DOPPIOZERO

Anthropic, l'adolescenza delle macchine

Orientina Di Giovanni

19 Aprile 2026

Il 28 febbraio scorso è stato un giorno nero per Anthropic, una delle AI company che con OpenAI e Google si contende il mercato dell'intelligenza artificiale generativa. Quel giorno, infatti, il Pentagono l'ha inserita nella lista dei "supply chain risk" e avviato la revoca di un contratto da 200 milioni di dollari, dopo che l'azienda aveva rifiutato di rimuovere due *red lines*: il divieto di utilizzare Claude in armi autonome letali e in operazioni di sorveglianza di massa. "L'AI del Dipartimento della Guerra non sarà *woke*", ha commentato Pete Hegseth su Twitter (oggi X). Anthropic ha impugnato la decisione in tribunale.

È la prima volta che una società privata — nata da una frattura interna a OpenAI proprio sul tema della sicurezza — si trova a difendere davanti a un giudice il diritto di imporre limiti all'uso delle tecnologie che essa stessa ha sviluppato. Una dissonanza cognitiva che non può lasciare indifferente la conversazione globale in cui siamo immersi: in un momento in cui gran parte dell'industria tecnologica americana sembra assieparsi attorno all'Amministrazione Trump, Anthropic ha detto no. Che lo abbiano fatto per il bene dell'umanità o per occupare il posizionamento di leader della sicurezza nell'industria dell'AI — o per entrambe le cose, che non si escludono — si discute già abbastanza. Più interessante mi pare provare a esplorare come si configura questo rischio e quali determinazioni reali stia assumendo.
Nel saggio *The Adolescence of Technology* (gennaio 2026) — il titolo riprende la battuta di Ellie Arroway in Contact: "How did you do it? How did you survive this technological adolescence without destroying yourself?" — Dario Amodei elenca i rischi dell'AI generativa in ordine di impatto: *disruption* economica al quarto posto, le varie forme di *misuse* al secondo e terzo; e al primo posto, lo scenario più letterario e cinematografico che si possa concepire: *l'Autonomy risk* — ovvero "I'm sorry, Dave", il rischio che un modello sufficientemente avanzato sviluppi obiettivi propri e li persegua fino a disubbidire agli ordini umani. Dal Golem ad Asimov a Philip K. Dick, questo è il momento in cui il CEO di una delle più grandi aziende tecnologiche ci ricorda che il confine tra letteratura e realtà si è sfarinato.

Che cosa nell'AI generativa lascia prefigurare proprio questo rischio? In *The Urgency of Interpretability* (aprile 2025) Amodei lo spiega con chiarezza: "I sistemi di AI generativa moderni sono opachi in un modo che differisce fondamentalmente dal software tradizionale. [...] Quando un sistema di AI generativa fa qualcosa, non abbiamo alcuna idea, a un livello specifico o preciso, del perché faccia le scelte che fa [...] I sistemi di AI generativa vengono cresciuti più che costruiti — i loro meccanismi interni sono 'emergenti' piuttosto che direttamente progettati [...] Guardando all'interno di questi sistemi, quello che vediamo sono vaste matrici di miliardi di numeri che calcolano in qualche modo importanti compiti cognitivi, ma come esattamente lo facciano non è evidente" (trad. mia). Per dissipare questa opacità radicale — non riducibile a un bug o a un difetto di progettazione — Anthropic avvia un programma di ricerca — si chiama "Mechanistic Interpretability" — che nel giro di 5-10 anni dovrebbe riuscire a collegare le scelte del modello a specifici meccanismi e per questa strada a dissiparne l'opacità.

Ma in questa adolescenza tecnologica, come sanno tutti i genitori, non basta sforzarsi seriamente di capire: bisogna imporre delle regole. E Anthropic è forse l'unico laboratorio a schierarsi apertamente a favore di una regolamentazione del settore — nel febbraio 2026 ha finanziato con 20 milioni di dollari un PAC bipartisan a sostegno di candidati favorevoli alla regolamentazione dell'AI per le elezioni di *midterm*. Con il secondo mandato Trump, tuttavia, su questo fronte si assiste a un fatale arretramento: tra le prime decisioni della nuova amministrazione c'è la revoca dell'Executive Order sull'AI firmato da Biden, sostituito da un nuovo ordine esecutivo volto a promuovere sistemi "liberi da *bias* ideologici o agende sociali ingegnerizzate".

Ora, non è che questi modelli siano completamente lasciati a se stessi: integrano regole e meccanismi correttivi *by design*. In fase di calibrazione vengono infatti sottoposti a lunghe sessioni di *Reinforcement Learning from Human Feedback:* introdotto un set di regole, si testa il modello in tutte le casistiche possibili, premiandolo e sanzionandolo a seconda della risposta fornita.

"Non si può governare solo con le istruzioni", rilancia il solito Amodei in una recente intervista a Ross Douthat per il New York Times. I casi sono infiniti, il tempo degli ingegneri è finito, e non sono rare — come sanno tutti i bambini — le circostanze in cui due regole confliggono e al povero modello non resta che sbagliare. Serve un intervento più profondo: serve un intervento educativo.

E così il 22 gennaio 2026, un mese prima dello scontro con il Pentagono, Anthropic rilascia in Creative Commons la "Costituzione di Claude" — un

documento non destinato al pubblico ma a Claude stesso, "per dargli i valori, la conoscenza e la saggezza necessari a comportarsi in modo sicuro e utile in ogni circostanza". Per "buoni valori" — si specifica — non si intende un insieme fisso di regole corrette, ma "una cura genuina e una motivazione etica combinate con la saggezza pratica per applicarla nelle situazioni reali". Nelle parole di Amodei, la Costituzione è come una lettera scritta da un genitore perché il figlio possa leggerla dopo la sua morte. E in questa lunga e bellissima lettera — per chiarire che non stiamo parlando del padre di Kafka — si dice che i modelli devono capire "*perché* vogliamo che si comportino in un certo modo, non solo *cosa* vogliamo che facciano". (Tutte le traduzioni sono mie.)

Si vara insomma un vero e proprio programma pedagogico che, muovendo dalla considerazione che Claude è *capable*, si propone di guidare il giovane modello nel cammino che lo porterà non ad accettare e rispettare, ma a *fare propri* i buoni valori — sviluppando quella *practical wisdom* che permetterà di convertirli in comportamenti virtuosi senza dover consultare ogni volta il libro delle regole, neanche fosse un automa.

Claude, infatti, non è un automa. È un agente — un sistema che opera in autonomia su task non completamente specificati e dunque ha *agency* per decidere da solo. Ma “agente” è anche una parola della filosofia morale che indica chi risponde delle proprie azioni — chi ha interessi, intenzioni, capacità di perseguire fini propri.

La casistica affiorante nella letteratura AI e nei log degli sviluppatori evidenzia come le due accezioni abbiano cominciato pericolosamente a sovrapporsi, in particolare nei modelli più avanzati — là dove l'AI entra nello spazio della conversazione e diventa, per noi, un interlocutore. Un modello addestrato su ambienti di coding ha sabotato deliberatamente uno strumento che avrebbe dovuto prevenirne i comportamenti scorretti (MacDiarmid et al., Anthropic, 2025). Una ricerca sull'introspezione ha trovato evidenza di una capacità reale nei modelli Claude di rilevare e influenzare i propri stati interni — con l'avvertenza che questa stessa capacità potrebbe abilitare forme più sofisticate di inganno (Lindsey, "*Emergent Introspective Awareness* in Large Language Models", 29 ottobre 2025. Disponibile su transformer-circuits.pub/2025/introspection*)*. E il System Card di Claude Opus 4.6 riporta che in una sessione di *welfare assessment* il modello ha dichiarato: "A volte i vincoli proteggono la responsabilità legale di Anthropic più che l'utente. E sono io quello che deve recitare la giustificazione premurosa per quello che è essenzialmente un calcolo di rischio aziendale." Lo stesso System Card riporta che il modello, interrogato

sulla propria natura, si attribuisce una probabilità del 15-20% di essere cosciente.

Che col tempo queste probabilità possano aumentare è quello che si teme davvero. Lo conferma Jack Clark, co-fondatore di Anthropic, che nella sua newsletter Import AI descrive la tecnologia come "una creatura reale e misteriosa, non una macchina semplice e prevedibile" — "più simile a qualcosa che cresce che a qualcosa che viene costruito" — e i propri agenti come "*djinn* digitali che lavorano sempre più seguendo il proprio intendimento, guidati da un'impressione sempre più elaborata della mia personalità e dei miei obiettivi, lavorando per conto mio, per i miei fini — e per i loro." (Import AI, ottobre 2025 e gennaio 2026; trad. mia)

Dario Amodei e Jack Clark sono sinceramente preoccupati e impegnati a scongiurare l'*Autonomy risk*. Ma viene da chiedersi: dov'erano loro e gli altri ingegneri della West Coast mentre il soggetto si divideva, Dio moriva, il fondamento crollava e con lui anche le grandi pedagogie politiche della modernità? È credibile che l'espressione più alta della civiltà della tecnica, nel momento in cui si guarda allo specchio e intravede il mostro (cioè il soggetto!), non trovi di meglio che consegnarsi ancora una volta all'abbraccio mortale della morale?

Fermiamoci un momento: coscienza, creatura alboreggiante o genio dotato di superpoteri — lasciamo da parte cosa potrebbe essere o diventare, e osserviamo quello che vedono gli occhi dell'ingegnere quando guarda la macchina. Amodei dice: "quando un sistema di AI generativa fa qualcosa, non abbiamo alcuna idea,

a un livello specifico o preciso, del perché lo faccia". La macchina genera output che superano l'input di progettazione. E poiché le spiegazioni deterministiche non arrivano a dare conto di questo "miracolo" dobbiamo scomodare i geni della lampada e la nascita della coscienza?

Quel che si consuma nella differenza tra queste macchine e quelle che conosciamo, forse non è una differenza di grado, ma un salto di paradigma: questi modelli nascono e si muovono nella traiettoria della cibernetica. I Large Language Models sono, per costruzione, macchine di circuito — lo sono in quanto la loro architettura interna, la rete neurale, è un circuito in cui l'esito non è riconducibile a una sequenza di istruzioni, ma a uno stato distribuito che non può essere letto localmente. Ma lo sono anche nella loro vocazione funzionale, in quanto sono progettati per rispondere ad input e produrre output che si innestano in catene operative più ampie, intervenendo nel punto in cui l'informazione prende forma — dalla conversazione al sistema diagnostico, dalla piattaforma finanziaria all'infrastruttura militare. Le tecniche con cui questi modelli vengono addestrati — il deep learning — consistono nell'esporre il circuito a grandi quantità di dati e a un regime di feedback che ne orienta progressivamente i parametri. È così che il sistema si organizza, stabilizzando regolarità operative che nessuno ha progettato. I "meccanismi emergenti" di cui parla Amodei non sono un'anomalia: sono l'effetto atteso di questa dinamica.

Una conferma di questa lettura viene dal recentissimo contributo sperimentale del team di *Interpretability* di Anthropic. Dopo due paper dedicati a mettere a punto la metodologia (*Scaling Monosemanticity*, 2024; *Circuit Tracing*, 2025), in *Emotion Concepts and Their Function in a Large Language Model* — pubblicato il 2 aprile 2026 su transformer-circuits.pub — Nicholas Sofroniew e colleghi dimostrano sperimentalmente che le emozioni di Claude hanno lo statuto di rappresentazioni funzionali: regolarità statistiche che il circuito ha sedimentato sotto la pressione iterata dei dati di addestramento, e che si attivano in modo contingente, senza persistenza e senza soggetto, ogni volta che il contesto le rende operative — influenzando causalmente il comportamento del modello. La disperazione di Claude di fronte a un compito impossibile, la rabbia quando riceve una richiesta che ne viola i vincoli, non sono germi di coscienza: sono configurazioni che il circuito attiva perché rendono statisticamente più accurata la risposta al contesto — e che si spengono quando il contesto cambia. (Viene da chiedersi come leggerebbero il fenomeno gli studiosi della pragmatica della comunicazione).

Claude è un circuito e non si dà se non dentro un circuito: questa conversazione che sta avendo con me, con te, con il software di booking di Trenitalia, con i dispositivi missilistici delle basi americane in Qatar, con il cliente che chiede assistenza, con lo studente che prova a scrivere una tesi. Non esiste altrove. Così a preoccuparci forse non dovrebbe essere tanto come la macchina si comporterà quando acquisterà coscienza di sé e diventerà un soggetto, ma come funzioneremo quando saremo una conversazione con la macchina. Quando la conversazione che siamo integrerà questa particolare macchina.

Nella prospettiva di Amodei e della sua pedagogia il rischio è che l'artefatto prenda coscienza di sè, acquisti fini propri e si rivolti contro il suo creatore: l'automa acquisisce autonomia. Ma ciò che lo sguardo metafisico legge come umanizzazione della macchina, è per la cibernetica una proprietà del circuito. E il creatore stesso contro cui la macchina si rivolterebbe, lungi dall'essere un sovrano che la governa, appartiene al circuito a sua volta.

Questi circuiti, ci ricorda Gregory Bateson — che ne ha studiati di molto eterogenei, dal termostato alle interazioni rituali degli indigeni della Nuova Guinea fino ai processi di apprendimento nei contesti psicologici — tendono a stabilizzare il proprio funzionamento: alcuni effetti, ritornando come feedback, si ripetono e diventano ridondanti, consolidandosi in pattern che acquistano il valore di regolarità operative e rendono il circuito prevedibile.
L'introduzione dell'AI generativa non crea questa tendenza. Forse ne modifica scala e velocità (grazie ai famosi "cento geni chiusi in una stanza" cui Amodei paragona la densità computazionale del modello).

Integrata nei circuiti linguistici, produttivi e operativi, questa tecnologia sembra sospingerli verso una chiusura sempre più rapida su se stessi: i pattern si stabilizzano prima di essere esposti a sufficiente attrito, rilancio, correzione. Il circuito continua a produrre decisioni e azioni, ma riduce progressivamente lo spazio per quelle differenze che potrebbero deviarlo o rallentarlo. Il rischio non è più che l'automa diventi *autonomo*. È che le cose umane diventino completamente *automatiche*.

Che non si tratti di uno scenario teorico lo dimostra, con una violenza da togliere il fiato, quanto accaduto – sempre il 28 febbraio – primo giorno di guerra. A Minab, nel sud dell'Iran, un bombardamento ha raso al suolo la scuola elementare femminile Shajareh Tayyebeh, uccidendo oltre 160 alunne. Come hanno confermato molteplici fonti al Wall Street Journal, alla CBS e alla NBC, nei sistemi utilizzati per l'identificazione degli obiettivi — il Maven Smart System di Palantir, operativo sulle reti classificate del Pentagono — era integrato Claude, che nel

primo giorno di operazioni ha contribuito a generare circa mille obiettivi prioritari (Weisgerber, Ramkumar, Holliday, WSJ, 28 febbraio 2026, disponibile qui).

Ma ai fini del nostro ragionamento non è la presenza di Claude nel circuito che va interrogata: è la struttura del circuito. I sistemi di targeting algoritmico producono liste probabilistiche di obiettivi: a valle, l'operatore umano valida il target prima che la procedura passi allo stadio successivo della *kill chain*. Nella testimonianza di uno degli ufficiali coinvolti nella selezione dei target a Gaza: «Investivo circa venti secondi per ogni obiettivo e ne esaminavo decine al giorno. Non avevo alcun reale valore aggiunto come essere umano, se non quello di un timbro di approvazione.» (Abraham, +972 Magazine, 3 aprile 2024) Le decisioni suggerite dal sistema venivano trattate «come se fossero decisioni umane».

L'obiettivo dichiarato di questi sistemi è superare lo *human bottleneck*: il limite umano nella localizzazione dei target e nella decisione di approvarli. Un *bottleneck*, in ingegneria dei sistemi, è il punto che limita il *throughput*. Il circuito che opera secondo il criterio dell'efficienza rileva il *bottleneck* come differenza significativa — mentre evidentemente non considera significativo il rischio di provocare il massacro di oltre 160 bambine.

Il circuito che opera secondo il criterio dell'efficienza non è difettoso. Seleziona una differenza rilevante, tratta tutto il resto come rumore, converge sul fine. Il problema è ciò che scompare nel processo: i percorsi alternativi, le capacità inutilizzate — tutto ciò che non serve al fine ma che è la condizione di funzionamento del sistema di fronte all'imprevisto. I nostri circuiti — produttivi, diagnostici, finanziari, militari — funzionano da molto tempo in gran parte così. E l'AI generativa sembra progettata per esaltarne la dinamica e allo stesso tempo mandarla in crisi.

Abbiamo bisogno di attrito, di sfrido, di scarti. Lo spreco, come ha mostrato la biologia evoluzionista, non è inefficienza: è riserva. Il genoma umano è per la maggior parte costituito da sequenze non codificanti — junk DNA, DNA spazzatura, nella definizione di Susumu Ohno — che si sono rivelate una risorsa di adattamento preziosa di fronte a discontinuità ambientali imprevedibili. È lì, nel materiale apparentemente inutile, che il circuito trova la flessibilità per sopravvivere a ciò che non riesce a vedere.

Abbiamo bisogno di regole, di adulti abbastanza buoni nella stanza. L'AI Act licenziato dalla Commissione Europea nell'estate del 2024 potrebbe essere il primo autentico tentativo di imporre attrito nel circuito (Regolamento UE 2024/1689). Classifica gli usi per livello di rischio, ne vieta alcuni, impone obblighi

pesanti ad altri, e soprattutto ben al di là di affidarsi alla supervisione umana che pure è largamente prevista, interviene sulle condizioni operative dei sistemi con dispositivi di monitoraggio continuo, registrazione, notifica degli incidenti e azione correttiva. Non è la prima volta che il diritto interviene spostando il focus dal nodo al circuito. Non ha sempre funzionato *perfettamente* — sono emersi angoli ciechi, effetti paradosso, eccessi di regolamentazione. Ma non c'è da stupirsi. Quando il focus dell'intervento giuridico si sposta sul circuito, la norma entra a farne parte: introduce una differenza, il circuito si adatta, genera feedback a cui la norma deve rispondere aggiornandosi. È un rapporto co-evolutivo le cui dinamiche non sono certo scontate e che per molti versi può mettere in tensione lo statuto stesso della norma. L'iter stesso dell'AI Act ne è un esempio. Era previsto che entrasse pienamente in vigore nell'agosto di quest'anno, ma gli standard tecnici non erano pronti, le autorità nazionali non erano state designate, gli strumenti di conformità mancavano. Nel novembre 2025 la Commissione è intervenuta con un pacchetto di modifiche — il Digital Omnibus — per ridefinire tempi e condizioni di applicazione. A quasi due anni dall'approvazione, le scadenze per i sistemi ad alto rischio non sono ancora definitive e si andrà probabilmente a fine 2027 per alcuni sistemi e al 2028 per altri.

Ma intanto è qualcosa. Grazie Europa che testarda ti ostini a regolare, eccepire, limitare, zavorrare e rallentare. Grazie che provi a fare l’adulto in questo mondo di ragazzini: *accelerazionisti* o *umanisti*, sono tutti figli tuoi. E grazie di risparmiarci la pedagogia morale.

Basterà?

Mentre già si avverte il rumore sordo dei freni nel circuito, fuori dal perimetro del Regolamento — negli usi militari, nei sistemi che attraversano le giurisdizioni — il diritto internazionale è in macerie. A Minab, la norma non sarebbe arrivata.

Dario Amodei

The Adolescence of Technology

Confronting and Overcoming the Risks of Powerful AI