di genomica BGI, il più grande al mondo, è stato in grado di ridurre il tempo di analisi dei lotti di dati di sequenziamento del DNA da quasi quattro giorni a sole sei ore. BGI, fondato nel 1999 come Beijing Genomics Institute per contribuire al Progetto genoma umano, si è evoluto come azienda di ricerca posseduta dai dipendenti e finanziata dal governo cinese e da altre entità (come la China Development Bank, che ha esteso una linea di credito di 1,5 miliardi di euro nel 2010). Nel 2007 ha trasferito il quartier generale a Shenzhen, l'area più dinamica della tecnologia cinese. Nel 2011, quando NVIDIA rivendica la loro collaborazione, BGI ha già una notevole presenza internazionale, in linea con la sua ambizione. Nel 2012, BGI annuncia l'acquisizione di un'azienda californiana di sequenziamento e analisi del genoma umano, Complete Genomics, per 117 milioni di dollari.

L'operazione suscita qualche protesta da parte del principale concorrente, Illumina, e di diversi politici, tra cui due membri della US-China Economic and Security Review Commission, che invitano il CFIUS – che dovrà dare la sua approvazione – a guardare con grande attenzione la vicenda, per le implicazioni militari e di sicurezza nazionale degli avanzamenti nella biotecnologia sintetica. Nonostante questa controversia, BGI e Complete Genomics ricevono una tranquilla approvazione del CFIUS e, una volta completati gli altri passaggi regolatori, l'acquisizione va in porto e l'azienda americana entra a far parte di MGI, di proprietà di BGI. Il gioiello della genomica cinese collabora anche col gigante tecnologico di Shenzhen, Huawei, che fornisce importanti soluzioni di supercalcolo e archiviazione.

Jensen racconta il lavoro con Hikvision, "leader dell'intelligenza artificiale applicata ai video: una rete di videocamere in tutto il mondo che può usare l'intelligenza artificiale per tenere le persone al sicuro". Con Hikvision, NVIDIA può adattare la DGX-1 (il supercomputer fornito nell'estate di quello stesso 2016 a Elon Musk negli uffici di OpenAI) per realizzare "un robot che pensa e respira per tenere le persone al sicuro". L'intelligenza viene applicata alla sicurezza attraverso il riconoscimento degli oggetti, degli ambienti e ovviamente dei volti, con l'enorme mole di dati forniti dalle videocamere Hikvision e processati dai server della stessa azienda cinese, denominati Hikvision Blade, grazie alla tecnologia di NVIDIA. Diventa possibile creare "centinaia di guardie di sicurezza virtuali" che controllano gli eventi più importanti; con Huawei, l'intelligenza può essere applicata al controllo del traffico.

Per Jensen, la fedeltà a Washington nel 2019 su un tema del genere è conveniente, anche perché ha molte frecce al suo arco e il destino del 5G non è certo determinante per i conti di NVIDIA. Eppure, lo stesso esito della guerra del 5G rimane ambiguo. Da un lato, è evidente che la riduzione del ruolo di Huawei in numerosi mercati, compresi quelli europei, abbia avvantaggiato Ericsson e Nokia. Dall'altro lato, negli anni seguiti all'assedio a Huawei la stessa azienda svedese si è trovata in una posizione paradossale, che ha evidenziato una volta di più l'importanza del mercato cinese. Quando nel 2020 la Svezia ha bandito Huawei dalla sua rete 5G, Ericsson ha fatto pressioni sul proprio governo per tornare indietro: il mercato interno per Ericsson al tempo valeva appena l'1% delle vendite, paragonato all'8% di quello cinese. Il giornale cinese "Global Times", con le sue formule iperboliche, ha avuto buon gioco nel raccontare della minaccia di Ericsson di lasciare il mercato svedese per non perdere le opportunità in Cina.

Pechino, come da copione, punisce Ericsson e Nokia nelle scelte per il suo 5G, sfruttando anche la concorrenza tra le due, mentre l'adozione dei nuovi standard in Occidente è lenta rispetto al ritmo delle potenze asiatiche. Ericsson deve ristrutturare le operazioni in Cina, perché non ha più un accesso stabile al suo secondo mercato. Huawei porta in tribunale le autorità svedesi sulla decisione basata sulla sicurezza nazionale, senza effetti. L'impatto cinese su Ericsson è molto significativo e si aggiunge alle debolezze dell'economia globale e del mercato delle telecomunicazioni. Così, nel 2023, il baluardo dell'Occidente destinato a sfruttare la caduta di Huawei deve licenziare 8500 persone, l'8% dei suoi dipendenti.

Il 25 settembre 2021 la principessa di Huawei Meng Wanzhou, figlia di Ren Zhengfei, scende le scale dell'aereo Air China che l'ha riportata in patria, a Shenzhen, dopo oltre mille giorni di arresti domiciliari in Canada. Il suo ritorno a casa chiude una lunga controversia politica, che ha coinvolto non solo gli Stati Uniti e la Cina, ma anche il Canada. Due cittadini canadesi, accusati di attività contrarie alla sicurezza nazionale di Pechino, sono infatti rimasti prigionieri nel paese, e anche loro possono finalmente essere liberati. Se la tecnologia vive di accelerazione, i tribunali hanno sempre e comunque un altro ritmo, e lo stesso governo degli Stati Uniti ha ammesso che il procedimento di estradizione avrebbe potuto durare mesi, o forse anni. Mesi e anni in cui Michael Spavor e Michael Kovrig, i cittadini canadesi messi sotto custodia in Cina a dicembre 2018, pochi giorni dopo l'arresto di Meng, sarebbero rimasti lontani dalle loro famiglie.

La direttrice finanziaria di Huawei il 22 settembre 2021 ha firmato una versione dei fatti che sintetizza l'impianto accusatorio degli Stati Uniti. Il gigante tecnologico cinese ha continuato a operare in Iran tra il 2010 e il 2014 attraverso una società, Skycom, controllata da un'entità a sua volta legata a Huawei, denominata Canicula. Meng ha mentito a un'istituzione finanziaria, HSBC, sul rapporto di Huawei con Skycom, nonostante conoscesse l'architettura societaria. Questa frode ha portato HSBC, che nel mentre è monitorata strettamente dal governo degli Stati Uniti per il supporto finanziario fornito ai cartelli della droga messicani, a violare a sua volta le sanzioni statunitensi verso l'Iran. Meng, pertanto, ammette l'accaduto e torna in Cina. Secondo i termini dell'accordo, le accuse vengono fatte poi cadere del tutto dal governo statunitense a fine 2022, ma si tratta solo di una formalità. Alcuni procedimenti verso Huawei continuano negli Stati Uniti, ma ormai Meng è libera.

Huawei così precipita in un cono d'ombra, priva anche dell'accesso al sistema operativo Android, parte della galassia Alphabet-Google e quindi sempre a stelle e strisce. Anche i danni economici sono notevoli, perché, dopo una tenuta

iniziale, Huawei nel 2022 precipita a 93,5 miliardi di dollari di ricavi, ben sotto la soglia di 100 miliardi indicata come "pessimistica" da Ren Zhengfei nel 2019. Ma già quello stesso Yan Xuetong, presidente dell'Istituto di relazioni internazionali dell'Università di Tsinghua, aveva detto chiaramente che Huawei non può morire: "Il popolo sa, il governo sa che, se Huawei non può sopravvivere, il paese perderà la speranza della rinascita nazionale".

Seagate vede nelle restrizioni a Huawei una grande opportunità di mercato: l'azienda firma un documento triennale di cooperazione strategica con Huawei, per diventare fornitore prioritario dell'azienda di Shenzhen. È facile immaginare i frequenti incontri tra Huawei e Seagate a Singapore, territorio neutro della guerra tecnologica, dove Seagate ha una sede importante. E senz'altro i tecnici di Huawei avranno visitato le strutture produttive di Seagate non solo in Cina, ma anche in Malesia e Thailandia, dove l'azione della "potenza del Pacifico" degli Stati Uniti ha limiti oggettivi. Mentre gli altri concorrenti hanno rinunciato a un'opportunità di mercato e a un grande cliente, Seagate è in grado di sfruttarli al meglio.

Si accumulano circa 1,1 miliardi di dollari di ricavi da Huawei per circa 7,4 milioni di dischi rigidi. Il gigante di Shenzhen ha così tanta fame di componenti che la capacità produttiva di Seagate non riesce a stargli dietro. Eppure, c'è sempre qualche inghippo nell'ingranaggio. Seagate, nelle sue fabbriche lontane dallo sguardo del governo, non opera da sola: per i processi produttivi, utilizza strumentazione fornita da due aziende statunitensi, sia per quanto riguarda i sistemi di ispezione basati su laser, sia per l'incisione e la deposizione di fascio ionico. Questi fornitori, in un certo senso, "intrappolano" Seagate, perché tali complessi procedimenti sono nell'elenco dei controlli sulle esportazioni, ed è vietato utilizzarli senza autorizzazione per realizzare materiali da vendere a Huawei.

Con Seagate e Micron-Fujian, Huawei svolge due ruoli differenti, seppur collegati al disegno più ampio della centralità dell'azienda di Shenzhen nel complesso industriale-tecnologico cinese. Huawei, che non opera solo attraverso divisioni e sussidiarie già finite nelle maglie delle sanzioni statunitensi, ma ovviamente può crearne di nuove (che vanno riconosciute come tali da Washington prima di poterle sanzionare), svolge un ruolo centrale di capofiliera. L'evoluzione dell'effetto Pechino è l'effetto Huawei: l'integratore di tecnologia interviene grazie alla propria scala. Da un lato, il gigante di Shenzhen tiene in vita aziende che altrimenti finirebbero fuori mercato; dall'altro lato, allarga la propria supply chain e il proprio campo d'azione.

Tutto questo riporta all'ambizione fondamentale: se la Seagate di turno non avesse dovuto operare, nella sua fabbrica in Cina (o in Thailandia o in Malesia, perché gli Stati Uniti non hanno né avranno le risorse per capire davvero cosa accade lì), attraverso processi produttivi dotati di tecnologia statunitense, allora difficilmente sarebbe stata scoperta, perché non ci sarebbe stato il gancio dei due fornitori che la mettono in trappola. Di conseguenza, l'ambizione finale è che la Seagate di turno non sia un'azienda americana, e venga sostituita da attori cinesi, come il resto della filiera: le sanzioni contengono questo incentivo implicito per Huawei. Anche per questo, Huawei non può morire.

L'effetto Huawei: a) mantenere, per quanto possibile, il rapporto con alcuni fornitori che per il momento non possono essere sostituiti, attraverso ogni mezzo, sfruttando anche le zone grigie e le ambiguità di leggi che, per essere efficaci (per esempio, per riconoscere nuove diramazioni di Huawei), debbono cambiare; b) soccorrere le capacità tecnologiche cinesi in difficoltà attraverso la propria scala e i propri ordini: Huawei, grazie a una probabile collaborazione con autorità pubbliche cinesi, a livello locale e nazionale, svolge così una continua attività di scouting che allarga il suo campo d'azione, con l'ambizione di sostituire nel giro di qualche anno le grandi aziende degli Stati Uniti; c) investire nella filiera interna e nelle nuove opportunità di mercato: il ritiro (temporaneo) di Huawei dagli smartphone fornisce importanti capitali per la sopravvivenza dell'azienda, su cui continuano a fluire risorse pubbliche.

Sappiamo che Huawei fa parte dell'approccio cinese all'intelligenza artificiale, concentrato in particolare su alcune applicazioni industriali in grado di fare leva su macrosettori e tendenze su cui la Cina ha già sviluppato una leadership generale. Considerare questi elementi ci aiuta a uscire da una caricatura del sistema cinese in cui si crede che lo stato faccia tutto, occupandosi nel dettaglio anche delle applicazioni industriali. Non è così: sono le specifiche applicazioni industriali della superpotenza manifatturiera a trainare l'innovazione privata, ovviamente in un sistema politicamente protetto e controllato, e dove ci sono obiettivi abilitanti generali che conosciamo, come la crescita dell'industria di macchinari per i semiconduttori per ridurre la dipendenza dalle aziende estere, in particolare statunitensi.

Huawei, come Hikvision, è attiva nei "porti smart": analisi dei dati, sorveglianza e miglioramento dell'efficienza sono possibili su vasta scala, anche per via del rilievo della Cina negli scambi commerciali e nella logistica globale.

Sfruttando il ruolo cinese di superpotenza manifatturiera. Prima di tutto nell'industria automobilistica, a partire da BYD, che ha assunto migliaia di ingegneri di software in pochi anni e che, secondo il fondatore Wang Chuanfu, il ricercatore di chimica che ha avviato l'azienda con una ventina di persone nel 1995 in un capannone di Shenzhen, a inizio 2024 ha oltre 90.000 dipendenti dedicati a ricerca e sviluppo degli oltre 600.000 complessivi. Nessuno, finora, ha eguagliato questi numeri e questi tassi di crescita. E per BYD vale lo stesso assunto di Huawei: per la sua importanza sistemica e il suo prestigio, qualunque cosa accada, non può cadere.

L'obiettivo di fondo di Pechino rimane lo sviluppo dei campioni nazionali, attraverso ecosistemi comunicanti, in cui convivono il coordinamento e la concorrenza, nelle telecomunicazioni, nell'automotive, nella robotica. La strategia è superare difficoltà sempre più evidenti, sempre più pesanti, tanto nel settore immobiliare (al quale Jensen guardava come grande cliente!) quanto nella demografia, rilanciando sempre sui punti di forza: sorprendere attraverso centinaia di nuovi "piccoli giganti" nei principali settori tecnologici, sfruttando il potere di rete dei "campioni manifatturieri" e dei grandi capifiliera, anche grazie al ruolo di uno "stato acceleratore". Questo stesso sistema di vasi comunicanti manifatturieri, che per Xi Jinping ha la priorità sugli altri aspetti dell'economia, è al centro della trasformazione dell'intelligenza artificiale in Cina.

Nel processo convivono le attività lecite, quelle illecite e le zone grigie, e tutto ciò si declina in innumerevoli storie, dal ricercatore di fluidodinamica che non fa assolutamente nulla di "duale" e non riesce ad avere un visto, fino a Linwei Ding, noto anche come Leon Ding, il dipendente di Google residente a Newark, California, che, secondo l'accusa del governo statunitense del 2024, mentre lavorava per una delle più grandi aziende del mondo e, in un parallelo illecito, per due aziende cinesi, vende i segreti industriali su "l'architettura e la funzionalità dei chip e dei sistemi GPU e TPU, il software che consente ai chip di comunicare ed eseguire attività e il software che orchestra migliaia di chip in un supercomputer all'avanguardia capace di operare nei settori dell'apprendimento automatico e dell'intelligenza artificiale.

NVIDIA è un avamposto per monitorare la frontiera tecnologica cinese. I supercomputer più avanzati vogliono avere il prodotto migliore e quindi scelgono NVIDIA. In questo modo, NVIDIA può sapere cosa accade e, se necessario, condividere le sue valutazioni col governo degli Stati Uniti, per ridurre la "nebbia di guerra". Non è il "commercio dei lumi" sognato da Leibniz, ma si mantiene comunque un canale di comunicazione, l'esigenza di sapere quello che succede, quali siano le necessità e gli avanzamenti reali. Il vincolo politico continua a esistere, ma è legato alla riduzione della sorpresa. La "cerimonia" della GTC di Jensen, con la sua sfilata di clienti e di prodotti, acquista quindi anche questa funzione. Kevin Xu ha riassunto bene questa prospettiva: "Gli affari di NVIDIA in Cina hanno un importante valore geopolitico per gli Stati Uniti. Non si tratta solo di un'azienda che fa più soldi. Si tratta di mantenere un certo accesso al vasto mercato di un concorrente che può rivelare tracce e informazioni sul suo progresso attuale e la sua direzione futura".

2. Chi dice umanità. La corporate governance di OpenAI

L'articolo fondamentale di Ronald J. Gilson del 1999 analizza come la Silicon Valley sia riuscita a emergere come un distretto industriale di alta tecnologia di successo, confrontandola con la Route nel Massachusetts. Il fattore chiave identificato è la diversa struttura legale della California rispetto al Massachusetts, in particolare riguardo all'applicabilità dei patti di non concorrenza post-impiego. In California, dove tali patti non sono applicabili, si è sviluppata una cultura di alta mobilità dei dipendenti tra le aziende, facilitando la diffusione di conoscenze e innovazioni e promuovendo una rapida evoluzione industriale.

Nella classica e sempre problematica divisione del diritto societario tra proprietà e controllo, l'ecosistema della Silicon Valley ha perfezionato l'istituto delle azioni multiclasse, alla ricerca di un equilibrio specifico tra la necessità di capitalizzazione delle aziende e il desiderio dei fondatori di mantenere il controllo decisionale. Google e Facebook sono due classici esempi in questo senso. Google, al momento della sua offerta pubblica iniziale, ha introdotto una struttura azionaria multiclasse che ha permesso ai fondatori, Larry Page e Sergey Brin, di mantenere un controllo significativo sull'azienda nonostante l'apertura al capitale. Facebook ha seguito un percorso simile, utilizzando azioni privilegiate per raccogliere fondi nelle fasi critiche della sua crescita, pur mantenendo il controllo decisionale nelle mani di Mark Zuckerberg. In vari scritti sulla corporate governance, l'avvocato David Berger dello studio Wilson Sonsini – che è basato a Palo Alto e nel rivolgersi ufficialmente al regolatore statunitense, la Securities Exchange Commission, si paragona ironicamente a Ron LaFlamme di Silicon Valley –, affronta complessivamente la questione di queste strutture azionarie.

Berger, nella sua difesa dell'istituto, propone tra l'altro di dare agli azionisti informazioni più dettagliate sulle potenziali estensioni delle classi azionarie duali già dalla quotazione, in modo da ridurre l'incertezza e migliorare la trasparenza. L'analisi di Berger evidenzia il conflitto intrinseco tra il desiderio dei fondatori di mantenere il controllo a lungo termine dell'azienda e la necessità di proteggere gli interessi degli azionisti minoritari. La creazione di ricchezza della Silicon Valley, con le sue strutture giuridiche, è un magnete anche per i giuristi, e alcuni di essi portano nell'Eldorado un bagaglio di competenze nuove. Uno dei casi più interessanti è quello di Joseph Grundfest.

Nato nel 1951 da immigrati ebrei polacchi e russi, Grundfest si forma nelle scuole talmudiche di New York, per poi studiare economia e cambiare direzione verso il diritto. Da democratico, collabora con l'amministrazione Reagan e, proprio per la sua affiliazione e le regole della rappresentanza della nomina, viene indicato come commissario della Securities Exchange Commission a soli trentaquattro anni. Nel corso del mandato spinge verso una deregolamentazione del mercato per trattenere i capitali degli investitori statunitensi, che negli anni ottanta erano intenti a investire in altri mercati come il Giappone, la Corea e il Regno Unito.

Anche la saga di OpenAI, intesa come crisi di corporate governance, finisce per riprendere questi dilemmi. La Carta di OpenAI è il documento, pubblicato nel 2018, che descrive i principi utilizzati per eseguire la "missione", ed è frutto di un processo di deliberazione interno, col concorso di esperti esterni. Come spiegato da Ilya Sutskever, OpenAI in quel momento adotta una definizione di intelligenza artificiale generale legata a una capacità superiore a quella degli esseri umani "nella maggior parte del lavoro dotato di valore economico". Il fine di OpenAI è assicurare che l'intelligenza artificiale generale così intesa "vada a beneficio di tutta l'umanità".

Nel primo manifesto di OpenAI, a fine 2015, non appare l'espressione "intelligenza artificiale generale". L'unico riferimento alla generalità riguarda i limiti e le delusioni dei risultati iniziali. Nel manifesto c'è l'obiettivo di far "avanzare l'intelligenza digitale nella maniera che porti maggiore beneficio all'umanità nel suo complesso". L'interesse dell'umanità non viene definito nel 2015, ma il manifesto rivendica la natura di non profit per "costruire valore per tutti invece che per gli azionisti". Il passaggio più significativo è quello in cui si afferma: "Il nostro dovere fiduciario primario è verso l'umanità". È il punto in cui emerge in modo chiaro l'ambiguità di una missione in cui si utilizzano i concetti propri del diritto societario, di un'azienda che mette il valore economico al centro della propria azione mantenendo però allo stesso tempo, e anzi rivendicando, il concetto di umanità. I "migliori interessi dell'umanità" richiedono un'accumulazione di potere da parte di OpenAI. Il potere del talento, il potere del calcolo, il potere dei clienti.

Gli studi di OpenAI sull'infrastruttura di calcolo necessaria per l'addestramento dei modelli, nell'accelerazione dopo AlexNet (il "big bang"), confermano la posizione di Bill Dally e di Jensen. Nel 2018, OpenAI indica quattro ere del calcolo (prima del 2012, 2012-2014, 2014-2016, 2016-2017). OpenAI, considerando l'analisi dei dati, i precedenti delle tendenze esponenziali nell'informatica, la specificità del machine learning, ma anche gli "incentivi economici", scommette nel prosieguo di questa tendenza. Lo stesso Gordon Moore, come abbiamo visto, chiarisce che la sua legge è anzitutto una questione economica, è un orizzonte di imprenditorialità che mette insieme una serie di incentivi in cui i vari nodi della rete si riconoscono, e in cui emerge un capofiliera, Intel. La legge di Huang accelera questo paradigma e vuole avvolgere intere industrie con i suoi prodotti e i suoi servizi.

OpenAI, secondo questo schema, gioca tutte le parti della commedia: il suo concetto di "umanità" è allo stesso tempo filo di seta e misura protettiva, e l'azienda di ricerca si pensa come reparto ben organizzato per spegnere l'incendio che deve e vuole continuare ad alimentare. La conseguenza della Carta del 2018, con la ricerca di risorse necessarie per pagare la capacità di calcolo, è l'accordo con Microsoft, il dinosauro che Nadella sta riportando al centro della geografia tecnologica. Il motore principale dell'accordo è Kevin Scott, direttore delle tecnologie di Microsoft. Nell'era di Nadella, Scott cerca di posizionare Microsoft per la competizione sull'intelligenza artificiale, dove Google sembra avere un ruolo dominante.

Nel momento in cui riconosce la legge di Huang, OpenAI non può più operare da sola e deve accordarsi con uno dei giganti dalle maggiori capacità di cloud: Amazon, Microsoft o Google. Quest'ultima va esclusa, perché OpenAI nasce in contrasto con Google, che peraltro possiede DeepMind. E con Microsoft si è già creato un rapporto nel 2016, che ha reso Azure "la principale piattaforma di cloud che OpenAI utilizza per l'apprendimento profondo e l'intelligenza artificiale". Ilya, Altman e Brockman sottolineano la capacità di Azure di fornire una notevole scala di hardware: nel dettaglio, le GPU K80 di NVIDIA e l'Infiniband di Mellanox. Servono capitali. Per questo, nasce OpenAI LP, società ibrida che è sia for profit che non profit. Nello specifico, il profitto è limitato nel ritorno (capped-profit) al massimo di cento volte per il primo round di investitori, e poi i profitti successivi dovranno essere reinvestiti nell'entità di controllo di OpenAI LP, la non profit, che a sua volta ha la maggioranza della società ibrida ed è governata da un

consiglio di amministrazione. I membri inizialmente sono otto: tre dipendenti di OpenAI (Ilya, Altman, Brockman), Adam D'Angelo, Holden Karnofsky, Reid Hoffman, Shivon Zilis, Tasha McCauley.

Quest'architettura deve giustificare contratti con interessi economici e obblighi giuridici, e ritornano le precisazioni "sul dovere fiduciario primario" degli obiettivi della Carta di OpenAI: di conseguenza, tutti gli investitori e i dipendenti firmano un documento secondo cui devono aderire ai doveri fiduciari nei confronti dell'umanità. A luglio 2019, per avanzare sulla strada verso l'intelligenza artificiale generale, viene annunciata la cooperazione con Microsoft, che investe un miliardo nella società for profit e diviene partner esclusivo per il cloud. Geoffrey Hinton ha descritto la determinazione di DeepMind per AlphaGo paragonandola al progetto Manhattan, ma secondo Sam Altman OpenAI è il progetto Manhattan.

Nel 2023 Microsoft estende la collaborazione per un importo "multimiliardario" non specificato (finché non c'è un documento pubblico, non si può sapere esattamente quanto sia l'investimento, e quanto sia in denaro e quanto in capacità di calcolo e servizi), ma stimato a 10 miliardi di dollari, per avere il 49% dell'azienda investibile e il 75% dei profitti di OpenAI fino al ritorno dell'investimento, OpenAI ha già raggiunto il successo del "momento iPhone dell'intelligenza artificiale". Questo successo deriva anche da una mentalità commerciale, dall'attenzione per il prodotto di manager come Altman, Brockman e Murati, da un equilibrio tra ciò che deve essere aperto, cioè diffuso (il prodotto "gettato" al pubblico e fatto testare quindi da milioni di persone, elemento che segna la superiorità della tesi di OpenAI rispetto al laboratorio scientifico "protetto" di DeepMind), e quello che deve essere chiuso, cioè riservato (per esempio, i contratti con Microsoft e ciò che essi regolano, comprese le modalità d'uso del codice dei modelli). Secondo i critici, col "cambiamento nella struttura organizzativa di OpenAI è avvenuta una transizione dall'apertura limitata e ambiziosa alla chiusura totale".

Joe Grundfest ha fornito un'acuta interpretazione dell'architettura di OpenAI, individuando tre problemi di fondo. Il primo è che si possano realizzare le protezioni (guardrails) a cui OpenAI fa riferimento nei suoi documenti. Il secondo è la struttura, che a detta del giurista è totalmente sbagliata. Il terzo problema sono le persone, anch'esse inadeguate. Da ultimo, riducendo all'osso il problema, Grundfest osserva: "Partiamo con la futilità dell'obiettivo. Dovrebbe essere: per il bene dell'umanità, giusto? Cosa diavolo significa?". Finché non si riconosce questa premessa, non è possibile avere un discorso informato sulla corporate governance di OpenAI e su ciò che ne consegue in termini di governo societario e di legittimazione. Nel momento in cui si continuano a mischiare i concetti propri del diritto societario col macroconcetto di umanità, non è possibile venire fuori da questo dilemma. Il richiamo all'umanità, oltre ad aumentare la confusione, può essere una copertura per qualunque altra azione.

A questo proposito vale, pur con una modifica necessaria, una massima di Proudhon: chi parla di umanità, vuol trarvi in inganno". Per Schmitt, "proclamare il concetto di umanità, richiamarsi all'umanità, monopolizzare questa parola", in un contesto di politica portata a estreme conseguenze nella guerra, sta a evidenziare che "al nemico va tolta la qualità di uomo", col risultato opposto della conduzione della guerra verso l'estrema inumanità. È però facile chiedersi a quali uomini toccherebbe il terribile potere che è legato a una centralizzazione economica e tecnica estesa a tutto il mondo". Questo potere di riorganizzazione attraverso il concetto di umanità, invece della politica come pluralità, implica una centralizzazione che viene decisa dai custodi del concetto di umanità. In un concetto portato all'estremo, in cui l'intelligenza artificiale, oltre al frutto in grado di generare abbondanza e quindi di liberare gli uomini attraverso la generazione di valore economico, contiene anche il pericolo più grande, apocalittico, davanti al quale si invoca sicurezza, la custodia diviene un dovere assoluto, perché si tratta di un dovere nei confronti dell'umanità.

La crisi permanente di un'umanità divisa sarebbe così risolta da chi possiede – dice di possedere – la conoscenza per dominare l'arma che può migliorare o distruggere tutto ciò che è umano. Da un lato, non posso definire ciò verso cui ho doveri fiduciari, l'umanità, e non posso "interpellarla", né stabilire chiaramente le modalità con cui "l'umanità" possa citarmi in giudizio nel momento in cui contravvengo ai miei doveri verso di essa. Dall'altro lato, siccome il mio dovere è verso "l'umanità", allora posso sempre richiamarmi a qualcosa, nel codice di condotta al quale mi devo conformare, che considero sovraordinato rispetto a qualunque altro obbligo. Singole decisioni sono basate su un assoluto, che per sua natura richiede una decisione esistenziale. Anche se OpenAI, tecnicamente, sta solo cercando di raccogliere soldi per comprare GPU (in sintesi, "Avevamo bisogno fin dall'inizio di almeno un miliardo per comprare capacità di calcolo, invece abbiamo avuto solo 130,5 milioni"), il suo percorso si inserisce nei macrotermini delle grandi imprese tecnologiche moderne, come quelle del suo creatore iniziale, Elon Musk. Secondo questa tendenza, la conquista di un mercato, obiettivo specifico e dimostrabile (il mercato dell'auto elettrica o delle costellazioni di satelliti, per esempio), rimanda a concetti e obiettivi indimostrabili, come "coscienza", "umanità", "specie multiplanetaria".

Certo, può esserci sempre un contenzioso che riporta questa discussione sulla terra, su temi come la proprietà intellettuale. Per esempio, quello che il "New York Times" nel 2023 ha intentato verso OpenAI e Microsoft, sostenendo tra l'altro che, "nonostante le sue premesse iniziali di altruismo, OpenAI è diventata velocemente un'impresa multimiliardaria votata al profitto, costruito in larga parte con lo sfruttamento non autorizzato di lavori sotto copyright posseduti dal 'Times' e da altri". Eppure, ogni discussione e processo decisionale teoricamente si basa sull'aderenza di dipendenti e amministratori a concetti come "umanità" e "intelligenza artificiale generale", che vengono inseriti nel circolo delle decisioni da parte dei "gigolò dell'umanità".

Microsoft non può più guadagnare, e non si capisce come OpenAI possa continuare a funzionare, visto che si basa su infrastrutture che non sono sue, e più prosaicamente sul consumo idrico e l'elettricità di una schiera di computer in Iowa. Nei contratti, Microsoft avrà comunque ottenuto garanzie su questo aspetto. Certo, esiste un possibile punto di equilibrio: OpenAI per aumentare il proprio valore e l'interesse degli utenti, non sulla base del nulla ma a partire da prodotti che hanno una resa e una pervasività sempre maggiori, continua a dire "ci stiamo avvicinando sempre di più, siamo quasi all'intelligenza artificiale generale", senza che avvenga l'annuncio faustiano con una conferenza stampa, e Microsoft nel mentre utilizza sempre più sia i prodotti di OpenAI nei suoi sistemi, sia l'alone stesso dell'intelligenza artificiale, così che i ricavi e la capitalizzazione aumentino in modo molto più consistente.

Il ponte per l'intelligenza, prima di raggiungere l'obiettivo della "generalità", è autoapprendere per saper fare soldi. Ciò non significa che sia possibile, né oggi né per il programma che eventualmente supererà il "test di Suleyman", scrivere in un prompt "indicami come migliorare la memoria a grande ampiezza di banda e realizza un prodotto in grado di farlo su vasta scala", oppure "siccome il titolo di Supermicro è salito in borsa, dammi due opzioni di un sistema di raffreddamento più competitivo", e ottenere una formula magica.

La stessa NVIDIA ha annunciato un innovativo processo di litografia computazionale, cuLitho, che mira a risolvere alcune delle sfide emergenti nella produzione di semiconduttori, che richiede enormi carichi di lavoro. CuLitho utilizza le GPU per accelerare notevolmente il processo di litografia inversa, con prestazioni fino a quaranta volte superiori rispetto ai metodi tradizionali basati su CPU. Per esempio, 500 sistemi NVIDIA DGX H100 possono svolgere il lavoro di 40.000 sistemi CPU, riducendo l'uso di energia e l'impatto ambientale. CuLitho può generare fotomaschere (i modelli per la progettazione di un chip) molto più rapidamente: ciò che richiede due settimane può essere elaborato durante una notte, con una grande accelerazione produttiva, a cui si aggiunge il miglioramento della resa attraverso varie tecniche di intelligenza artificiale. Nel progetto, NVIDIA collabora coi leader di tre segmenti della filiera dei semiconduttori: ASML (leader dei macchinari), TSMC (leader della produzione), Synopsys (leader dell'electronic design automation). Sembra incredibile e rivoluzionario, no? A settembre 2023, Jensen utilizza questo stesso progetto, su cui NVIDIA è al lavoro da tre anni, per dire che spesso "si è sul binario giusto ma non si ha successo, semplicemente perché è difficile": "We haven't shipped a thing", il prodotto non esiste ancora.

Il 17 novembre 2023, con uno scarno comunicato diffuso attraverso il blog, il consiglio di amministrazione della non profit di OpenAI annuncia a sorpresa l'uscita di Sam Altman come amministratore delegato e membro del consiglio, e la nomina come nuovo CEO ad interim di Mira Murati, direttrice delle tecnologie. Alla descrizione dei meriti di Murati segue una spiegazione poco chiara: "L'uscita di Altman fa seguito a una revisione deliberativa da parte del consiglio, che ha concluso che non è stato coerentemente sincero nelle sue comunicazioni con il consiglio, impedendo a esso di esercitare le proprie responsabilità. Il consiglio non ha più fiducia nella sua capacità di continuare a guidare OpenAI.

Il consiglio di OpenAI, dal 2015 fino a quel momento, è cambiato in modo significativo. All'inizio i due membri sono Elon Musk e Sam Altman, nel 2017 affiancati da Ilya, Greg Brockman, Holden Karnofsky dell'organizzazione filantropica Open Philanthropy – che dà a OpenAI 30 milioni di dollari ed è finanziata in gran parte dal co-fondatore di Facebook, Dustin Moskovitz, e dalla moglie Cari Tuna – e Chris Clark, imprenditore con esperienza politica come sindaco di Mountain View, che lavora per OpenAI e Y Combinator (l'incubatore da cui proviene Altman). Nel 2018 avvengono numerosi cambiamenti: anzitutto l'uscita di Elon Musk e quella di Chris Clark, che comunque resta a lavorare con OpenAI come responsabile non profit e iniziative strategiche; l'ingresso di Reid Hoffman, fondatore di LinkedIn; di Adam D'Angelo, ex capo delle tecnologie di Facebook e amministratore delegato di Quora; di Sue Yoon, che già nel 2020 non farà più parte del consiglio e in seguito andrà a lavorare per Google; e infine di Tasha McCauley, imprenditrice nell'ambito della robotica.

Nel 2020, nel consiglio è presente anche Shivon Zilis, tecnologa canadese, poi dirigente delle aziende di Elon Musk e madre di due dei suoi figli. Nel 2021 entrano a farne parte due figure dell'ambito politico: Will Hurd, già agente della CIA e parlamentare repubblicano, che si dimetterà nel 2023 per la sua campagna presidenziale (si ritirerà

presto dalle primarie repubblicane, dando il proprio sostegno a Nikki Haley) e Helen Toner, ricercatrice australiana che lavora al Center for Security and Emerging Technology (CSET), un think tank nato all'Università di Georgetown sull'analisi delle tecnologie emergenti e del loro impatto politico, finanziato prevalentemente da Open Philanthropy, cioè, in ultima analisi, sempre dai capitali di Dustin Moskovitz. Il CSET, con decine di milioni di risorse, ha già realizzato in pochi anni un lavoro significativo per la competizione tra Stati Uniti e Cina, per esempio traducendo e analizzando i principali documenti strategici cinesi disponibili sulla politica industriale e tecnologica per l'attuazione del piano Made in China 2025. Moskovitz fa parte del movimento del cosiddetto "altruismo efficace", che sulla base del pensiero del filosofo Peter Singer ha elaborato una pseudovariante dell'utilitarismo per suggerire alle persone facoltose come regalare con maggiore efficienza o con maggiore "impatto" i loro soldi. Nel corso del 2023, per varie ragioni, si dimettono dal consiglio Hoffman, Zilis e Hurd. Restano quindi, a novembre 2023, Altman, Brockman, Sutskever, D'Angelo, McCauley e Toner. I primi due vengono sfiduciati dagli altri quattro membri, perché Brockman seguirà Altman.

Non vengono fornite spiegazioni esaurienti per il cambio repentino di leadership, e – qui è appropriato il giudizio di Grundfest su un consiglio di persone inadeguate – con una totale ignoranza del contesto, in particolare di due fattori: i rapporti con l'azionista Microsoft, da cui dipende la sussistenza di OpenAI, e le reazioni dei dipendenti. I dipendenti di OpenAI, che sono un asset molto importante dell'azienda in uno scenario fortemente competitivo su risorse umane limitate nel settore e contese dai principali attori, hanno un forte interesse al mantenimento del successo dell'azienda, misurato come valutazione in round degli investimenti, perché a esso è legata una parte ingente dei loro compensi.

I dipendenti, come del resto buona parte dell'ecosistema della Silicon Valley, da Eric Schmidt a Brian Chesky, co-fondatore e amministratore delegato di Airbnb, appoggiano Altman. Più di 700 dipendenti di OpenAI, che a quel punto sono circa 770, firmano una lettera che accusa il consiglio di incompetenza e minaccia dimissioni di massa, a meno che non se ne vada il consiglio stesso, sostituendo gli attuali membri con "consiglieri indipendenti come Bret Taylor e Will Hurd e riportando Sam Altman e Greg Brockman". La lettera è firmata dallo stesso Ilya, che in un tweet si pente della sua partecipazione alle azioni del consiglio nei panni di Microsoft. La sorpresa del successo di OpenAI è un passaggio decisivo del vostro ciclo di crescita, comunque già in corso. Chi vi chiamava "dinosauro di Detroit" ora prega per la vostra considerazione. Se OpenAI salta per aria, perdete il vostro investimento, ma in ogni caso, con una spesa che siete in grado di assorbire, potete comprare tutti i talenti di OpenAI, dire che sono parte di Microsoft oppure inventare un altro nome, se le vostre ricerche vi dicono che conviene fare così. Qualunque cosa accada, vincete, a meno che qualcuno dei concorrenti al vostro stesso livello – quello dell'oligarchia digitale – possa impadronirsi di OpenAI, ma questo non è possibile, perché avete diritti che derivano dal vostro accordo e dal fatto che avete in mano la dimensione materiale della start-up, che, oltre alla proprietà intellettuale e alle persone, è letteralmente composta dai vostri computer.

Nadella esprime questi concetti in termini molto chiari, come un manager spietato: Domani, se OpenAI scomparisse, non vorrei che nessuno dei nostri clienti si preoccupasse, in tutta onestà, perché abbiamo tutti i diritti per continuare l'innovazione, e non solo per servire il prodotto, ma possiamo andare avanti e fare da soli tutto ciò che stavamo facendo in collaborazione. Abbiamo le persone, abbiamo la capacità di calcolo, abbiamo i dati, abbiamo ogni cosa. Nel giro di pochi giorni, Altman e Brockman tornano ai loro ruoli di amministratore delegato e presidente, e viene costituito un nuovo consiglio a tre, in cui del gruppo precedente resta solo Adam D'Angelo. I due nuovi membri sono Bret Taylor, che ha esperienza in società come Google, Facebook, Twitter e Salesforce, e l'ex segretario al Tesoro ed economista Lawrence Summers. Il consiglio, nella configurazione di fine novembre 2023, è quindi composto solo da membri esterni a OpenAI. A marzo 2024, Altman ritorna nel consiglio, che coopta altre tre personalità; un'altra grande vittoria del capitale in una lotta sull'etica dell'intelligenza artificiale.

La "saga di Altman offre una risposta: non possiamo contare su regole etiche, strutture di corporate governance o anche consiglieri di amministrazione con saldi principi per tenerci al sicuro. Ci hanno provato, e questo va loro riconosciuto, ma non era abbastanza; c'è un'altra chiave di lettura, se vogliamo più "nobile", per leggere il tumulto di OpenAI, e riguarda il dilemma del rapporto tra l'uomo dell'impresa e l'uomo della scienza: in fondo, lo stesso dilemma della "scienza come professione e vocazione" nel nostro tempo. Come abbiamo visto, figure come Bill Dally, Demis Hassabis, Geoffrey Hinton, Shane Legg, Fei-Fei Li, Ilya Sutskever sono i filosofi-scienziati del nostro tempo. Il loro dramma è dover navigare tra legge di Moore, legge dei ritorni acceleranti, legge di Huang: in parte perché credono a queste modalità di organizzare la logica della loro ricerca, in parte perché non è possibile scappare dall'accelerazione economica degli stessi obiettivi scientifici se non facendo un passo indietro, come Alex Krizhevsky.

In particolare, la mente di una “impresa di ricerca” non può essere estranea al meccanismo dell’accelerazione capitalista mentre viene sempre più “strattonata” dalla potenza della sicurezza nazionale, anche se la scienza come professione è la ricerca continua di una liberazione intellettuale. Solo che non ti puoi liberare né dalla sicurezza nazionale né dal capitale: o meglio, puoi farlo a parole. Nei fatti, no.
La caduta di Sutskever, che a maggio 2024 lascia formalmente OpenAI per annunciare poco dopo una nuova impresa, Safe Superintelligence Inc. (SSI), al fine di perseguire il suo scopo scientifico puro, simboleggia proprio la tensione che la mente inquieta della scienza non può risolvere.

In termini economici, quello di Musk non si può considerare un buon affare. L’imprenditore che ha riportato l’America a “costruire”, l’ingegnere capo che ha aperto strade significative a imprese che in tanti consideravano impossibili, dall’auto elettrica allo spazio, non ha finora messo a punto una vera superapp (everything app) paragonabile al ruolo che ha assunto WeChat in Cina. Il giro d’affari di X non aumenta. Non ci sono profitti. Il valore, a fine 2023, è più basso rispetto all’acquisizione del 2022. Uno degli investitori, Fidelity, ha ribassato il valore del proprio investimento del 71,5% a novembre 2023. La piattaforma non è indispensabile per gli utenti. Non è di certo l’unica porta d’ingresso della connettività.

In che modo attori e spettatori dell’intelligenza artificiale possono partecipare al banchetto? Qui entra in gioco X, e Musk ne ha avuto la conferma. Non solo per lo sviluppo di un modello, Grok, della sua start-up xAI, che, oltre a utilizzare l’ampia disponibilità di dati del social network, vuole fornire risposte sarcastiche e irriverenti per denudare la pretesa delle altre società di intelligenza artificiale di costruire spazi protetti da una libertà di espressione assoluta. In Musk c’è anche, innegabile, la ragione personale. Se il successo di OpenAI rappresenta per lui una grande ferita, in riferimento a uno dei temi al quale ha dedicato la sua attenzione pubblica, con una sequela impressionante di dichiarazioni sui pericoli dell’intelligenza artificiale e il rischio esistenziale per l’umanità, va ricordato che le vicende societarie di OpenAI sono state comunicate e commentate, in grande prevalenza, attraverso X. In un anno, si è passati da un uso tutto sommato strumentale della piattaforma (riprendere i comunicati di OpenAI su ChatGPT) a un uso sempre più personale e pervasivo: al comunicato che presenta l’uscita di Sam Altman corrisponde una sequela di tweet dove i vari protagonisti dell’azienda commentano, chiariscono, puntualizzano, polemizzano. Tutto è stato reso pubblico attraverso X.

A febbraio 2024, Elon Musk, assistito dallo studio legale della Silicon Valley Irell & Manella (specializzato nella proprietà intellettuale, che ha seguito anche ASML), denuncia presso la corte di San Francisco Sam Altman, Greg Brockman e OpenAI, in tutta la sua conformazione giuridica, umanitaria e non. Nella denuncia, Musk riprende la sua versione della vicenda di OpenAI, già nota attraverso gli articoli dei giornali e le sue biografie: la preoccupazione verso Page, Google e DeepMind, l’idea di un’impresa di ricerca aperta e non profit, il reclutamento dei talenti a partire da Ilya, il tradimento dell’ispirazione iniziale attraverso l’accordo con Microsoft. Musk ricorda di aver dato “più di 44 milioni a OpenAI tra il 2016 e il 2020”, e vuole che OpenAI torni all’ispirazione iniziale, di ricerca aperta al pubblico, e che smetta di lavorare per gli interessi finanziari di Microsoft.

Nonostante abbia ricevuto l’assistenza di uno studio legale esperto, il caso di Musk sembra essere molto debole: per esempio, identifica il “contratto” tra i fondatori con uno scambio di mail e con un documento che contiene la formula “quando applicabile”. Pertanto, l’interesse di Musk è più quello di esporre in pubblico alcune questioni, che di ottenere un risultato concreto. Pochi giorni più tardi, OpenAI risponde attraverso il suo blog, riprendendo una versione della vicenda già intuibile attraverso gli articoli dei giornali. Per via dell’infrastruttura di calcolo necessaria, OpenAI nel 2017 si trova davanti a una scelta: cessare di essere rilevante o ottenere capitale. Musk vede il futuro di OpenAI all’interno di Tesla, per avere infrastruttura, dati, capitali e controllo: la tesi viene espressa da Andrej Karpathy, l’allievo di Geoffrey e Fei-Fei che, come abbiamo visto, passa da OpenAI a Tesla per poi tornare a OpenAI, e infine lasciare quest’ultima a inizio 2024.

In quest’opera di trasparenza selettiva, OpenAI afferma che la non profit ha avuto “meno di 45 milioni da Elon e più di 90 milioni da altri donatori”, ma soprattutto dice che l’aggettivo “aperto” ha senso solo in riferimento alla diffusione dei benefici della loro opera (l’intelligenza artificiale generale), su cui è OpenAI a giudicare. Come scrive Ilya a Musk a inizio 2016: “Mentre ci avviciniamo alla costruzione dell’intelligenza artificiale generale, avrà senso cominciare a essere meno aperti. Open, in OpenAI, significa che tutti devono trarre beneficio dai frutti dell’intelligenza artificiale dopo che viene costruita, ma è perfettamente corretto non condividere la scienza”. Anche se, aggiunge Ilya, dire che si è completamente aperti è lo stesso una giusta strategia per attrarre le persone. “Cambiate il nome in ClosedAI e ritirerò la causa”. Il nome non viene cambiato ma Musk ritira lo stesso la causa. E poi ne intenta un’altra ancora, perché questo è il suo modo di operare, mentre infiamma la vicenda, e tutto il resto, attraverso la sua piattaforma.

3. Nani sulle spalle di Lewis Strauss

Nel 2016, dopo che Jensen consegna la sua infrastruttura a Elon Musk negli uffici di OpenAI, Google continua i suoi sforzi per mantenersi al centro della ricerca e dello sviluppo sull'intelligenza artificiale. Per rafforzare le sue attività cloud, assume due scienziate: Fei-Fei Li, che si prende un anno sabbatico dall'Università di Stanford, e Jia Li, sua dottoranda dal 2006 al 2011, interessata soprattutto alle implicazioni mediche della visione artificiale.

Gli anni dieci sanciscono una divergenza di interessi tra le aziende tecnologiche statunitensi sulla questione cinese; Apple è senz'altro la più coinvolta in Cina, il luogo fondamentale e al momento insostituibile della sua supply chain: senza la Cina non c'è l'enorme ricchezza di Apple, non c'è il momento iPhone. Un momento paradigmatico è quello dell'uscita di Google dalla Cina, nel 2010: dopo una lunga battaglia col leader cinese del mercato delle ricerche, Baidu, Google è vittima di un attacco hacker, Operation Aurora, che colpisce sia la proprietà intellettuale dell'azienda che gli account di attivisti di diritti umani cinesi. Segue un inasprimento dei rapporti col Partito Comunista, che porta Google ad abbandonare di fatto la Cina per continuare a operare attraverso Hong Kong.

La Cina ospita molti dei massimi esperti mondiali di intelligenza artificiale (AI) e machine learning. Tutti i gruppi vincitori della sfida di ImageNet negli ultimi tre anni erano in gran parte composti da ricercatori cinesi. Nel 2015 gli autori cinesi hanno contribuito al 43% dei contenuti delle 100 migliori riviste sull'intelligenza artificiale e, quando l'Associazione per l'avanzamento dell'intelligenza artificiale ha scoperto che il loro incontro annuale coincideva con il Capodanno cinese di quest'anno, hanno cambiato data".

Quali sono i confini della collaborazione di Google col suo governo? È l'altra domanda che, in parallelo, intrappola l'avventura di Fei-Fei Li con l'azienda. L'uscita dalla caverna dei pixel avviene attraverso il riconoscimento di gattini, che ci fanno sorridere e divertire. Ecco un gatto avvolto da una coperta! Ecco un gatto rannicchiato vicino a un albero! La macchina distingue le ombre, si fa largo tra le ambiguità, e alla fine riconosce che si tratta di un gatto. Immaginiamo Lewis Strauss mentre gli scienziati gli mostrano una schiera di gatti. Il privilegio di essere gli Stati Uniti d'America, per cui nessun sacrificio è troppo grande, si basa forse solo su questo? Non sarà utile avere uno strumento che, così come riconosce i gatti, è in grado di monitorare e controllare i pericoli, i nemici dell'America? Questo è ciò che pensa Jack Shanahan, alto ufficiale dell'aeronautica che nel 2015, come direttore dell'intelligence della difesa per il supporto al combattimento, si trova letteralmente inondato da una marea di dati ottenuti dalle risorse di sorveglianza degli Stati Uniti (come aerei spia, droni, satelliti). Come orientarsi? Shanahan ne parla con Eric Schmidt ed elabora un primo progetto pilota di apprendimento supervisionato su alcuni dati limitati, una sorta di "AlexNet per il campo di battaglia".

È l'inizio del cosiddetto progetto Maven, formalmente noto come Algorithmic Warfare Cross-Functional Team, che viene avviato nell'aprile 2017 dal vicesegretario alla Difesa, Robert Work, proprio per gestire l'enorme mole di dati attraverso l'intelligenza artificiale e ottenere indicazioni operative. L'articolo del "New York Times" che approfondisce le proteste dei dipendenti di Google sul progetto mette al centro proprio la figura di Fei-Fei, rivelando una sua mail in cui, interpellata su come gestire pubblicamente Maven, da un lato sostiene che il contratto è una vittoria per la piattaforma Cloud di Google, e dall'altro lato suggerisce di "evitare a ogni costo qualunque menzione o implicazione dell'intelligenza artificiale", perché teme che la narrazione dell'intelligenza artificiale umanistica, da lei avanzata, venga colpita dall'uso militare, il tema più sensibile in assoluto per il pubblico. Ma non c'è scampo: il progetto Maven riguarda, con ogni evidenza, l'intelligenza artificiale. Ciò che segue è la protesta di migliaia di ricercatori di Google, che chiedono all'amministratore delegato l'annullamento di questo piccolo contratto, per tenersi fuori da ogni sviluppo di tecnologie belliche. Alcuni dipendenti si dimettono.

Google abbandona formalmente il progetto Maven, che viene portato avanti sotto diverse etichette da altre aziende, tra cui Microsoft e Amazon Web Services, e in seguito assegnato in buona parte alla National Geospatial-Intelligence Agency, che ha una storica esperienza nel trattamento delle immagini satellitari sensibili. Il progetto Maven è significativo perché alla protesta di Google – che fornisce un modello per proteste successive – corrisponde una forza uguale e contraria: l'accusa di ipocrisia da parte della comunità militare. A svolgere questa tesi è direttamente il vicesegretario alla Difesa Work, che guiderà dal 2019 con Schmidt la National Security Commission on Artificial Intelligence. Work critica i dipendenti di Google sul piano del calcolo rischio-opportunità, perché, sulla base dei pericoli alle vite umane, non vedono che i programmi "possono salvare 500 americani o 500 alleati o 500 civili innocenti dall'essere attaccati". E soprattutto, Work critica l'ipocrisia di Google sulla Cina: "Tutto quello che accade nel centro cinese sull'intelligenza artificiale va al governo cinese e in ultima analisi finisce in mano all'esercito cinese. E non ho sentito nessun dipendente di Google dire: mmh, forse non dovremmo farlo".

Da un lato, il modello cinese di “fusione militare-civile” rende ogni ricerca, effettuata da attori cinesi o da imprese straniere, potenzialmente militare per la Cina. Dall’altro lato, per contrastare questo modello, è richiesto un rinnovamento della “fusione militare-civile” americana, su cui non si può essere neutrali.

Mordecai Richler scrive su “Playboy” nel 1975: L’America pullula di sbarellati che vedono complotti ovunque, ma anche di alacri studiosi dei medesimi, in un campo che si estende dalla destra più estrema alla sinistra più improbabile, dove chiunque spara ad altezza d’uomo teorie che magari dicono tutto e il contrario di tutto, ma hanno almeno un tratto in comune: la certezza con cui vengono enunciate.

Nel 2021, l’anno in cui BlackRock supera i 10.000 miliardi di dollari gestiti, Soros interviene sul “Financial Times” e sul “Wall Street Journal” per dire che BlackRock non ha capito niente e continua a investire sulla Cina, con l’aggravante ipocrita di fondi basati sui cosiddetti criteri ESG (Environmental, Social and corporate Governance), per immettere risorse in un mercato finanziario – quello cinese – dove la “governance” è basata sulle decisioni di una persona. Il linguaggio di Soros verso BlackRock è, per usare un eufemismo, severo. Oltre a suggerire velatamente che Fink voglia aiutare Xi Jinping a cavarsela nella crisi immobiliare, afferma che, siccome Cina e Stati Uniti sono in conflitto mortale, BlackRock mette in pericolo la sicurezza nazionale degli Stati Uniti e delle democrazie mondiali.

Pertanto, Soros invita il Congresso e la SEC a intervenire per proteggere gli investitori americani. I provvedimenti del governo hanno obbligato da poco BlackRock e altri gestori a vendere le partecipazioni nelle aziende di telecomunicazioni cinesi accusate di collaborare con l’esercito, quindi perché non allargare i divieti, visto che tanto c’è un uomo solo al comando e l’esercito cinese è ovunque? Tra la fine del 2023 e l’inizio del 2024, BlackRock chiude il suo più controverso fondo cinese, citando il disinteresse degli investitori, e mette in vendita, a un prezzo inferiore del 30% dei 168 milioni spesi nel 2018, i suoi uffici di Shanghai.

Gina Raimondo è il primo segretario al Commercio a intervenire al Reagan National Defense Forum e, come dice lei stessa, non sarà certo l’ultimo. Raimondo afferma che “la sicurezza nazionale si basa sulla sicurezza economica”. La forza dell’America è alimentata dalla competitività dell’economia e dal “motore dell’innovazione che guida il mondo”. Gina Raimondo si rivolge direttamente alle aziende americane: L’America è leader mondiale nell’intelligenza artificiale. Punto. L’America è leader mondiale nella progettazione di semiconduttori avanzati. Punto. Questo grazie al nostro settore privato. Perché abbiamo grandi innovatori. Ed è anche merito del nostro settore pubblico, che investe in questi campi. Siamo un paio d’anni avanti alla Cina. Non possiamo permettere che ci raggiungano. Non possiamo permetterle di raggiungerci. Quindi negheremo loro la nostra tecnologia più avanzata. So che tra il pubblico ci sono amministratori delegati di aziende produttrici di chip che erano un po’ irritati quando l’ho fatto, perché stavano perdendo ricavi: proprio come la vita, la protezione della nostra sicurezza nazionale è più importante dei ricavi di breve termine.

Questa roba – e con “questa roba” intendo supercomputer, tecnologia di intelligenza artificiale, chip per l’intelligenza artificiale – nelle mani sbagliate è letale quanto qualsiasi arma che potremmo fornire. Perciò dobbiamo essere seri se vogliamo affrontare questa minaccia ed essere seri nell’applicazione della legge. L’altra cosa per cui abbiamo bisogno di risorse al Dipartimento del Commercio è l’applicazione della legge. Ogni minuto di ogni giorno, la Cina si sveglia cercando di capire come aggirare i nostri controlli sulle esportazioni.

Quando gli ingegneri del Dipartimento del Commercio che lavorano per Gina Raimondo bussano alla porta di NVIDIA per dire “collaboriamo, lavoriamo per la sicurezza nazionale”, dopo aver ottenuto l’agognato autografo di Dally e Buck, al di là dei sorrisi di facciata, viene loro sempre riservato il trattamento della battuta di Reagan: “Le parole più terribili della lingua inglese sono: Vengo dal governo e sono qui per aiutare”. E quando Dally va nelle università a parlare – e non smette mai di farlo – dice chiaramente come stanno le cose: Anche se ho trascorso molto tempo a Washington per spiegare che impedirci di vendere le schede H100 in Cina è una pessima idea, l’unico vero effetto di questa politica di controllo delle esportazioni è stato far sì che migliaia di programmatori cinesi che scrivevano software per le nostre macchine adesso lo scrivano per le macchine di Huawei e Biren. In sintesi, questo farà male all’industria degli Stati Uniti nel lungo termine senza rallentare per nulla il progresso cinese nell’intelligenza artificiale. Ma alla gente di Washington non piace sentirselo dire.

Serve sicurezza, ma dobbiamo essere molto attenti, perché non possiamo esagerare altrimenti soffochiamo l’innovazione. L’America ha raggiunto la sua posizione di leader grazie all’innovazione e dobbiamo continuare a coltivare questo approccio. L’Europa è molto indietro rispetto a noi, la Cina è ancora indietro. Ancora una volta, è tutto delicato e complicato.

La leader statunitense ribadisce una questione che gli europei sottovalutano sistematicamente, ovvero la percezione di un loro ritardo incolmabile su queste tecnologie da parte di Washington. C'è un punto di incontro del "calcolo parallelo" del governo statunitense e di NVIDIA, nello spazio aperto dal momento iPhone dell'intelligenza artificiale e dalla debolezza finanziaria cinese. Da un lato, come abbiamo visto, più si rivendica il dominio americano e la "distanza abissale", più i lettori cinesi del Problema dei tre corpi si impegnano per mutare l'arretratezza, grazie al potere di mercato, in opportunità di crescita: ne conseguono innovazioni continue nella zona grigia tra il lecito e l'illecito, dalle preziose macchine di Applied Materials che arrivano a SMIC attraverso la Corea del Sud, alla collaborazione con le monarchie del Golfo, ansiose di convertire il petrolio in GPU.

Il fondo di Abu Dhabi sull'intelligenza artificiale, G42, guidato dall'amministratore delegato Peng Xiao, rassicura gli Stati Uniti e vende le sue quote in aziende cinesi, tra cui ByteDance. Di più: il 15 aprile è direttamente Microsoft ad annunciare una partnership con G42, con un investimento di 1,5 miliardi. "Nelle tecnologie emergenti, non si può essere allo stesso tempo nel campo cinese e nel nostro campo," commenta Gina Raimondo, mentre la stessa Microsoft rivendica il ruolo del governo statunitense nell'affare.

La principessa di Huawei, Meng Wanzhou, mentre le aziende come Baidu iniziano a ordinare nel 2023 i nuovi chip di intelligenza artificiale Ascend, sa di non poter competere con Bill Dally, e non presenta alcun grafico comparativo sulle prestazioni dei suoi prodotti rispetto a NVIDIA. Si limita a dire che "Huawei è impegnata a costruire una solida base di potenza di calcolo in Cina – e una seconda opzione per il mondo". L'orgoglio del piano B.

4. Palantir e la prima colazione

Nel 2023, al World Economic Forum di Davos, Alex Karp risponde così alla domanda sull'origine del nome della sua azienda, Palantir: "Nel Signore degli Anelli c'è un globo che permette alle forze del bene di vedere cosa sta succedendo e di organizzarsi". Il pubblico non obietta su questa sua definizione del Palantir nell'opera di J.R.R. Tolkien, così Karp riprende a discutere del futuro dell'Occidente e del ruolo della tecnologia, frustando le élite della Silicon Valley.

Lo stregone Gandalf spiega agli hobbit che sette pietre furono fabbricate dall'antico elfo Fëanor in un tempo molto lontano. Le pietre sono utilizzate dagli umani "per vedere lontano e trasmettersi i pensieri": insieme, costituiscono una sorta di rete che contribuisce a rafforzare i legami dei loro regni. Nemmeno quest'uso iniziale riguarda quelle che Karp definisce "le forze del bene": quei regni umani in Tolkien non sono il "bene", sono realtà ambigue nei loro comportamenti. Inoltre, per Tolkien essere connessi non è un valore positivo. Proprio con quell'esempio, Gandalf fa capire agli hobbit che il potere di scrutare, di muoversi nello spazio e nel tempo e così accorciare le distanze, è un grande pericolo per chi lo adopera. A esso nessuno è in grado di resistere. Le pietre veggenti non sono estranee al disastro dei regni degli uomini, che le hanno utilizzate per accrescere il proprio benessere e la propria potenza. Agli hobbit, proprio perché rappresentano l'innocenza, questo può essere svelato, ma anche loro devono stare attenti al pericolo che accompagna la volontà di vedere.

Lo stesso "occhio scrutatore di Saruman", che di nascosto si impadronisce di un Palantir quando ormai tutti, in una nuova epoca, hanno dimenticato la sua esistenza, è "intrappolato e ipnotizzato". "È pericoloso per chiunque servirsi degli artifizi di un'arte di cui non sappiamo scandagliare gli abissi". Chi cerca negli abissi vuole vedere il tutto. Chi vuole conoscere qualcosa troverà sempre un occhio che lo guarda, e che lo mette a nudo. Nel mondo di Tolkien è l'Occhio del Male, l'Occhio del Signore Oscuro di Mordor, a cui non si può sfuggire. Così, la pietra veggente accompagna la rovina dello stregone Saruman. Ti conosco", "ho letto la tua mente", "ho appreso": tutti i termini con cui Denethor si rivolge a Gandalf, nel dialogo decisivo, mostrano il senso profondo del Palantir per Tolkien. La rivendicazione della conoscenza, la conoscenza come eccesso di dati, è senz'altro qualcosa di ambiguo, ed è la strada più breve verso la schiavitù del male. Sapere troppo, rimuginare sui dati della nostra esperienza, vuol dire diventare schiavi del potere e, alla fine, incontrare sempre l'Occhio che ti vede dall'altra parte: l'Occhio di Sauron. Pochissimi possiedono la solidità necessaria per sopravvivere a questa prova. Per resistere, non bisogna mettere il sapere al primo posto, altrimenti non si sarà mai soddisfatti.

Alla Stanford Law School degli anni novanta avviene l'incontro fra Peter Andreas Thiel e Alex Karp. I due studenti cementano un'amicizia basata su comuni interessi intellettuali. Negli anni dell'università, tra gli studi di filosofia e diritto, Thiel si forma nel clima delle guerre culturali degli Stati Uniti, di cui diviene precoce protagonista. A vent'anni, nel 1987, fonda "The Stanford Review", rivista studentesca conservatrice e provocatoria, che critica con violenza il dominio della sinistra sulle politiche universitarie. Essenziale testimonianza di questa fase militante e decisiva della vita di Thiel, è un libro del 1995 che pubblica con David Sacks, contro "il mito della diversità".

Nel 1998, co-fonda Fieldlink, poi rinominata Cofinity, una società di pagamenti online che, dopo la fusione con X.com di Elon Musk nel 2000, prende il nome di PayPal e viene acquisita da eBay nel 2002 per 1,5 miliardi di dollari. Sono risorse importanti per i personaggi più ambiziosi di quel gruppo (noto come "PayPal mafia"), Musk e Thiel. Mentre Musk inizia a dare forma al suo impero elettro-spaziale, nel 2004 Thiel investe 500.000 dollari in Facebook per una quota del 10,2%, diventando il primo investitore esterno della società. L'investimento si rivela ovviamente molto redditizio, e Thiel diviene un mentore di Mark Zuckerberg, nonché membro del consiglio di amministrazione di Facebook/Meta fino al 2022. Oltre a Facebook, Thiel ha investito in numerose altre start-up tecnologiche attraverso Founders Fund, un fondo di venture capital che ha co-fondato nel 2005 e che investe su aziende come SpaceX, LinkedIn, Yelp e, come abbiamo visto, DeepMind. Inoltre, Thiel partecipa all'acceleratore di start-up Y Combinator, dove si forma Sam Altman.

La consacrazione dello spirito provocatorio di Thiel avviene durante la campagna presidenziale del 2016, col suo sostegno convinto e militante a Donald Trump. Dopo l'elezione di Trump, Thiel non ha grande influenza, e nel corso del tempo abbandona la scialuppa, senza però abdicare da una più ampia volontà di influenza politica. Durante l'amministrazione Trump, il suo vero obiettivo è la preparazione della quotazione di Palantir. Palantir nasce dall'esperienza di PayPal, da ciò che Thiel ha "visto" guardando nell'intreccio di due sfere: la decentralizzazione dei pagamenti e la centralizzazione del controllo.

Da un lato, il sistema dei pagamenti online si alimenta nella controcultura che, in ottica libertaria, cerca di superare i controlli sulle transazioni e da ultimo il controllo per eccellenza, quello della moneta. Dall'altro lato, far funzionare in modo efficace una piattaforma sui pagamenti significa trovare sistemi per valutare le minacce, anzitutto di contraffazione e cybercriminalità, e rispondervi con efficacia. Il Palantir di PayPal, in questo senso, è la visione delle minacce attraverso l'analisi dei dati, che porta con sé una "maledizione" del controllo che indebolisce la spinta verso la decentralizzazione.

La scommessa di Thiel e Karp avviene dopo l'11 settembre, scintilla per la costruzione di uno "stato della sicurezza nazionale" del XXI secolo, matrimonio tra l'aumento della capacità di analisi dei dati e l'allargamento della sicurezza nazionale. Un processo che richiede la trasformazione degli apparati dello stato. Perché accade tutto questo? Con gli anni novanta, finisce la breve stagione illusoria della sicurezza occidentale. Nel 1993, una famosa cena dell'allora segretario alla Difesa William Perry con le principali industrie militari passa alla storia come "l'ultima cena": col "dividendo della pace" ci saranno meno soldi. Nel 1994, come abbiamo visto, Danny Hillis può lamentarsi per la fine della guerra fredda perché perde i fondi che tengono in piedi il suo progetto di supercalcolo, che non sa vivere sul mercato. Se si considera l'11 settembre 2001 come fine della fine della guerra fredda, si comprende quanto quella parentesi sia stata breve.

Se non possiamo dare la sicurezza per scontata, dobbiamo investire per essere sicuri, e dobbiamo aggiornare il nostro investimento sulla base delle minacce. Se questa tesi è vera – e lo è –, allora Karp e Thiel hanno davanti un mercato. Come nel caso dei videogiochi di NVIDIA, quel mercato ancora non esiste, ma è essenziale agire per motivare la comunità degli utenti. Perciò, Thiel e Karp devono convincere il cliente, in questo caso non un anonimo videogiocatore ma il governo degli Stati Uniti, che il suo lavoro di sicurezza non può essere solo umano, e che non deve limitarsi a comprare sensori, e in generale dispositivi di captazione, ma deve concentrarsi sull'analisi dei dati, pena l'incapacità di "pensare" il dato stesso. La tesi si corrobora se nel nuovo secolo gli Stati Uniti dovranno stare sempre in stato di guerra e se vi saranno sempre minacce di ordine tecnologico. L'effetto di questo processo non può che essere la "grande trasformazione" del complesso militare-industriale.

Si inserisce in alcune operazioni controverse dell'amministrazione Trump, per esempio il muro al confine col Messico. Ovviamente, non è compito di Palantir la costruzione letterale di un muro, bensì l'organizzazione di uno smart wall, cioè di un software per profilare gli immigrati. Strutturale, in Palantir, è la sfacciataggine con cui Thiel rivendica il rapporto con la sicurezza nazionale americana. In un AMA (Ask Me Anything) su Reddit, cioè un confronto con la comunità di utenti in cui un personaggio si mette a disposizione per rispondere a varie curiosità, Thiel davanti alla domanda secca "Palantir è una copertura per la CIA?", ribatte: "No, la CIA è una copertura per Palantir".

Durante l'amministrazione Trump, Palantir realizza la quotazione in borsa a lungo rimandata. Nella prima giornata di contrattazioni, a settembre 2020, il titolo chiude con circa 20 miliardi di capitalizzazione, e a febbraio 2024 la capitalizzazione è di circa 50 miliardi, tornata ai valori del 2021 dopo una discesa consistente tra 2022 e 2023. Quando il valore di borsa di Palantir scende, i fondatori mantengono la loro presa sull'azienda, anche per via delle loro azioni privilegiate, di classe F. Il sistema è progettato per concentrare il controllo decisionale nelle mani di Peter

Thiel, Alex Karp e Stephen Cohen, garantendo loro il 49,999999% del potere di voto in perpetuo, indipendentemente dalla proprietà effettiva delle azioni. Questa struttura, chiamata dai suoi critici di "imperatore aziendale a vita", amplifica una tendenza tipica della Silicon Valley che abbiamo già analizzato, e la applica al caso estremo di chi vuole cambiare e costruire il mercato per la difesa e la sicurezza, al riparo da influenze esterne.

Palantir può contare su una pattuglia di attivismo politico, perché Thiel, come consentito dal sistema, può finanziare le campagne elettorali di personalità a lui vicine. Il principale esempio è J.D. Vance, che, dopo aver lavorato con Thiel nelle sue società di investimenti, diviene celebre per il libro autobiografico in cui racconta la difficoltà di crescere in Ohio (uno degli Stati più politicamente rilevanti, segnato dal degrado sociale e dall'uso crescente di droghe), e si candida con successo nel 2022 a rappresentare lo stato in Senato. Nel 2024, Vance viene scelto da Donald Trump come candidato vicepresidente. Inoltre Palantir, come le altre aziende tecnologiche, fa uso frequente di porte girevoli, cioè di reclutamento di figure politiche e amministrative per aumentare le sue capacità di mercato, dall'ex cancelliere austriaco Sebastian Kurz al politico australiano Mike Kelly, ai dirigenti del National Health Service del Regno Unito.

Sul piano pubblico, Palantir, così come vuole richiamare l'alone di mistero attraverso l'investimento della CIA, esalta i suoi risultati – ottenuti con attività che non possono essere rese pienamente trasparenti, ma che hanno comunque salvaguardato la sicurezza – attraverso una precisa narrazione. Secondo Karp, "Palantir ha, almeno nel Regno Unito e negli Stati Uniti, cambiato la traiettoria della pandemia con l'organizzazione della distribuzione di mascherine, ossigeno, vaccini". Palantir all'inizio si è mossa in un deserto, dove la difesa non aveva ancora corpi specializzati nell'innovazione e i fondi di venture capital non avevano capito le opportunità delle tecnologie per la sicurezza: "In-Q-Tel era stata appena fondata (1999) per concentrarsi sugli investimenti in nuove capacità tecnologiche per la comunità dell'intelligence (e meno male, altrimenti Palantir probabilmente non esisterebbe oggi, e In-Q-Tel non avrebbe avuto un ritorno sull'investimento più di cento volte maggiore, per finanziare un numero ancora maggiore di attori nell'ecosistema)".

In questo senso, nella storia (ri)scritta da Palantir, la "prima colazione" è l'innovazione del complesso militare-industriale occidentale, guidato dagli Stati Uniti e reso possibile da talenti con una visione non neutrale, bensì patriottica, della tecnologia. Come afferma Karp: "Vogliamo persone che vogliano stare dal lato dell'Occidente, che vogliano rendere l'Occidente una società migliore, più capace di difendersi". L'epoca in cui viviamo, secondo Karp, è un "momento Oppenheimer", in cui i magnati della Silicon Valley non possono dimenticare che esistono grazie alla nazione che li ha resi possibili, e quindi alla sua infrastruttura militare. Karp richiede "una collaborazione più intima tra lo stato e il settore tecnologico", con un maggiore allineamento dei loro interessi "per porre vincoli ai nostri avversari nel lungo termine". Il mantra di Palantir è: "Le precondizioni per una pace duratura spesso vengono solo da una credibile minaccia di guerra".

L'equazione tra minaccia e realtà della guerra è l'ambiente in cui Palantir può ampliare il suo mercato di riferimento; il pianeta diviene resiliente e sostenibile attraverso l'allargamento della monarchia tecnologica degli Stati Uniti, con un nuovo matrimonio tra difesa e tecnologia. Palantir avverte che "l'ecosistema tecnologico della difesa oggi non ha lo stesso lusso del tempo concesso dalla Seconda guerra mondiale […]. Dobbiamo ancora una volta mobilitare la nostra base industriale in un modo nuovo, ma in questo caso riconoscendo che l'acceleratore è il software". Peter Thiel, "intellettuale" interessato di questa mobilitazione, chiarisce l'importanza di individuare gli avversari, altrimenti la "prima colazione" non può diventare un appuntamento regolare.

Thiel ha reso celebre la formula: "Volevamo le auto volanti, invece abbiamo avuto i 140 caratteri", per paragonare la riduzione delle aspettative sulla tecnologia. Nel libro Zero to One, Thiel cita la legge di Moore al contrario (la "legge di Eroom"), secondo cui il numero di nuovi farmaci approvati dalla Food and Drug Administration (FDA) statunitense per miliardo di dollari di investimenti industriali in ricerca e sviluppo adeguati all'inflazione è diminuito di circa cento volte dal 1950 al 2010. Per Thiel lo stato delle biotecnologie, come quello dei trasporti, dell'energia, dello spazio, mostra che l'innovazione si è rifugiata quasi esclusivamente nell'informatica.

La sua osservazione del 2013, durante un dibattito con Marc Andreessen: "Abbiamo una grande computer rust belt [cintura di ruggine dei computer] e a nessuno piace parlarne. Sono aziende come Cisco, Dell, Hewlett Packard, Oracle e IBM. Penso che il modello sarà quello di diventare merci senza avere più capacità di innovare. Ci sono molte aziende che sono su questo crinale. Microsoft è probabilmente vicina alla computer rust belt. L'altra azienda che si trova su questo crinale, sorprendentemente, è Apple". Thiel commette l'errore comune di sottovalutare l'hardware, e crederlo inevitabilmente incapace di creare valore: tesi rigettata dalla crescita dei data center.

Vi è una possibile distinzione tra “intelligenza artificiale” e “intelligenza aumentata”, ovvero tra sistemi in grado di operare in effettiva autonomia, di eseguire compiti e prendere decisioni senza l’intervento umano, e sistemi progettati per estendere e potenziare le capacità umane, piuttosto che sostituirle. Questa “intelligenza aumentata”, volta a conservare e sfruttare i “vantaggi comparati” degli uomini e delle macchine, è essenziale per l’operato di Palantir. Thiel lo presenta ai suoi studenti insieme a un ospite dell’azienda, il quale afferma: “Sarebbe una pessima idea costruire un’intelligenza artificiale dedicata solo a individuare i terroristi. Bisognerebbe far sì che una macchina pensasse come un terrorista. Probabilmente ci vorrebbero vent’anni per arrivare a qualcosa di simile. Ma i computer sono abili nell’elaborazione dei dati e nella corrispondenza dei modelli. E le persone sono abili nella comprensione dei concetti. Se si mettono insieme questi pezzi, si ottiene l’approccio dell’intelligenza aumentata”.

Infine, Thiel paragona gli investimenti sulle biotecnologie a quelli sull’intelligenza artificiale, evidenziando tre vantaggi di quest’ultima: la libertà ingegneristica, la libertà normativa e le opportunità di un campo ancora sottoesplorato e che va controcorrente. Ricordiamo che tutto questo avviene nell’anno “decisivo”, il 2012. Thiel stesso ha già investito in DeepMind, e Fei-Fei Li è da poco arrivata a Stanford con ImageNet, su invito di Bill Dally, nello stesso dipartimento dove Thiel tiene il suo corso. Nel giro di poche settimane, AlexNet segnerà il big bang dell’intelligenza artificiale. Nel 2019, Thiel ritorna sull’incontro con Demis nel lontano 2010 e sul suo “progetto Manhattan”. Quest’immagine, scrive Thiel, conta per come viene letta nelle “capitali straniere”.

DeepMind non si è avvicinata alla formula di Shane, l’intelligenza artificiale generale, “mentre sta finalmente diventando chiaro che, come nel caso della fissione nucleare, i primi utenti degli strumenti di apprendimento delle macchine creati oggi saranno i generali piuttosto che gli strateghi dei giochi da tavolo. L’intelligenza artificiale è una tecnologia militare”. Thiel procede a descrivere gli usi della tecnologia su cui si fonda il modello di business di Palantir, poi concede che, ovviamente, gli strumenti possono essere anche utilizzati in ambito civile, e quindi sono classicamente duali. TikTok, spauracchio di Thiel – anche perché concorrente di Facebook –, è “un’arma della Cina comunista” usata per accentuare debolezze (come la criminalità giovanile e la diffusione delle droghe) già individuate dal principe dei consiglieri del Partito Comunista, Wang Huning, nel libro del 1991 America Against America.

“Dietro la competizione tecnologica c’è quella politica. […] Per superare gli americani, bisogna fare una cosa: sorpassarli nella scienza e nella tecnologia.” Agli occhi di Wang Huning, nel 1991 gli statunitensi non possono sostenere il pensiero di perdere, con “un senso di superiorità tecnologica che gradualmente è divenuto la sensazione di superiorità nazionale”. Anche l’idea “che qualunque nazione possa superarli per loro è inimmaginabile”, il successo economico del Giappone ha rappresentato un risveglio e, aggiunge il principe dei consiglieri, “credo che gli americani incontreranno di nuovo una situazione simile”. Quando ciò accadrà, potrà essere utile sfruttare una consapevolezza: “A volte non è l’uomo che controlla la tecnologia, ma la tecnologia che controlla l’uomo”. Anche in quest’ambito, la tesi di Thiel è vincente. La comunità di intelligence degli Stati Uniti ritiene che “la Repubblica Popolare Cinese possa cercare di influenzare le elezioni del 2024 a qualche livello per via del suo desiderio di mettere ai margini i critici della Cina e amplificare le divisioni sociali degli Stati Uniti”. Questa tattica sarà portata avanti attraverso esperimenti con l’intelligenza artificiale generativa e attraverso TikTok.

Marc Andreessen nasce in Iowa nel 1971 e cresce a New Lisbon, un paesino del Wisconsin di poco più di mille abitanti. Pur vivendo in luoghi rurali, lontani dalle direttrici dello sviluppo tecnologico, Andreessen matura un interesse per l’informatica. Quando si trasferisce nella Silicon Valley, sviluppa il browser Netscape, che si quota in borsa nel 1995 e viene venduto ad AOL nel 1999 per oltre 4 miliardi di dollari. Andreessen, nel nuovo secolo, si reinventa come investitore in start-up tecnologiche, e nel 2009 dà vita ad Andreessen Horowitz, che ha investito in società come Airbnb, Facebook, Github, Instagram, Oculus, Slack, Skype, e oggi gestisce 35 miliardi di dollari nei suoi fondi. Andreessen sta descrivendo CUDA, senza sapere nel dettaglio che la stessa NVIDIA sta già elaborando quelle soluzioni, ha già messo al lavoro Buck e Dally per essere padrona della “distruzione creatrice” e distruggere o sottomettere, dall’interno, le start-up in cui lui sta investendo.

D’altra parte, questo processo è lungo e non immediatamente riconoscibile. Dal 2010 al 2012, il titolo di NVIDIA scende e l’azienda, come spesso rivendicato da Jensen, affronta la sua traversata nel deserto per via della scommessa di CUDA e dei nuovi mercati ipotetici. Nel 2016, Andreessen riconosce invece in modo esplicito la centralità di NVIDIA: “Abbiamo investito in molte start-up che applicano il deep learning a vari ambiti, e tutte si basano sulla piattaforma di NVIDIA. È come quando tutti si basavano su Windows negli anni novanta o sull’iPhone alla fine degli anni 2000. Per divertimento la nostra azienda ha un gioco interno su quali società quotate investire, se fossimo un hedge fund. Metteremmo tutti i nostri soldi in NVIDIA”. In questo modo, Jensen completa la profezia di Andreessen: il software sta mangiando il mondo, l’intelligenza artificiale mangerà il software, e questo stesso

processo si basa sull'hardware programmabile e personalizzabile di NVIDIA, offerto con complementi e servizi che lo rendono sempre più indispensabile. "L'hardware non è mai stato così importante."

Andreessen, con cui dibatte Peter Thiel nel 2013 durante la sua sfortunata previsione su Microsoft, ha invece una pretesa diversa. Anche lui vuole essere non solo attore, ma pensatore di questo processo. Nel 2023 pubblica un "manifesto tecno-ottimista": l'accumulazione di articoli di fede che evidenziano il valore positivo di ogni sviluppo tecnologico, introdotti dalla formula "noi crediamo", che nel manifesto appare volte. I tecno-ottimisti, secondo Andreessen, credono in due elementi principali: la continua combinazione fra tecnologia e mercati, che chiama "macchina del tecno-capitale"; e nell'accelerazionismo, che è "la propulsione cosciente e deliberata dello sviluppo tecnologico" – per assicurare il compimento della legge dei ritorni acceleranti. Credono poi, tra l'altro, che l'intelligenza artificiale sia "l'alchimia, la pietra filosofale – perché stiamo letteralmente facendo pensare la sabbia". Credono che "ogni decelerazione dell'intelligenza artificiale costerà vite" e che "non ci sia conflitto tra la macchina del tecno-capitale e l'ambiente naturale". E credono che "l'America e i suoi alleati dovrebbero essere forti e non deboli", con una forza economica, culturale e militare che deriva da quella tecnologica.

Il lettore del manifesto tecno-ottimista può sentirsi vittima delle app nate per condensare i libri in poche sentenze, al fine di sembrare intelligenti a cena (un'evoluzione della lettura veloce di Guerra e pace secondo Woody Allen, compressa nelle tre parole "Riguarda la Russia"). Per questo, il tecno-ottimismo richiede il complemento di uno scritto di Carlo Emilio Gadda, dedicato a L'uomo e la macchina (1940). Gadda inizia ricordando il motore dell'età delle macchine, il grido "Vogliamo il Prodotto!" pronunciato da "Moltitudini dai mille sensi e dai mille appetiti". Il Prodotto toglie la fame e la sete, toglie di dosso il freddo, cancella la paura delle notti.

Con la Repubblica Popolare Cinese ci troviamo in una competizione persistente e generazionale per il vantaggio, e dobbiamo insistere con urgenza e fiducia." Alle start-up del dinamismo americano, Hicks dice: "Sapete qual è l'alternativa. Sapete chi volete che vinca. Sono felice che abbiate scelto il team USA. Perché il Dipartimento della Difesa ha bisogno di voi, con noi". Il cambio di paradigma di questa fase del conflitto tra Stati Uniti e Cina è che sempre più aziende vengono allo scoperto, con la certificazione del lavoro con gli apparati statunitensi, e questo fa parte della loro tesi di investimento, seguendo la strada inaugurata da Palantir.

Questa strada è battuta anche da Palmer Luckey, rigorosamente in infradito. Luckey è un investitore e imprenditore nato nel 1992, qualche mese prima di NVIDIA. Indossa quasi sempre una camicia hawaiana (ne possiede circa settanta). Fin da bambino ha imparato a smontare, rimontare e adattare dispositivi elettronici e informatici. A sedici anni in un furgoncino costruisce visori di realtà virtuale che attirano l'attenzione, tra gli altri, del leggendario programmatore di videogiochi della id Software, John Carmack, la mente di serie sparatutto come Doom e Quake che negli anni novanta hanno segnato i primi passi di NVIDIA.

I visori portano alla creazione di Oculus, sostenuta da Peter Thiel e Marc Andreessen. Carmack diventa direttore delle tecnologie di Oculus: durante una convention sui videogiochi di NVIDIA in Canada, nel 2013, promuove i prodotti di Oculus, che viene acquisita da Facebook nel 2014 per due miliardi di dollari. Luckey continua a lavorare nell'impero di Zuckerberg, fino a quando una sua donazione a un gruppo che promuove meme offensivi contro Hillary Clinton porta nel 2016 alla sua uscita, con una coda di polemiche nella Silicon Valley che discute animatamente del legame tra Thiel e Trump. Come Thiel, Luckey cita Norman Angell per spiegare che l'interdipendenza economica non genera pace e che la realtà in cui dobbiamo vivere è armata, dove armi sempre più sofisticate e intelligenti alimentano la deterrenza, per parlare con i forti il linguaggio della forza.

Mentre fomenta il pubblico attraverso la creazione del Game Boy perfetto, Luckey – che possiede la più grande collezione di videogiochi al mondo, conservata in una ex base di missili nucleari – parla per ore degli investimenti in ricerca e sviluppo della difesa statunitense, della storia dei sistemi d'arma e del dominio sotterraneo dei conflitti, dei teatri delle guerre in Europa e nel Medio Oriente, dei rischi su Taiwan, dell'importanza delle Filippine e del Vietnam. E di come l'America non debba avere paura del talento cinese, ma debba anzi costruire programmi sempre più ambiziosi di attrazione e di "esfiltrazione" dei ricercatori di Pechino. La debolezza dell'America, secondo Luckey, sta nella sua industria della difesa, nei contractor tradizionali. I suoi nemici sono Boeing, Lockheed Martin, Raytheon e le altre aziende del "capitalismo militare": un oligopolio che non funziona più perché non produce innovazione e che quindi, nel nuovo capitalismo politico statunitense, deve essere distrutto per non soccombere davanti alla Cina. E fare spazio al giovane oligopolio della prima colazione.

La visione di Luckey, che sviluppa il pensiero di Thiel e il dinamismo americano, può essere così sintetizzata: "Gli uomini in giacca e cravatta, gli incapaci che sottovalutano i videogiochi e mi prendono in giro per le camicie

hawaiane, sono i parassiti dell'America. Le loro cravatte ci strozzano. Ci hanno portato a pagare i russi per portare gli americani sulla Stazione spaziale internazionale, fanno cadere gli aerei, sono pieni di fornitori cinesi che li tengono in pugno. O prendiamo il potere noi pazzi, noi disadattati, noi ribelli, noi casinisti, oppure vincono i cinesi. Dobbiamo prendere il potere perché abbiamo ragione, però, per non sbagliare, siccome siamo miliardari pagheremo la politica ed eserciteremo così la nostra libertà di espressione: dirigere l'arsenale della democrazia, che gli altri hanno svuotato, è il nostro dovere patriottico". Non a caso, Luckey identifica la sua missione con una parola: "arsenale".

Nell'estate 2023, pochi mesi dopo essere stato sanzionato dalla Cina per la fornitura di armi a Taiwan, l'amministratore delegato di Raytheon Technologies, Greg Hayes, chiede pubblicamente al governo americano di trovare un modus vivendi con la Cina. La separazione dalla Cina, a suo avviso, è "impossibile". Hayes aggiunge che in Cina la sua azienda, cioè un tassello fondamentale del sistema della difesa statunitense nato dall'"ultima cena", ha "varie migliaia di fornitori". Hayes, amministratore delegato di un'azienda che è stata fondata dall'autore della "frontiera infinita", Vannevar Bush, e che fornisce i missili alle forze armate degli Stati Uniti, ci tiene a ricordare che dipende dall'avversario cinese.

Il potenziale vantaggio nell'industrializzazione, grazie alla possibilità di saltare le fasi iniziali dello sviluppo industriale e adottare direttamente le tecnologie e i metodi organizzativi più avanzati. Un simile salto tecnologico, correttamente praticato attraverso le istituzioni finanziarie, la scala industriale e la qualificazione della manodopera, può permettere rapidi progressi economici. Gerschenkron stesso riconosce i limiti del "vantaggio dell'arretratezza" (e usa con parsimonia questa formula), anche perché è evidente che, se l'arretratezza è eccessiva, si trasforma in una trappola, ma la sua teoria è rimasta influente nello sviluppo industriale dell'Asia orientale nella seconda metà del Novecento, prima con la crescita giapponese, poi con le cosiddette quattro tigri asiatiche (Taiwan, Corea del Sud, Singapore, Hong Kong), e infine con l'ascesa cinese.

Gerschenkron rimane un riferimento negli studi economici cinesi importanti per il Partito Comunista: per esempio, è sovente citato nelle opere di Justin Yifu Lin, professore della Peking University che è stato capo economista e vicepresidente senior della Banca Mondiale dal 2008 al 2012, e che ora è una voce influente sulle politiche economiche cinesi. Pechino non persegue un disegno onnicomprensivo di sostituzione di capacità industriale di base, o se vogliamo "arretrata", con industrie "di frontiera". La sua scommessa è, appunto, un altro "vantaggio dell'arretratezza": il mantenimento di una capacità produttiva soverchiante in alcuni ambiti industriali, che può essere utilizzata, assieme alle scorte materiali e a una capillare industria chimica e della trasformazione, per rendere dipendenti e vulnerabili altre economie, a partire dall'avversario strategico statunitense. Il surplus produttivo garantisce il mantenimento di un ruolo centrale nel prezzo di alcuni beni, e per questo vantaggio politico e tattico il sistema cinese è disposto a pagare, se necessario, costi economici considerevoli.

Lo scopo è evitare che l'avversario possa riprendere il ruolo di "arsenale". L'accelerazione può alimentare l'idea astratta che il software possa mangiare il mondo da solo. Invece, il software ha comunque bisogno dell'hardware, che costruisce, in ogni fabbrica o "mulino satanico", la materia del mondo. Per riprendere i termini di Gadda, non basta essere "celebratori del futuro". Bisogna "sapere, dal di dentro, come è fatta" la propria filiera. Con tutti i pezzi che la compongono. Il "vantaggio dell'arretratezza" cinese è anche una scommessa sulla pigrizia e sull'inefficacia degli altri: gli altri annunceranno, proclameranno, ma non sapranno realizzare, non faranno le cose in tempo, non apriranno miniere, espelleranno le aziende chimiche per ragioni ambientali, non concederanno permessi. E così via.

C'è un altro punto debole. Se gli Stati Uniti colpiscono la cosiddetta "frontiera" dei semiconduttori e poi la microelettronica relativa a quasi tutti gli armamenti e le applicazioni industriali, comprese le piattaforme militari, viene considerata una parte "arretrata" della filiera, allora la Cina si specializza ancora di più in quegli ambiti, costruendo in questo modo altre dipendenze. Quando gli Stati Uniti si svegliano e allargano il cortile e la recinzione di cui parla Jake Sullivan, è troppo tardi o è troppo complicato, perché le operazioni su cortili e recinzioni sono pratiche, richiedono persone disposte a svolgerle, capitali da portare da un obiettivo all'altro: si scontrano con la realtà di imprese quotate che assicurano di essere superoccidentali e poi si lamentano ogni giorno per la perdita di quote del mercato cinese, e non c'è un test di Turing o di Suleyman che sia stato superato affinché ciò avvenga in automatico e faccia national security design automation.

5. Il lungo periodo e l'intelligenza sovrana

Le tre A: auto, acciaio, aglio. La sicurezza nazionale degli Stati Uniti è un'arma potente, cuore della compenetrazione tra economia e politica del nostro tempo, il capitalismo politico. L'agitazione dell'arma della sicurezza nazionale, davanti alla questione cinese, l'ha resa ormai simile alla "notte in cui tutte le vacche sono nere" criticata da Hegel: l'unità indifferenziata in cui, senza distinzioni, non c'è esperienza della coscienza. E dunque non può esserci scienza, azione, trasformazione del mondo. È facile illustrare con pochi esempi questa "notte in cui tutte le vacche cinesi sono nere".

Il 29 febbraio 2024 il presidente Biden ha applicato la sicurezza nazionale all'industria automobilistica. Biden ha affermato: "Le Big Three e i lavoratori dell'automobile guidano il mondo della qualità e dell'innovazione. Un'industria automobilistica dinamica è vitale per l'economia degli Stati Uniti". La sicurezza nazionale statunitense è messa in pericolo dalle automobili cinesi, perché esse sono "connesse" e quindi "possono ottenere dati sensibili". Perciò, per ragioni di sicurezza nazionale, il governo deve continuare ad analizzare lo sviluppo del settore, e prendere provvedimenti.

L'allargamento della sicurezza nazionale incide così sulle acquisizioni degli alleati per eccellenza contro la Cina, i giapponesi, subordinando alla politica interna la logica del cosiddetto friendshoring, la riorganizzazione delle supply chain industriali e tecnologiche attraverso un sistema di alleanze. Nippon Steel deve attendere e assumere intanto parecchi lobbisti, tra cui l'ex segretario di stato Mike Pompeo. Anche l'attenzione è un'arma. Punto che non va mai dimenticato, nel conflitto tra Stati Uniti e Cina. I controlli sulle esportazioni sono misure che spostano e divergono l'attenzione sulla risposta ai problemi dell'economia e della società cinese, mentre il tentativo impossibile di rendere l'America "una grossa fabbrica", non per un solo grande obiettivo, ma per una molteplicità di scopi, emerge come un'altra deviazione, per chi deve rispondere a problemi che non si riducono solo alla competizione tecnologica.

Geoffrey attribuisce la sua "conversione" a due "fatti": l'esposizione al modello linguistico PaLM di Google, che l'ha sorpreso per la capacità di spiegare una battuta, cosa che non credeva sarebbe stato in grado di fare, e, soprattutto, il rovesciamento della sua storica convinzione sui limiti della retro-propagazione e di altri strumenti rispetto al cervello umano. Nelle sue conferenze, Geoffrey paragona le connessioni di GPT-4, i pesi, stimati a circa 2000 miliardi, a quelli del cervello umano, le sinapsi, stimate a circa 100.000 miliardi. Pur avendo meno connessioni, il modello linguistico "ha molta più conoscenza", perché "ha visto enormemente più dati rispetto a quelli che ogni persona potrebbe vedere". In questi termini, "il calcolo biologico richiede molta meno energia, ma è molto peggio per la condivisione della conoscenza". Nei modelli linguistici di grandi dimensioni, ogni agente individuale cerca di imitare quello che dicono le persone, prevedendo una parola dopo l'altra, ma "con la capacità di combinare in modo molto efficiente quello che apprendono".

A suo avviso le reti neurali "diventeranno molto più intelligenti di noi" e poi "prenderanno il controllo", sia per la loro superiorità evolutiva, sia perché hanno appreso la logica del controllo dall'esperienza umana (che Geoffrey riassume con uno sbrigativo "beh, hanno letto tutti i libri di Machiavelli"). Infine: La mia ipotesi è che prenderanno il sopravvento, ci terranno in giro per far funzionare le centrali elettriche, ma non per molto, perché saranno in grado di progettare computer analogici migliori. Saranno molto, molto più intelligenti di quanto le persone siano mai state. Noi siamo solo una fase di passaggio nell'evoluzione dell'intelligenza. Questa è la mia ipotesi su ciò che accadrà. E spero di sbagliarmi.

Nel documentario realizzato nel 2019 da una regista norvegese, iHuman, Ilya descrive la sua visione del futuro in termini non difformi dall'ipotesi di Geoffrey: Le prime intelligenze artificiali generali saranno enormi data center pieni di processori specializzati nelle reti neurali, che lavorano in parallelo e consumano l'energia di dieci milioni di case. […] Le credenze e i desideri delle prime intelligenze artificiali generali saranno decisive; quindi, è importante programmarle correttamente. Credo che, se questo non sarà fatto, allora la natura dell'evoluzione della selezione naturale favorirà i sistemi che danno priorità anzitutto alla propria sopravvivenza.

Non è che le intelligenze artificiali generali "odieranno" gli esseri umani o vorranno far loro del male, ma saranno troppo potenti. Penso che una buona analogia sia il modo con cui gli esseri umani trattano gli animali. Non è che li odiamo. Anzi, li amiamo. Ma quando arriva il momento di costruire un'autostrada tra due città, non chiediamo agli animali il permesso, lo facciamo e basta perché è importante per noi. Penso che questo sia il genere di relazione che ci sarà tra noi e intelligenze artificiali davvero autonome che operano da sole.

Nel lungo periodo saremo tutti morti. Gli economisti si attribuiscono un compito troppo facile e troppo inutile, se, in momenti tempestosi, possono dirci soltanto che, quando l'uragano sarà lontano, l'oceano tornerà tranquillo. La decisione del futuro si compie nel presente, che, in termini agostiniani, è il tempo che mi chiama. Un commento dell'estate 1937 sulla politica estera britannica: in riferimento alla guerra civile spagnola, Keynes si richiama alla poesia di W.H. Auden, Spain, per tornare sul suo concetto di tempo. "Ieri, tutto il passato. Domani, forse il futuro. Ma oggi, la lotta." Per Keynes, in quel caso, la priorità del presente è il mantenimento della pace. A questo proposito scrive: È nostro dovere prolungare la pace, ora dopo ora, giorno dopo giorno, il più a lungo possibile. Non sappiamo cosa porterà il futuro, tranne che sarà molto diverso da qualsiasi cosa potessimo prevedere. In un altro contesto ho detto che è uno svantaggio del "lungo periodo" il fatto che nel lungo periodo saremo tutti morti. Ma avrei potuto dire altrettanto bene che è un grande vantaggio del "breve periodo" il fatto che nel breve periodo siamo ancora vivi. La vita e la storia sono fatte di brevi periodi.

Nella sua denuncia contro OpenAI del 2024, Elon Musk dice, tra le altre cose, che l'intelligenza artificiale generale (definita come "una macchina che ha intelligenza come quella umana per un'ampia varietà di compiti") è stata già raggiunta. Il "grave pericolo per l'umanità", nell'argomento di Musk, sta anche sul piano strettamente economico: "La nostra intera economia si poggia sul fatto che gli umani lavorano insieme e arrivano alle migliori soluzioni per un compito difficile. Se una macchina può risolvere quasi ogni compito meglio di noi, quella macchina diviene più economicamente utile di noi". Questa posizione rappresenta quindi l'opposto della tesi di investimento di OpenAI, espressa da Ilya nel 2018: il superprodotto che, attraverso la scala, diviene sempre più utile e necessario perché sa eseguire meglio dell'uomo ogni suo lavoro.

Musk sostiene che l'accordo tra i fondatori di OpenAI, pur non sostanziato da un contratto, viene tecnicamente rotto non nel 2019 – quando si avvia formalmente la partnership di OpenAI con i "mulini satanici" di Microsoft –, ma nel 2023. Questo perché "GPT-4 è un algoritmo di intelligenza artificiale generale". Per via delle performance del modello, Musk sostiene che esso si trovi fuori dall'accordo di licenza tra OpenAI e Microsoft, che non si applica all'intelligenza artificiale generale. Il procedimento riprende a questo proposito un paper di Microsoft Research, che vede in GPT-4 "scintille di intelligenza artificiale generale", o comunque una sua forma iniziale, che potrà essere seguita da forme più profonde.

Quando ha parlato per la prima volta di "intelligenza artificiale generale", Shane, il matematico neozelandese finito a Lugano grazie ai soldi del Cynar, non poteva immaginare la sequela di processi scatenata dalla sua espressione e dalle previsioni del suo blog. Ormai l'intelligenza artificiale generale si presenta in una doppia forma: la prima è una modalità attraverso cui le aziende aumentano la loro valutazione da parte degli investitori e/o guadagnano, fornendo prodotti ai clienti e ricevendo investimenti e fiducia per ipotetici guadagni successivi; la seconda è il sogno indifferenziato di un calcolo universale, approdo definitivo capace di dischiudere l'ignoto. Nella sua seconda forma, l'intelligenza artificiale generale è, e resta, come altri elementi della nostra civiltà, un'ossessiva forma di secolarizzazione di concetti teologici.

Nell'intelligenza artificiale generale si può riconoscere un Anticristo secolarizzato, cioè senza garanzia di salvezza, con il pericolo dell'estinzione. Ciò porta alla firma di "appelli" per la "sicurezza" e per "trattenere", in un processo in cui il katechon ha comunque assunto già un volto chiaro, al di là del mistero: quello delle aziende che trainano questo sviluppo, e che quindi "custodiscono", seppur nella loro conoscenza limitata, la possibilità di accelerare la fine, di far finire il mondo. Così, noi viviamo una nuova "età assiale" (Achsenzeit), il termine con cui il filosofo Karl Jaspers identifica il periodo in cui "si concentrano i fatti più straordinari".

Bill Dally, nella sua testimonianza al Congresso del 2023: Alcuni commentatori hanno espresso il timore che i modelli di intelligenza artificiale di frontiera diventino incontrollabili prodotti di "intelligenza artificiale generale" che sfuggiranno al nostro controllo e causeranno danni. Fortunatamente, l'intelligenza artificiale generale incontrollabile è fantascienza, non realtà. Anche se potremmo non essere in grado di prevedere tutto ciò che farà un determinato modello di intelligenza artificiale, tuttavia il modo con cui i modelli vengono costruiti e collegati limita ciò che possono fare. Fondamentalmente, l'intelligenza artificiale è un programma software, non un reattore nucleare. È limitata dal suo addestramento, gli input forniti e la natura dell'output. Quando creiamo un'intelligenza artificiale, prima decidiamo quali compiti vogliamo che esegua e la addestriamo a svolgere tali compiti.

Per esempio, supponiamo di addestrare un'intelligenza artificiale a rispondere a domande: digita una domanda e fornirà una risposta. Quell'intelligenza artificiale non è addestrata o connessa in modo da consentirle di guidare automobili, pilotare aeroplani o controllare la rete elettrica. E anche se potesse rispondere alle domande su come svolgere tali compiti, non avrebbe la capacità di requisire un'auto o un aereo. L'intelligenza artificiale sta

esattamente dove la mettiamo, può fare solo ciò che la addestriamo a fare, e può influenzare solo ciò a cui sono connessi i suoi output. Di conseguenza, saremo sempre noi esseri umani a decidere quanto potere decisionale cedere ai modelli di intelligenza artificiale. Essi non prenderanno mai il potere da soli. Avranno solo il potere che diamo loro.

Immaginiamo che OpenAI, DeepMind, Anthropic o un'altra casa o bottega di Salomone, che cavalca un'onda continua di aspettative e arricchimento, ottenga una "svolta decisiva" che suscita un giudizio quasi unanime di un pericolo generale. In tal caso, qualcuno che ha conoscenza di ciò che sta accadendo, per esempio il generale Paul Nakasone, ex capo della NSA – entrato nel 2024 nel consiglio di OpenAI per portare al comando lo "stato profondo" di Washington –, informerà il governo degli Stati Uniti. Gina Raimondo, o il suo successore, insieme al capo del Pentagono, farà stendere ai suoi uffici il testo dell'ordine esecutivo con cui il presidente degli Stati Uniti metterà sotto il diretto controllo del governo l'impresa in questione.

Come Truman l'8 aprile 1952 per il controllo della produzione di acciaio durante la Guerra di Corea, il presidente invocherà il Defense Production Act del 1950, una "legge poco conosciuta di grande importanza per la nazione", richiamata anche nell'ordine esecutivo del 2023 sull'intelligenza artificiale. In questo scenario, l'ala del "futurismo assoluto" rappresentata dai Marc Andreessen cercherà di impedire l'intervento, e alcuni membri del Congresso protesteranno con veemenza, così come l'industria dell'acciaio avviò nel 1952 una campagna pubblicitaria contro la scelta di Truman. La decisione presidenziale porterà a un contenzioso nelle corti per stabilire, con un'accelerazione dei tempi della legge, se si è trattato di un abuso del potere presidenziale, come nel caso di Truman.

Mentre la legge farà il suo corso, gli ufficiali statunitensi si recheranno fisicamente nei "mulini satanici", cioè negli uffici di OpenAI a San Francisco, nei data center in Iowa e altrove (ma in un "altrove" dove il potere pubblico statunitense potrà comunque arrivare, che riguardi OpenAI, Google DeepMind, Anthropic o chiunque altro stia in America). Ciò confuterà l'ipotesi di Geoffrey per cui il programma, per quanto potente, sia in grado di prendere il controllo in modo automatico e di indurre gli esseri umani a realizzare il suo volere egemonico. È probabile che si tratti di un falso allarme, ma questa "esercitazione" servirà a mostrare il controllo politico della tecnologia, rendendo ancora più stretta l'interconnessione tra le aziende e il governo, nell'unico contesto in cui ciò è possibile, quello della politica nazionale, nell'unico luogo in cui ciò sembra possibile, gli Stati Uniti d'America. A seconda dell'accelerazione che ci aspetta, vedremo una proliferazione à la carte di questi strumenti di controllo sulla base della sicurezza nazionale in molti altri paesi, in analogia a quello che è avvenuto sul controllo degli investimenti esteri e sui controlli delle esportazioni.

 # Conclusione

In Europa sarebbe auspicabile l'esistenza immediata di un'unione dei mercati dei capitali, oltre a una politica quanto più aggressiva di attrazione dei talenti e di riduzione radicale del peso regolatorio, ma è essenziale capire che i capitali europei, privi di ambizione e di scala, hanno già perso l'occasione: non si possono invertire o fermare le leggi di Moore e di Huang perché bisogna attendere il Godot di un pezzetto dell'Occidente. Dove si può creare e innovare? Negli Stati Uniti. Dove si può costruire e scalare? In Asia.

Il discorso pubblico europeo privilegia la lamentela attorno al mantenimento di un proprio ruolo nel mondo alla conoscenza di paesi la cui crescita è protagonista in questo secolo e in questa fase storica, perché semplicemente ospitano più persone istruite e più capacità produttiva oppure perché sono in diverso modo "connettori" o "frontiere" nel conflitto tra Stati Uniti e Cina. Questa conoscenza non è più opzionale. È necessaria. La domanda "Cosa possiamo fare?" è dunque inutile, provinciale e patetica, se si ignora ciò che accade non solo in Giappone, Corea del Sud, Taiwan, India, ma anche in luoghi come Singapore, Malesia, Vietnam.

Nel "nuovo modo di produzione asiatico" tutti gli ingranaggi funzionano. Quando c'è un problema, i manutentori entrano a correggere i difetti delle macchine, le riparano, e la produzione riprende con maggiore vigore. Nel 1983, Samsung inizia il suo percorso nell'industria dei semiconduttori e sviluppa la sua prima DRAM da 64 kB. Quel primo risultato è celebrato con una curiosa cerimonia: in onore dei 64 kB, gli ingegneri coreani fanno un'escursione in montagna di 64 km. In quarant'anni, la capacità di Samsung nella DRAM crescerà di un fattore di 500.000. Eppure, il 7 giugno 2024 i dipendenti di Samsung non marciano più tra le montagne. Alcune migliaia sfilano col pugno chiuso per il primo sciopero nella storia dell'azienda, guidato dal più grande sindacato, National Samsung Electronics Union (Jeonsamno). Lo sciopero va avanti a fasi alterne per venticinque giorni, ma la partecipazione diminuisce e il sindacato non si accorda col management sull'entità dell'aumento dei salari.

Samsung – che domina la Corea come un conglomerato ha dominato solo le entità politiche della fantascienza – si è trovata a inseguire la più piccola SK Hynix (anch'essa da qualche anno alle prese con le richieste dei sindacati) nella fornitura di memorie per alimentare i "mulini satanici" dell'intelligenza artificiale. Lo sciopero di Samsung non porta ad alcun risultato e i lavoratori tornano tutti al loro posto all'inizio di agosto. Cosa ha pensato Morris Chang leggendo degli scioperi del suo avversario coreano, alle prese col sindacato, e vedendo le immagini della nuova marcia? Bisogna immaginare il suo sorriso mentre si compiace della fragile ricchezza di Taiwan, alimentata dall'ecosistema che lui ha costruito.

Secondo le stime di SemiAnalysis, i data center nel 2030 potrebbero utilizzare il 4,5% dell'energia generata a livello globale.