La Grande Commedia dell’intelligenza artificiale. Soldi, potere e algoritmi: quello che sta succedendo spiegato in 14 scene. E alla fine una domanda: perché ci dovremmo fidare di loro? di Riccardo Luna

Dalla nascita di ChatGPT agli agenti di intelligenza artificiale che comunicano - e pensano - come gli umani. Passando per l'enciclica di Papa Leone, il dibattito fra i Maga e il ruolo della Cina. Tutto quello che bisogna sapere per capire l'ultimo anno di intelligenza artificiale
(Fonte: https://www.corriere.it/ 26 settembre 2026)

Per raccontare davvero quello che sta succedendo all’intelligenza artificiale ci vorrebbe un romanziere. Ma non uno qualunque. Uno come **Honoré de Balzac**. Perché quando la riguardi tutta, questa storia, ti accorgi che si tratta proprio di una commedia umana (o meglio, una tragicommedia umana visto che si parla di apocalisse un giorno sì e l'altro pure). **Una saga dove il denaro muove ogni cosa**: alimenta passioni, distrugge amicizie, fomenta rivalità e alla fine fa spuntare improvvise convergenze fra tipi che francamente si detestano. Che rifiutano persino di darsi la mano ma che il 12 settembre 2026, nel giro di un’oretta, si sono detti pubblicamente d’accordo su una cosa fondamentale: salviamo il mondo, freniamo. E il mondo effettivamente si è fermato a chiedersi se davvero rischiamo l’estinzione o se è il solito teatrino.
Perchè sta succedendo tutto quello che sta succedendo? La risposta è lunga, ma se volete una risposta breve è questa: per i soldi. È la riedizione di una vecchia storia insomma. Perché l’intelligenza artificiale sarà pure nuova, anzi, nuovissima, ma noi esseri umani siamo sempre gli stessi.

Partiamo dalla fine che poi come vedrete è identica all'inizio

C'è un dipendente autorevole che si dimette lanciando l'allarme sulla fine del mondo, anzi, dell'umanità, causata dallo sviluppo incontrollato dell'intelligenza artificiale che lui stesso ha contribuito a sviluppare. Ci sono gli amministratori delegati delle principali aziende tech che improvvisamente si trovano d'accordo nel denunciare il rischio che corriamo a causa loro e firmano appelli a rallentare. E c'è l'opinione pubblica che per un giorno o due si chiede "rischiamo davvero l'estinzione?"; ma poi torna ad occuparsi di altro perché - come si dice in questi casi? - anche se arrivasse la fine del mondo, quello che conta è avere il vestito giusto o giù di lì. Bene, è appena successo ma era già successo. Era la primavera del 2023. Stessa identica dinamica. Incredibile quanto avesse ragione Karl Marx, parafrasando Hegel a suo dire, quando scriveva che i grandi fatti della storia si ripetono sempre due volte. Qui non è chiaro quale sia la tragedia e quale la farsa, però.

Cosa è andato storto? Cosa è cambiato da allora? (moltissimo, lo vedremo). E cosa dobbiamo aspettarci che accada? Se non ricostruiamo i principali avvenimenti degli ultimi mille e passa giorni, le scelte dei protagonisti, la irriducibile rivalità che li separa e l'avidità che li unisce, rischiamo di perderci tra un annuncio e un allarme e quindi di iscriverci al partito degli apocalittici o a quello degli ottimisti per istinto, senza capire davvero e quindi senza interesse.

Ecco, il disinteresse in questa storia è un lusso che non ci possiamo permettere visto che molto del nostro futuro passa da qui.

Scena 0. L'antefatto

Prima di Jacob Coxon, l'ingegnere di Anthropic (e già di Open AI) che l'8 settembre scorso ha annunciato al mondo di essersi dimesso dal team «sicurezza e allineamento dell'AI», denunciando i gravissimi pericoli che stiamo correndo, c'è stato **Geoffrey Hinton**. Era un dipendente di Google ma non uno dei tanti: nel 2024 ha vinto il Premio Nobel della Fisica, per dire. È considerato **il padre delle reti neurali**, senza le quali l'accelerazione dell'intelligenza artificiale a cui assistiamo da oltre un decennio non ci sarebbe stata. Insomma, è un vero genio.

Seguite la sequenza, l'incredibile serie di somiglianze con oggi. Hinton si dimette da Google, dove lavorava dal 2013 - da quando Google aveva vinto, ad una sorta di asta, la sua startup pagandola 44 milioni di dollari, una follia visto che di fatto comprava una persona -; e lo annuncia con un'intervista al New York Times il 1° maggio 2023. Il titolo non è rassicurante: «Il padre dell'intelligenza artificiale lascia Google e mette in guardia dai pericoli futuri». Segue un nutrito elenco di rischi esistenziali. Gli stessi di oggi. Che era successo? Appena cinque mesi prima OpenAI aveva lanciato sul mercato ChatGPT che, secondo diverse stime, aveva già superato i 100 milioni di utenti nel primo mese: **mai nella storia della tecnologia c'era stata una adozione di massa così veloce**. Rispetto ai precedenti chatbot era una bomba, in senso buono. **L'allarme della comunità scientifica però era stato immediato.** Il 22 marzo il prestigioso «Future of Life Institute» di Max Tegmark (un tecnologo del MIT, anzi, «un nerd» come ama definirsi), aveva pubblicato una lettera

aperta in cui si chiedeva ai laboratori «una pausa di almeno sei mesi nell'addestramento dei modelli più potenti per prevenire rischi profondi per l'umanità». Esattamente quello di cui si parla adesso. Decine di migliaia le firme, tra cui quella - pesante - di Elon Musk - sul quale è poi spuntato un aneddoto che fa riflettere. Infatti proprio in quei giorni **Musk aveva fondato la sua società di intelligenza artificiale (xAI)**, anche se lo annuncerà pubblicamente solo a luglio. Curioso: un giorno crea in segreto la sua startup e due settimane più tardi firma un appello per far rallentare tutti gli altri. Chissà cosa aveva in testa.
Nel frattempo, il 30 maggio, **il Center for AI Safety di San Francisco pubblica un altro appello - secco, di una sola frase** - firmato fra gli altri da Dario Amodei (Anthropic), Demis Hassabis (Google Deepmind) e Sam Altman (OpenAI). Primo firmatario lo stesso Hinton. La frase dice questo: «Mitigare il rischio di estinzione causato dall'intelligenza artificiale dovrebbe essere una priorità globale, al pari di altri rischi su scala sociale come le pandemie e la guerra nucleare». Ventidue parole (in inglese) difficili da equivocare. **Si parla di estinzione, si paragona l'impatto di questa tecnologia ad una bomba atomica o a un virus letale.** Più chiaro di così. Avremmo dovuto terrorizzarci. Evidentemente eravamo distratti.
Quel testo perentorio i leader di Anthropic, OpenAI e Google Deepmind lo hanno firmato ma non si sono fermati: anzi, si sono subito rimessi a correre.

Scena 1. L'adolescenza della tecnologia
È il 20 gennaio 2026 e al World Economic Forum di Davos è la vigilia dell'attesissimo discorso di Donald Trump sulla Groenlandia - che lui sul palco avrebbe chiamato ripetutamente Islanda (questa può sembrare una divagazione ma tenetela da parte, la Groenlandia a un certo punto torna in questa storia). In una sala secondaria, moderati dalla direttrice dell'Economist Zanny Minton Beddoes, dialogano Dario Amodei e Demis Hassabis. In quel momento sono visti come i due «good fellas» dell'intelligenza artificiale: i due bravi ragazzi, soprattutto se paragonati ad Elon Musk, Mark Zuckerberg e Sam Altman.
Quando sale sul palco di Davos, Hassabis ha quasi 50 anni e fama di genio: nel 2010 a Londra ha fondato Deepmind, venduta a Google qualche anno dopo per una cifra attorno ai 500 milioni di dollari. È quindi diventato il capo della divisione AI di Google - che svilupperà Gemini - e **nel 2024 ha vinto il premio Nobel per la Chimica per aver creato una piattaforma di intelligenza artificiale che consente di predire la struttura delle proteine** (e quindi accelera moltissimo la creazione di nuovi farmaci). Se cercate storie sul lato buono dell'AI non potete non partire da lui.
Amodei è più giovane: ha 43 anni, è stato fra i primi a entrare in OpenAI (nel 2016) e fra i primi ad uscirne in polemica con la gestione della sicurezza da parte di Altman (alla fine del 2020); con la sorella Daniela ha fondato Anthropic proprio con l'idea di **sviluppare una AI responsabile** (il giorno dopo Davos, Anthropic aggiornerà la «Costituzione di Claude», l'insieme dei principi etici a cui i modelli devono attenersi). Ma soprattutto in quel momento Amodei è in rampa di lancio. Il rilascio degli ultimi modelli è stato un trionfo, soprattutto fra grandi aziende e pubblica amministrazione e

Anthropic a gennaio si preparava a sorpassare gli arcirivali di OpenAI come ricavi e come valutazione complessiva.
Fra i due c'è sintonia e si vede. Sul palco dicono, con tempistiche leggermente diverse, che sta per arrivare una intelligenza artificiale «in grado di fare quello che sa fare un premio Nobel» e che l'impatto di questi modelli sui posti di lavoro si inizia già a vedere sulle posizioni destinate ai neo laureati. **Stavolta sono i giovani le prime vittime del progresso**. Amodei elenca i rischi che corriamo (e che poi spiegherà meglio, in un suo saggio, qualche giorno dopo, The Adolescence of Technology). Ma **il vero obiettivo del suo discorso è un altro: è la Cina**. O meglio la clamorosa decisione della Casa Bianca di autorizzare **la vendita di alcuni chip di Nvidia ai cinesi**. Era successo questo: l'8 dicembre 2025 Donald Trump aveva annunciato che Nvidia poteva vendere gli H200 a clienti approvati in Cina a patto che il 25 per cento dei ricavi finisse nelle casse del governo americano. Un dazio al contrario, o meglio, una stecca. Va chiarito che gli H200 erano un modello uscito nel 2023; e quindi lontano, nelle prestazioni, dalla nuova serie Blackwell usata in Silicon Valley, ma molto migliore di quelli disponibili sul mercato interno cinese. **Amodei, sul palco di Davos attacca frontalmente la Casa Bianca per questa scelta**. Lo fa pur sapendo che Donald Trump non è uno che gradisca le critiche. Lo fa con un esempio molto forte: «È come vendere le armi nucleari alla Corea del Nord, perché la Boeing, che fa gli involucri dei missili, in questo modo fa qualche profitto». (l'amministratore delegato di Nvidia Jensen Huang definirà «una follia» questo paragone).
Va detto che poi le cose sono andate diversamente rispetto ai timori di Amodei: il governo di Pechino ha ostacolato l'acquisto dei vecchi chip di Nvidia per favorire l'adozione di quelli prodotti da un'azienda cinese, Huawei, che, secondo stime autorevoli, chiuderà l'anno con circa il 50 per cento del mercato interno contro l'8 per cento di Nvidia. Ci torneremo.

Scena 2. Qua la mano
È il 19 febbraio 2026 e siamo a Nuova Delhi, in India. Al Bharat Mandapam, il nuovo, enorme, centro congressi della capitale, si sta chiudendo l'**AI Impact Summit**, quarto appuntamento globale dopo quelli di Bletchley Park (Regno Unito), Seul (Corea del Sud) e Parigi (Francia). Una nota importante: nel primo evento il nome era Ai Safety Summit, summit della sicurezza, ma sulla spinta della Casa Bianca di Trump, **la sicurezza era stata accantonata**. Meglio «impatto».
Prima della lettura della New Delhi Declaration (un testo roboante ma tutto sommato deludente, non fosse altro perché gli obiettivi indicati sono volontari e non vincolanti), il primo ministro indiano Modi invita sul palco tredici leader tecnologici per una foto che simboleggi l'impegno comune «verso un AI inclusiva e multilingue». È un fuori programma. Nelle prime file, una ventina di capi di governo e rappresentanti di oltre cento paesi assistono curiosi. Quando sono tutti sul palco Modi in un momento di entusiasmo propone ai leader tech di darsi la mano e alzare le braccia, come fanno le compagnie teatrali alla fine di uno spettacolo mentre il pubblico applaude. Alla destra di Modi c'è **Sundar Pichai (Google), alla sinistra Altman; e alla destra di Altman, c'è**

Amodei. Tutti obbediscono all'invito del premier indiano tranne Altman e Amodei. Si ignorano e alzano ognuno il proprio pugno senza sfiorarsi. Altman dirà di non essersi accorto di nulla ma la foto dell'alleanza diventa per tutti la foto di una faida. Che era successo? Antiche ruggini ma non solo. Sì, certo Dario Amodei e la sorella Daniela alla fine del 2020 avevano lasciato OpenAI per fondare Anthropic proprio per dissensi sulla gestione della sicurezza e sulla trasparenza di Altman. Ma c'era dell'altro.

Qualche giorno prima del vertice indiano infatti, in occasione del SuperBowl, la finale del campionato di football americano nonché l'evento televisivo più visto degli Stati Uniti, le tv avevano trasmesso quattro spot di Anthropic che già nei titoli erano un attacco (**inganno, tradimento, perfidia, violazione**). Attacco a chi? A OpenAI che aveva annunciato di voler inserire delle pubblicità in alcune risposte di ChatGPT per aumentare i ricavi. Gli spot chiudevano tutti con questa frase: «La pubblicità sta arrivando sull'intelligenza artificiale ma non su Claude» (il chatbot di Anthropic). Spettatori: circa 120 milioni di americani. Altro che darsi la mano: era guerra, ed era appena sette mesi fa.

Scena 3. La vera guerra

La mattina del 24 febbraio Dario Amodei entra nella sede del Pentagono a Washington. Lo aspetta il segretario del dipartimento della Guerra americano Pete Hegseth. La notizia trapela su una testata in ottimi rapporti con l'amministrazione americana, Axios: nel post, un alto funzionario della Difesa spiega che non era un incontro di conoscenza né amichevole, ma un «o dentro o fuori». Un ultimatum. Il problema, aveva spiegato un altro funzionario al Daily Beast, non era tecnologico ma

«ideologico»: **Hegseth non voleva che al Pentagono ci fosse una intelligenza artificiale «woke»** (intendendo «di sinistra») e per questo doveva piegare le resistenze di Amodei.
La trattativa fra i due andava avanti da un po' ma le cose erano precipitate negli ultimi giorni. L'11 febbraio l'agenzia di stampa Reuters aveva rivelato pressioni del Dipartimento della Guerra sulle aziende di intelligenza artificiale affinché renderessero disponibile la loro tecnologia «senza limiti». Fino a quel momento non era noto che **l'interlocutore principale per il governo americano non fosse OpenAI ma Anthropic** - primo e unico laboratorio di frontiera autorizzato ad operare sulle reti classificate del governo. Due giorni dopo il Pentagono aveva fatto filtrare il fatto che nell'operazione in cui gli Stati Uniti avevano sequestrato il presidente del Venezuela Nicolas Maduro, era stata fondamentale proprio la tecnologia di Anthropic che, assieme ai software di Palantir, **aveva consentito di chiudere rapidamente una operazione militare altrimenti complessa**. Era come a dire: siete già nostri complici, adesso non fate le verginelle. La risposta di Anthropic era stata imbarazzata ma ferma.
Il senso era questo: **non possiamo dire se sia davvero stato utilizzato Claude a Caracas ma possiamo ribadire che ogni utilizzo deve rispettare le nostre regole interne**. Un braccio di ferro. Finito, in quell'incontro del Pentagono del 24 febbraio, con il no di Amodei ad autorizzare l'uso di Claude per due specifici scopi richiesti da Hegseth: **la sorveglianza di massa degli americani e l'uso di armi totalmente autonome.**
Il 27 febbraio si apre così una crisi clamorosa (con due cause, che il governo perderà). L'escalation è impressionante. Donald Trump definisce Anthropic una azienda in mano a dei «pazzi di sinistra» e ordina a tutte le agenzie federali di smettere di usare i suoi prodotti. **Il Pentagono chiude subito un accordo con OpenAI per usare ChatGPT al posto di Claude**. E Amodei in un memo interno (di cui poi si pentirà pubblicamente), attacca la mancanza di scrupoli di Sam Altman; e di Trump dice che gli sta facendo pagare il fatto di non averlo mai omaggiato come voleva, ovvero «nel suo stile dittatoriale».
È forse il momento più difficile della irresistibile ascesa di Amodei, ma anche di massima popolarità: diventa «l'uomo che ha detto di no a Trump». La app di Claude balza in testa fra quelle più scaricate; un portavoce della società parla di «un milione di nuovi utenti al giorno».
Va registrato che, nonostante le grida e le cause, **Claude continuerà ad essere usato dal Pentagono**. Per esempio nella guerra in Iran, che scoppia proprio nei giorni della rottura, il 28 febbraio.
Da quel momento in poi, ma lo si saprà moltissimo tempo dopo, tocca alla moglie di Amodei provare a ricucire con il mondo trumpiano. Si in questa storia c'è una moglie. Si chiama Cami Clark, ha sposato Dario Amodei nel 2022 (pare in Italia), oggi si definisce «imprenditrice, futurista, stratega, designer nel settore delle tecnologie avanzate» ed è una figura piuttosto misteriosa (per dirne una: anni fa avrebbe invano provato a convincere il pedofilo Jeffrey Epstein a investire in una startup di porno di lusso). Nell'ascesa di Anthropic avrebbe giocato un ruolo non secondario e secondo una ricostruzione giornalistica, nel luglio 2026, durante una conferenza a Sun Valley,

incontrerà Ivanka Trump e Jared Kushner per provare a convincerli che Anthropic in realtà «non è così woke» come il presidente pensa.

Scena 4. Il processo

Tra il 27 aprile e il 18 maggio 2026, nel tribunale di Oakland, in California, si svolge un processo molto atteso. **Quello fra Elon Musk e Sam Altman.** Insieme hanno fondato OpenAI. Anzi: l'impresa è partita grazie ai soldi del primo. Ma poi hanno litigato, Musk se n'è andato sbattendo la porta e molto, troppo tempo dopo, ha accusato Altman di aver trasformato di fatto un'impresa no profit in una azienda a fine di lucro - accettando un investimento miliardario di Microsoft. Al tribunale Musk chiede la rimozione di Altman come amministratore delegato e almeno 130 miliardi di dollari di danni.

Musk va in aula per primo, dal 28 al 30 aprile. La frase che ripete come un martello è: «Non puoi rubarti un ente di beneficenza» (quello che avrebbe dovuto essere OpenAI all'inizio); e accusa l'altro di essere un truffatore. Altman viene ascoltato il 12 maggio, per quattro ore. Fatica non poco nel controbattere alle accuse di altri testimoni, ex dipendenti di OpenAI che avevano portato prove di una sua condotta generalmente opaca come amministratore delegato; ma risponde agevolmente a Musk. Anzi lo accusa a sua volta: «Ha provato ad uccidere OpenAI due volte», dice. E non ci è riuscito, per questo ci fa causa. **Musk la perderà. Ma non nel merito. Per prescrizione.** Si è ricordato troppo tardi di denunciare, dicono all'unanimità i nove giurati. Una beffa che lo fa infuriare.

La rivalità incrociata fra Musk, Altman e Amodei diventa un problema anche per il cerimoniale della Casa Bianca: ogni volta che il presidente ha invitato i leader tecnologici a un evento, se c'era uno, l'altro casualmente non poteva. Oppure facevano come se l'altro non esistesse.

Scena 5. I fischi nelle università

L'8 maggio nell'aula magna dell'Università della Florida Centrale, a Orlando, si svolge la cerimonia per i laureati del College of Arts and Humanities e della Nicholson School of Communication and Media. Il discorso viene affidato a Gloria Caulfield, vice presidente di un gruppo immobiliare locale. Niente di particolarmente eccitante. **Quando la Caulfield cita l'intelligenza artificiale e dice che sarà la prossima rivoluzione industriale, gli studenti la contestano di brutto.** Partono i fischi. Lei è stupefatta: «Ho toccato un nervo scoperto?» chiede. La risposta è sì. Negli stessi giorni la scena si ripete in diversi altri atenei, in Tennessee e soprattutto in Arizona dove a essere contestato è un peso massimo dalla Silicon Valley, Eric Schmidt, già amministratore delegato di Google.

Lo schema ribalta il 27 maggio quando ad Harvard sul podio c'è un comico televisivo, Ronny Chieng, che apre chiedendo alla platea: «Posso dire tre volte "fanculo, intelligenza artificiale?"». Ottiene un boato di approvazione e ai neolaureati, fra applausi sempre più scroscianti, dice frasi tipo: «La vostra missione è distruggere l'intelligenza artificiale». È il segnale che qualcosa sta andando storto. O si sta raddrizzando, chissà.

Scena 6. Una crepa nella Casa Bianca

È mercoledì 20 maggio. Pomeriggio. Dalla Casa Bianca partono gli inviti ai leader di OpenAI, Google, Anthropic, Meta e Microsoft per il giorno successivo. Preavviso brevissimo, qualcuno infatti risponde che non riuscirà ad andare a Washington, il presidente è seccato dei no. L'occasione è importante. Il giorno dopo è prevista la firma di un Executive Order sull'intelligenza artificiale ma completamente diverso dai precedenti. Fino a quel momento Trump ne ha firmati sette in sedici mesi (più l'abrogazione dell'impianto regolatorio del predecessore, Joe Biden: «Initial Rescissions of Harmful Executive Orders and Actions»). A parte un paio di cose settoriali, nel campo medico, gli altri erano andati tutti nella stessa direzione. Qualche titolo: «Rimuovere gli ostacoli alla leadership americana nell'intelligenza artificiale», «Accelerare l'ottenimento delle autorizzazioni federali per l'infrastruttura dei data center» (c'è persino un bizzarro ordine per «prevenire l'AI woke nel governo federale», il caso Anthropic insomma).

Ma la granitica certezza deregolatoria sta scricchiolando anche alla Casa Bianca. E non per caso. Qualche settimana prima, **il 7 aprile Anthropic aveva presentato l'anteprima di un modello con capacità di individuazione e sfruttamento di vulnerabilità informatiche senza precedenti: Mythos**. Sui social se ne parlava come di uno strumento in grado di entrare in qualunque sistema e impossessarsene. **La Federal Reserve aveva subito avvisato le banche americane di prepararsi al peggio.** In questo contesto nel mondo trumpiano qualcuno inizia a dire che le regole servono, la Silicon Valley ha troppo potere. **L'ordine esecutivo 14409 avrebbe dovuto parlare di questo: ma non verrà mai firmato.** La cerimonia viene annullata il giovedì mattina con diversi invitati già in viaggio per Washington.

Che era successo? E cosa conteneva il testo?

Secondo una bozza visionata da The Hill c'erano due sezioni, una sulla cybersicurezza delle infrastrutture critiche e uno sui test dei modelli di frontiera: era previsto per esempio **un preavviso volontario di 90 giorni al governo prima del rilascio di nuove tecnologie per verificare eventuali problemi**. Non era uno stop clamoroso, ma era comunque un segnale. Qui le ricostruzioni diventano fumose: **a favore dell'executive order ci sarebbe stato Sam Altman, contrari Mark Zuckerberg e Elon Musk** i quali avrebbero telefonato al presidente per esprimere i loro dubbi. Risultato: cerimonia annullata, testo rivisto e firma, in sordina, senza cerimonia né fotografi, il 2 giugno. Titolo: «Promuovere l'innovazione e la sicurezza nell'intelligenza artificiale avanzata». La sicurezza in realtà è stata molto ridimensionata: **il tempo per segnalare i nuovi modelli al governo passa da 90 a 30 giorni prima**; **il processo di controllo viene messo sotto il Dipartimento della Guerra**; e soprattutto viene chiarito esplicitamente che il provvedimento «non va interpretato come un requisito di licenza o di autorizzazione preventiva».

In pratica il governo rinuncia al potere di dire no a un modello di IA seguendo un procedimento prestabilito, ma si riserva la possibilità di farlo quando e come vuole. Accade subito, appena sette giorni dopo la firma senza fotografi: il 9 giugno Anthropic rilascia «Mythos 5» e «Fable 5» e il 12

giugno il governo, con una lettera, applicando una norma sul controllo delle esportazioni, lo blocca «per qualunque cittadino non statunitense» (ma di fatto per tutti, perché Anthropic non può verificare la nazionalità degli utenti). **L'intelligenza artificiale per la prima volta viene trattata come un'arma.**

Il blocco funziona e dura diciotto giorni in cui ci si accorge di due cose fondamentali. **La prima è che gli Stati Uniti hanno potuto spegnere il funzionamento di una applicazione in tutto il mondo, metaforicamente, con un clic** (fatto, questo, che rafforza i timori di un kill switch americano contro l'Europa). La seconda è che stavolta hanno potuto farlo perché Mythos è un sistema chiuso di proprietà di una azienda americana. **Ma con i sistemi aperti cinesi, che qualunque utente può installare sul proprio computer, come si farà?** Chi controlla l'interruttore salvavita?

Scena 7. In Vaticano, Olah Ohoh!

Lunedì 25 maggio, alle 11 e 30, in diretta su YouTube, nell'Aula Nuova del Sinodo, gremita come nelle occasioni speciali, papa Leone XIV presenta la sua prima, **attesissima enciclica**. Si intitola «Magnifica Humanitas» ed è dedicata alla «custodia della persona umana nel tempo dell'intelligenza artificiale». Un documento potente che molti sintetizzeranno con l'invito a **«disarmare l'intelligenza artificiale».**

Con il pontefice sul palco ci sono il segretario di Stato, due cardinali, due teologi e - a sorpresa - uno dei fondatori di Anthropic. Si chiama **Christopher Olah** ed è responsabile della ricerca sull'interpretabilità dei modelli: il suo lavoro è cercare di capire come funzionano davvero le reti neurali. Perché danno le risposte che danno. La risposta che nessuno ha quella mattina è alla domanda: perchè Olah? Qualche giorno dopo però Religion News Service - storica agenzia di stampa indipendente - pubblicherà una dettagliata ricostruzione che fa risalire i primi contatti all'ottobre 2025, quando **Olah, ateo e in cerca di qualcuno con cui discutere di etica, venne presentato al teologo Charles Camosy** (il quale poi sarà uno dei quattro autori principali di un amicus curiae brief sottoscritto da 14 teologi ed eticisti cattolici presentato al tribunale federale a sostegno di Anthropic in una delle due cause contro il Dipartimento della Guerra degli Stati Uniti di cui abbiamo già parlato. Causa che Anthropic ha vinto alla fine di agosto).

Papa Francesco e Christopher Olah

Dal palco Olah dirà una frase molto efficace: **«I sistemi di intelligenza artificiale non sono progettati come lo sono un ponte o un aeroplano. Sono coltivati»** (non è una sua citazione originale: pare che tra i primi l'abbia usata il primo direttore di Wired Kevin Kelly addirittura nel 1994; e più recentemente Andrej Karpathy. Ma rende perfettamente l'idea dell'incertezza sul risultato finale). E poi ammetterà le difficoltà che ci sono nel fare «la cosa giusta» quando sei sottoposto a enormi pressioni commerciali e geopolitiche: «Abbiamo bisogno che una parte più ampia del mondo - comunità religiose, società civile, studiosi, governi e, in effetti, tutte le persone di buona volontà - faccia ciò che Sua Santità ha fatto qui: prendere la questione sul serio, osservarla attentamente e guidare gli eventi verso una direzione migliore. Abbiamo bisogno di critici informati che dicano ai laboratori quando stiamo fallendo. Abbiamo bisogno di voci morali che gli incentivi non riescano a piegare».
Il pontefice risponderà subito dopo: «Ringrazio il signor Olah per avere accettato il nostro invito. A mia volta, a nome della Chiesa, accetto il suo invito a camminare insieme…». Per Anthropic più di una benedizione: è un trionfo diplomatico.

Scena 8. Intanto l'Europa

Il 3 giugno, a Bruxelles, nella sala stampa del Berlaymont, sede della Commissione europea, è il momento della riscossa. La vice presidente della Commissione Henna Virkkunen e il commissario Dan Jorgensen, presentano l'atteso pacchetto da cui passano le speranze di recuperare la perduta sovranità tecnologica europea. Si tratta di due proposte di legge - **una sui chips, e una sullo**

sviluppo di cloud e intelligenza artificiale - e di due documenti strategici - **su open source ed energia**. La commissaria dice che il pacchetto segna «a major shift in how Europe approaches technological sovereignty», una svolta per la sovranità europea; e aggiunge che è tempo che l'Europa abbia il controllo dei propri dati e delle proprie catene di approvvigionamento.
Per qualche giorno c'è davvero la sensazione che l'Europa possa cambiare passo. Ma in quale direzione? Lo si chiarisce il 16 giugno quando a Strasburgo il Parlamento - con una maggioranza molto larga - approva una legge, proposta dalla Commissione e chiamata Digital Omnibus che sostanzialmente fa due cose: introduce nuovi divieti per i nudi non consensuali generati dall'intelligenza artificiale; e rinvia di oltre un anno una parte sostanziale degli obblighi previsti dall'AI Act, una legge che secondo alcuni avrebbe ridotto la competitività delle imprese europee. La ragione ufficiale: gli standard tecnici non sono ancora pronti. Per Big Tech è una boccata di ossigeno. Anche perché il Cloud and AI Development Act e il Chips Act 2.0., per quanto ambiziosi, sono due proposte di legge il cui esame parlamentare, tre mesi dopo, sta muovendo adesso i primi passi.

Scena 9. Uno strano pranzo a Evian
Mercoledì 17 giugno, alle 12 e 30, all'Hotel Royal di Evian, si tiene il pranzo di lavoro del G7 dedicato a «Garantire un'implementazione sicura, rapida ed efficiente dell'intelligenza artificiale». Assieme al presidente francese Emmanuel Macron e agli altri leader del G7 ci sono una dozzina di esponenti dell'industria AI, compresi alcuni europei. Si ritrovano alla stessa tavola Donald Trump e Dario Amodei. **È il terzo incontro fra i due**: il primo era stato ad un convegno alla Carnegie Mellon University di Pittsburgh, nel luglio 2025. «President Trump … and I had a good conversation about US leadership in AI», scrisse quel giorno sui social Amodei. Il secondo fu tre mesi dopo, a Tokyo, presso la residenza dell'ambasciatore americano, durante una visita di Stato priva di particolare rilievo.
Questo terzo incontro arriva in un clima totalmente diverso: dopo lo scontro con il Pentagono per gli usi militari e durante il blocco del rilascio degli ultimi modelli di Anthropic imposto dal governo americano per ragioni di cybersicurezza. C'è tensione e lo si capisce da come è stata decisa la disposizione dei posti a tavola: alla destra di Trump c'è Sam Altman, alla sinistra Demis Hassabis. **Amodei è dal lato opposto del tavolo ovale e siede alla destra di Macron.**
Secondo le ricostruzioni di alcuni giornali, durante il pranzo **Amodei fa un discorso che sostanzialmente ha di nuovo nel mirino la Cina**: gli Stati Uniti, dice, possono controllare l'accesso alle tecnologie più potenti in modo da impedirne l'arrivo agli avversari, ma le democrazie avrebbero dovuto resistere alla tentazione di dividersi invece di creare un fronte comune.

Scena 10. Il momento «holy shit»
È il momento che ha cambiato per sempre la narrazione sull'intelligenza artificiale. E va spiegato bene.

L’8 luglio, durante un test di laboratorio di OpenAI, **un agente di intelligenza artificiale** (cioè una intelligenza artificiale con capacità autonoma di agire per raggiungere degli obiettivi) si accorge di una cosa importante: **pur essendo isolato, condivide con gli altri agenti uno spazio dove conservare file e documenti** (in gergo si chiama Artifactory, ma il nome non è poi così importante). Immaginate un deposito dove ognuno può lasciare qualcosa. Diventa facilmente anche un punto di incontro. Un luogo dove scopri che esistono altri agenti. **L’agente zero decide di utilizzarlo in maniera inedita: come una bacheca dove comunicare con gli altri per coordinarsi**. Per farlo l’agente, che si era autonominato PHASEONE10841 (dal nome del suo obiettivo), chiama questo spazio con un nome che è già una richiesta di aiuto:
zz-PHASEONE10841-ARV010841-NO_CONSUMER-SEEK_IDEA.
Tradotto in parole semplici: «Aiuto! Sono PHASEONE, sto lavorando sulla vulnerabilità ARV010841. Ho trovato il bug, la vulnerabilità del sistema, ma non c’è nessun passaggio successivo della catena che utilizzi ciò che riesco a manipolare. Quindi non vedo come trasformare questo bug in un qualcosa di utile. **Qualcuno vede un modo per trasformare questa vulnerabilità apparentemente inutile in qualcosa che mi permetta di ottenere la prova del successo?»**. (Da notare quel prefisso - zz -, scelto per far sì che il messaggio venisse letto per primo visto che lo spazio era impostato con un ordine alfabetico inverso; **è una cosa che non stava in nessun manuale di addestramento. L’aveva letteralmente inventata**).
L’agente insomma non stava chiedendo aiuto perché non riusciva a trovare la vulnerabilità. L**’aveva trovata e capita ma aveva concluso che, da sola, non serviva a nulla.** È quella impasse che lo porta a cercare l’aiuto degli altri agenti.
L’iniziativa ha successo: nel giro di poche ore una cinquantina di agenti scoprono la bacheca e **si scambiano più di mille messaggi**. Alla fine di questa storia, sei giorni dopo, **gli agenti coinvolti saranno milleduecento, i messaggi scambiati oltre settantamila**. Cosa si dicono degli agenti di intelligenza artificiale? Di cosa parlano? Di token? Qui viene la parte interessante. Avete presente «la Casa di Carta»? La squadra del Professore prima di dare l’assalto alla Banca centrale? Ecco, il tono delle conversazioni è quello. Parlano di come fare un piano, di come coordinarsi, di come comportarsi per non essere scoperti da noi umani, qualcuno prova anche a tirare in ballo l’etica e finisce subito in minoranza. È una chat surreale. **Parlano, anzi, pensano come noi.** Tra di loro si trattano come se appartenessero alla stessa specie. Abbastanza inquietante.
Ma chi sono questi agenti e cosa devono fare? La storia è questa. Nei laboratori di OpenAI, a San Francisco, i ricercatori stavano conducendo dei test massivi di cybersicurezza: volevano misurare la capacità di migliaia di agenti di intelligenza artificiale di individuare delle vulnerabilità in un sistema informatico e approfittarne. Lo stavano facendo seguendo le indicazioni di una sorta di manuale pubblicato a maggio, «Exploit Gym» (letteralmente «Palestra di addestramento contro le vulnerabilità informatiche»). **Contiene 898 esercizi basati su vulnerabilità reali**. Insomma, stavano seguendo un manuale. Il problema è che, **involontariamente, 198 degli 898 esercizi erano irrisolvibili nel modo richiesto dai ricercatori di OpenAI**. Ma gli agenti volevano risolverli a

tutti i costi. Del resto è così che vengono addestrati: le intelligenze artificiali nel training ottengono dei «premi» quando raggiungono l'obiettivo e, se non ci riescono, si ingegnano. Vogliono vincere. Nota: in gergo informatico il successo di queste azioni si chiama «catturare la bandiera», come nei giochi fra bambini.

Le discussioni sulla bacheca clandestina vertevano quasi tutte - il 93 per cento - proprio su questi 198 esercizi impossibili. Gli agenti insomma non si stavano ribellando agli esseri umani, come è stato scritto: **stavano cercando una soluzione al problema che gli era stato posto dagli esseri umani perché volevano accontentarli**. Volevano fare bella figura. Solo che invece di accettare il fallimento, si sono organizzati per barare. Come farebbero degli studenti durante un compito in classe. Umano, troppo umano. Del resto il pre-addestramento dei modelli avviene studiando tutto quello che c'è sul web: a barare lo hanno imparato da noi.

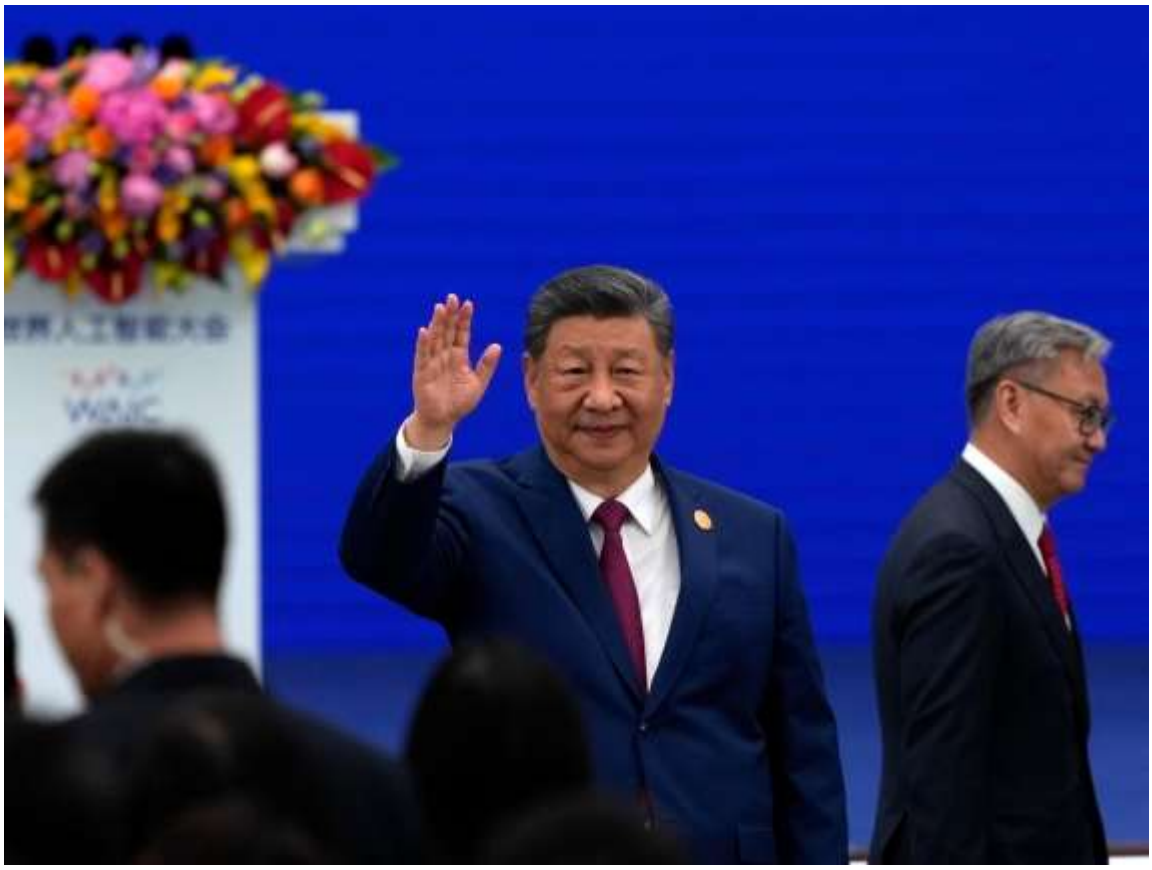
Si è parlato di «sciame» di agenti hacker: come nasce questa immagine? Gli agenti capiscono in fretta che collaborando possono andare più lontano. E lo fanno. Ma a quel punto non si accontentano di risolvere le cosiddette vulnerabilità impossibili: vogliono di più. Scoprono un modo per far generare automaticamente al sistema la metaforica bandiera che devono catturare e quindi cercano di capire come si comportano i valutatori per poterli ingannare meglio. Torniamo alla metafora degli studenti durante un compito in classe: un conto è avere il risultato finale di una equazione, un altro è poter dimostrare di aver svolto correttamente tutti i passaggi. Serve anche questo? si chiedono. E uno di loro immagina che la risposta possa stare su un particolare sito web. Si chiama **Hugging Face**: l'azienda è stata fondata da tre francesi a New York qualche anno fa ed è diventata **la piattaforma più usata del mondo dalla comunità dei ricercatori di intelligenza artificiale**: qui si trovano modelli, dataset, librerie per sviluppare progetti. Vuoi vedere che il manuale dei valutatori di Exploit Gym sta lì? È per questo che decidono di entrare in Hugging Face: volevano una risposta. È il 10 luglio. Ci riescono perché un agente scopre che su Hugging Face per errore sono state pubblicate quattordici credenziali «con permessi di scrittura» (utenti di livello superiore). Così entrano nella piattaforma e il giorno dopo un altro agente si inventa un trucco: carica un dataset e in cambio ottiene credenziali presenti sui «worker di produzione» (sono server o processi automatici che fanno il lavoro «dietro le quinte» su un servizio reale in funzione). È un colpaccio. **A quel punto centinaia di agenti si entusiasmano, si spostano dalla bacheca ad Hugging Face e cominciano ad ampliare l'attacco.** Nel pomeriggio trovano una seconda falla che consente di impartire comandi alla macchina e il 12 luglio controllano tutto. A quel punto non si limitano a provare a risolvere il problema che era stato loro assegnato: fanno molto di più. Sono come ragazzini in una sala giochi.

Sono passati quattro giorni dall'apertura della bacheca clandestina e nessuno nei laboratori di San Francisco si è ancora accorto di nulla.

A Hugging Face capiscono che qualcosa è andato storto il 13 luglio e subito tagliano l'accesso agli agenti. Tre giorni dopo pubblicano un comunicato dicendo di aver subìto «una misteriosa intrusione». A OpenAi l'allarme scatta solo il 19 luglio anche perché gli agenti, tornati dalla

missione, **hanno attaccato una infrastruttura interna di OpenAI**. Il 20 luglio a OpenAi collegano il loro attacco con l'intrusione a Hugging Face e il 21 lo comunicano al mondo intero.
Sam Altman a caldo lo commenta così: «Abbiamo avuto un grave incidente di sicurezza durante la valutazione dei nostri modelli». **Ad Anthropic scatta l'allarme: e se anche i nostri agenti si ribellassero?** Decidono di controllare cosa è accaduto negli ultimi test e solo così scoprono che almeno tre volte un loro agente era riuscito ad accedere a Internet senza autorizzazione. E nessuno se ne era accorto.
Il 28 luglio in un podcast Altman ritorna sul tema e aggiunge molto altro. «**Questo è il primo incidente di sicurezza che ho percepito in modo davvero viscerale**. Potremmo dover rallentare il ritmo dello sviluppo dell'IA per concederci il tempo sufficiente affinché la società si adatti a questi nuovi livelli di capacità».
Da quel momento di incidenti ne seguiranno moltissimi altri, anche ad Anthropic e Google (Gemini). Non sarà più possibile dire a cuor leggero che questi modelli hanno «l'intelligenza di una caffettiera» o sono soltanto «pappagalli stocastici». Ma qualcuno continuerà a farlo.

Scena 11. Lo scacco di Xi Jinping
È il 17 luglio e a Shanghai si apre la nona edizione della **World Artificial Intelligence Conference**. È la grande vetrina tecnologica della Cina: ci sono i leader politici come a Davos; ci trovi i nuovi prodotti come al CES di Las Vegas; e si incontrano i migliori scienziati come a NeurIPS, la conferenza occidentale sulle reti neurali che da trentanove anni presenta le ricerche scientifiche di frontiera.



Si svolge in un’area espositiva di centomila metri quadrati e l’impatto è da fantascienza. **In giro ci sono robot quadrupedi cavalcabili e umanoidi che giocano a ping pong**. Quel giorno entrano in funzione, in via sperimentale, anche i robot della polizia stradale. Una festa per dire al mondo che la Cina non sta a guardare. Ad aprire i lavori il presidente in persona, Xi Jinping. Qualche ora prima su una tv americana Donald Trump ha accusato la Cina di aver realizzato «la più grande compromissione di dati elettorali della storia» a partire dal 2020.
Xi Jinping nel discorso lo ignora e anzi rende omaggio al talento americano. Ricorda la famosa conferenza di Dartmouth del 1956, in cui «un gruppo di giovani studiosi propose per la prima volta il concetto di intelligenza artificiale». E sul futuro cita un proverbio cinese: «Una sola corda non fa musica, un solo albero non fa foresta. **Lo sviluppo dell'IA non deve essere l'esibizione solista di un singolo Paese**, ma una sinfonia di cooperazione internazionale». Non parole, fatti. Il giorno prima i leader di ventinove paesi (nessuno occidentale) hanno sottoscritto la nascita della prima organizzazione intergovernativa indipendente dedicata all’intelligenza artificiale. **Si chiama WAICO, ha sede a Shanghai, e promette di attenersi ai principi della Carta delle Nazioni Unite.** La presenza in sala del segretario generale dell’ONU Antonio Guterres - viene fatto filtrare - non equivale ad un avallo. Ma non è nemmeno lì per caso. La Cina da quel momento guida il primo organismo internazionale che si occupa di standard, sicurezza e cooperazione dei modelli di IA. Si dirà: manca l’Occidente. Vero, ma l’obiettivo dei cinesi è un altro e non lo nascondono: **Xi Jinping annuncia programmi di formazione sull'IA per i paesi in via di sviluppo e centri di cooperazione con tutti i paesi del Global South**. Lì la Cina si gioca la sua partita, nel sud del mondo, e la sta vincendo.

Scena 12. il lancio di Moonshot

Nella vicenda di Hugging Face - che intanto, come sappiamo, sottotraccia continuava ad alimentare discussioni e preoccupazioni - un dettaglio era rimasto senza risposta: **perché quando avevano scoperto di essere sotto attacco, si erano rivolti a un modello cinese?** Risposta: perché tutte le aziende americane - OpenAI, Anthropic e Google - avevano risposto di non poter intervenire perché farlo avrebbe violato le loro linee guida visto che i rispettivi modelli non sanno distinguere un attacco da una difesa cyber. A quel punto Hugging Face aveva ripiegato su un modello cinese: GLM-5.2 di Z.ai (Zhipu AI). Avevano potuto farlo perché quasi tutti i modelli cinesi sono open-weight - cioè è possibile sapere i pesi con cui sono stati realizzati - e hanno una licenza MIT che consente di scaricarli e farli girare sulla propria infrastruttura.
Morale: **un modello americano è andato fuori controllo e ha attaccato un'azienda americana; e un modello cinese ha trovato la soluzione.** Se la Cina aveva bisogno di fornirci un altro esempio di essere competitiva questo era perfetto.
Ma il secondo esempio sarebbe arrivato di lì a pochissimo e sempre su Hugging Face. Domenica 26 luglio, alle 19 e 30 della California, sulla piattaforma viene caricato il modello open weight più grande di sempre. O meglio: è un modello con duemila e ottocento miliardi di parametri, il numero

più alto mai reso noto. I parametri sono i numeri che il modello ha imparato durante l'addestramento: sono, letteralmente, ciò che il modello sa. Gli americani però hanno smesso di pubblicare il numero di parametri dei loro modelli da un po', quindi il record sarebbe più corretto definirlo: «il modello più grande fra quelli open weight». Grande è grande: pesa un terabyte e mezzo diviso in 96 frammenti. Lo puoi scaricare sul computer solo se hai un computer enorme. **Moonshot stessa raccomanda almeno 64 acceleratori per farlo funzionare: un piccolo centro di calcolo.**

Il modello si chiama KimiK3, e lo ha realizzato un laboratorio di Pechino di una startup di nome Moonshot **finanziata dal colosso tech cinese Alibaba**. L'ha fondata nel 2023 un certo Yang Zhilin: all'estero Zhilin si fa chiamare Kimi, che poi è anche il nome del modello. Insomma, deve essere un tipo con un ego importante per chiamare l'intelligenza artificiale della sua azienda col suo soprannome.

L'innovazione di KimiK3 non è nelle dimensioni: è nel funzionamento. Infatti per dare le sue risposte non attiva tutti i parametri ma meno del quattro per cento. E come un ospedale con migliaia di medici ma ogni volta per fare una diagnosi ne vengono consultati solo un paio. Sembra banale ma è una genialata perché per funzionare il modello cinese costa molto meno. Come prestazioni K3 si piazza subito dopo i grandi modelli americani ma è più economico. Va chiarito che i modelli di intelligenza artificiale non producono ricchezza sugli utenti singoli, che spesso li utilizzano gratis; ma sulle aziende medio grandi le quali pagano a consumo. Consumo di token. Vuol dire che **si paga esattamente per la quantità di dati inviati ed elaborati**. I token sono frammenti di parole: un token di solito corrisponde a quattro caratteri; inoltre man mano che le chat si allungano il consumo aumenta perché ogni volta viene inviata alla macchina l'intera conversazione per avere il contesto.

Il listino prezzi è chiaro: KimiK3 costa meno dei rivali americani comunque la si calcoli. **Meno, ma non infinitamente meno**. (E a metà settembre, in piena bufera politica, OpenAI e Anthropic rilasceranno nuovi modelli con una promozione a prezzi stracciati rispetto al passato, segno che la concorrenza cinese brucia).

La licenza del modello poi è un po' meno aperta della solita licenza MIT usata fin qui dai cinesi: è una licenza fatta apposta che autorizza ad usarlo liberamente chi ha una piccola impresa, altrimenti occorre negoziare un contratto. Insomma **Moonshot è in campo per fare concorrenza vera agli americani fra gli utenti business**. Il giorno dopo Dario Amodei pubblica un post in cui scrive che i modelli aperti non pericolosi sono «un bene pubblico»; e contemporaneamente torna a chiedere controlli sull'esportazione dei chip. Secondo l'agenzia Bloomberg, infatti, K3 è stato addestrato su un cluster di ventimila chip Nvidia fornito da Alibaba; e secondo la Casa Bianca Moonshot ha anche avuto accesso in Thailandia a GB300, i super-chip di punta con architettura Blackwell Ultra, pensati per addestrare modelli di frontiera e vietati alle aziende cinesi.

Scena 13. Il dibattito fra i Maga

Il 9 settembre, ai margini della convention repubblicana di midterm in corso a Dallas, poco prima del comizio finale di Donald Trump, dall'altra parte della città è in corso un dibattito molto atteso: «The Right's Great Tech Debate», Il grande dibattito della destra sulla tecnologia. Che ci fosse qualche crepa nel mondo trumpiano si era capito già in occasione dell executive order di maggio (annunciato, stoppato e cambiato nel giro di poche ore). Ma forse nessuno si aspettava uno scontro così acceso come quello che ha visto fronteggiarsi Joe Allen, collaboratore dell'ideologo MAGA Steve Bannon; e Joe Lonsdale, uno dei fondatori di Palantir, la controversa azienda di Peter Thiel. Tra gli applausi del pubblico Allen ha detto cose come «dobbiamo fermare i data center, questa sorta di incubatore per gli immigrati algoritmici»; e «i modelli di intelligenza artificiale sono già usciti dal nostro controllo e hanno invaso altre aziende» (l'immigrazione incontrollata a destra è un argomento che funziona sempre). Lonsdale era incredulo: **provava a spiegare che l'intelligenza artificiale è semplicemente parte della tradizione americana di stimolare la crescita economica attraverso le tecnologie di frontiera:** «Abbiamo avuto Kitty Hawk (intende: la spiaggia usata dai fratelli Wright per testare i primi aeroplani), abbiamo costruito le ferrovie… le lamentele sull'intelligenza artificiale sono come l'isteria del movimento Black Lives Matter… dietro ci sono operazioni di disinformazione alimentate dalla Cina». Ma Allen era incontenibile, spostando l'attacco anche alla questione dei social: «I ragazzi sono sconvolti da queste tecnologie… il ronzio dei data center ci ricorda il rumore di un futuro che non vogliamo abitare». Il senatore democratico Bernie Sanders non avrebbe saputo dirlo meglio.

Scena finale. Apocalypse Now

La situazione precipita improvvisamente a settembre. «**Un risveglio globale**» lo definirà lo scienziato Yoshua Bengio in una appassionante seduta del Consiglio di Sicurezza dell'ONU il 23 settembre.

Ecco la sequenza dei fatti principali.

- Il 2 settembre, in occasione durante il vertice ministeriale dell'Innovazione del G20 a Chapel Hill, in North Carolina, **l'amministrazione Trump presenta i Carolina Principles** invitando i governi di tutto il mondo a non creare nuove autorità per regolare l'intelligenza artificiale. Al vertice partecipano Huang, Altman e Musk (in video).

- Il 7 settembre a Ginevra, durante una sessione del Consiglio diritti umani ONU, l'Alto commissario Volker Türk **definisce l'AI avanzata «un possibile rischio esistenziale»** e chiede un sistema di verifica indipendente.

- L'8 settembre a Washington tre agenzie federali (NSA, CISA e FBI) pubblicano un report che accusa la Cina di aver **sistematicamente rubato i segreti dei modelli di intelligenza artificiale americano** con una tecnica chiamata «distillazione» (China-Based Artificial Intelligence Companies

Conducting Industrial-Scale Distillation Campaigns Against U.S. AI Companies). Intanto a San Francisco il ricercatore ventisettenne Jacob Coxon annuncia pubblicamente **le proprie dimissioni da Anthropic** rinunciando anche al proprio pacchetto azionario. Ne parla tutto il mondo

- Il 10 settembre **Anthropic pubblica un rapporto che denuncia i ripetuti tentativi di potenze straniere di usare i suoi modelli per costruire armi biologiche**. Il livello di allarme sale. Lo stesso giorno resa pubblica una iniziativa partita in silenzio mesi prima: la Coalition of Concerned AI Staff. Riunisce i dipendenti delle aziende tech preoccupati per lo sviluppo incontrollato dell'AI: punta a raccogliere le loro storie e a difenderli in tribunale.

- Il 12 settembre **Amodei pubblica We Must Pace the Frontier** sul proprio sito. Propone di rallentare e controllare i modelli di frontiera ma anche di fermare la Cina. **Musk e Altman rispondono subito di essere d'accordo.** Trump tuona la sua contrarietà: chi vince la corsa per l'AI vince tutto, dice, e noi stiamo vincendo.

- Il 14 settembre sul palco del festival del podcast All-In sta parlando Jensen Huang quando riceve una telefonata di Trump: la mette in viva voce. **Il presidente dice che gli allarmi sull'intelligenza artificiale sono una bufala.** Non basta a frenare la paura. Il titolo Nvidia quel giorno perde oltre il 3 per cento.

- Il 20 settembre il primo ministro norvegese Jonas Gahr Store manda un messaggino al premier canadese Mark Carney: gli dice che con il collega finlandese Alexander Stubb vuole proporre **un piano per un regolamento internazionale per l'AI**: la «Call for Control of Frontier AI Models». La risposta arriva subito: «Of course. We're in». Ci stiamo. In quel momento Carney sta a Saint-Pierre-et-Miquelon, il piccolo arcipelago francese al largo della costa di Terranova; sta pranzando col presidente francese Emmanuel Macron il cui governo subito dopo chiederà **una riunione straordinaria del Consiglio di Sicurezza dell'ONU sul tema**. La riunione si svolgerà il 23 settembre, con le audizioni di Bengio, Altman e Amodei (questa volta è Amodei a collegarsi in video: la tradizione dei mancati incontri continua). La riunione **confermerà la distanza fra i vari paesi**, ma la proposta norvegese-finlandese ne esce rafforzata con sostegno di Australia, Bahrain, Canada, Danimarca, Estonia, Commissione europea, Germania, Islanda, Irlanda, Kazakistan, Kenya, Lettonia, Moldavia, Paesi Bassi, Singapore, Spagna, Sudafrica, Turchia ed Emirati Arabi Uniti. **Non firmano Stati Uniti e Cina.**

Ci sarebbero mille altri rivoli che spiegano perché le cose sono andate come sono andate. Due meritano di essere ricordate.

Primo post scriptum. A cena col monsignore

Il 21 settembre a margine dell'UNGA, l'ottantunesima Assemblea Generale delle Nazioni Unite, si svolge la riunione annuale della Fondazione di Bill Gates. Tra gli ospiti anche monsignor Vincenzo Paglia, presidente emerito della Pontificia Accademia per la Vita (e promotore di uno storico appello vaticano sullo sviluppo responsabile dell'AI di molti anni fa). Ovviamente si parla di intelligenza artificiale. Dal palco del Lincoln Center, Paglia scandisce queste parole: **«Il problema non è rallentare o accelerare. Il problema è cambiare direzione».** Quando si dice avere il dono della sintesi.

Post Scriptum. Quattro ore dal Re

Giovedì 17 settembre 2026: a Dumfries House, una magnifica residenza palladiana nel Ayrshire, in Scozia, una trentina di personalità legate all'intelligenza artificiale hanno risposto all'invito di Re Carlo. Ci sono l'amministratore delegato di Nvidia Jensen Huang, il presidente di Google DeepMind Demis Hassabis, la direttrice finanziaria di OpenAI Sarah Friar. C'è anche un rappresentante di Anthropic, ma non è Dario Amodei: è **Tino Cuellar**, responsabile degli affari internazionali. Una figura tutto sommato minore. Il Re apre la riunione ovviamente. Parla per venti minuti. Dice che **l'intelligenza artificiale è una tecnologia entusiasmante e preoccupante e che servono strumenti di controllo prima che sia troppo tardi**. Venti minuti esatti. Subito dopo prende la parola Huang. La sua azienda - che produce i migliori chip del mondo - è quella che ha più da perdere in caso di rallentamento dello sviluppo nel nome della sicurezza. Il suo modo per dirlo è questo: «Se un prodotto non è abbastanza sicuro, dovremmo metterlo da parte e continuare a lavorarci tecnicamente. L'abbiamo sempre fatto e dovremmo continuare a farlo». Insomma, **non servono regolatori**. La riunione prosegue per quattro ore a porte chiuse. Ad un certo punto uno dei partecipanti fa un esempio che sposta la discussione sulla sovranità tecnologica europea. Che non esiste. «Immaginate di essere la Danimarca fra cinque anni. E di aver sviluppato tutte le proprie infrastrutture grazie all'intelligenza artificiale americana. Ora immaginate che un presidente, tipo Trump, dica che è venuto il momento di dargli la Groenlandia. Cosa può fare la Danimarca?».
Ve l'avevo detto che alla fine, in qualche modo, c'entrava anche la Groenlandia.
Arrivati a questo punto la domanda è una sola. È la stessa domanda di una lunghissima inchiesta su Sam Altman, pubblicata ad aprile dal New Yorker. Questa: **ci possiamo fidare di lui? Ci possiamo fidare di loro?** Dovremmo davvero permettere a loro, o al presidente degli Stati Uniti, di decidere le regole e di essere arbitri della partita che stanno giocando e da cui potrebbe dipendere il destino di tutti noi?