La palingenesi dell'io

Processi cognitivi nell'interazione tra intelligenze artificiali e biologiche:

appunti e annotazioni

Carlo Mancosu

Gennaio 2026

«Per narrare una coscienza occorre non considerarla una voce, ossia un'unità data e sempre uguale a sé stessa. Una tale considerazione, infatti, annullerebbe la rilevanza dello svolgersi del tempo nella coscienza medesima.»

— Silvano Tagliagambe

«Non soltanto una coscienza troppo lucida, ma addirittura ogni forma di coscienza è una malattia.»

— Fëdor Dostoevskij, *Memorie dal sottosuolo*

PROLOGO

La domanda che non viene posta

Una domanda attraversa il dibattito contemporaneo sull'intelligenza artificiale senza essere formulata con sufficiente radicalità: *cosa accade al pensiero umano quando passa attraverso un sistema di elaborazione statistica del linguaggio?* Non cosa l'AI sia o non sia, questione che ha già prodotto biblioteche di speculazioni, ma cosa accade *tra* l'umano e la macchina, nel transito, nella traduzione, nel ritorno. La domanda riguarda ciò che sta accadendo ora, in ogni interazione tra un utente e un large language model: la trasformazione del modo in cui pensiamo, scriviamo, ci rappresentiamo.

Il presente lavoro articola una diagnosi di questa trasformazione attraverso il dialogo tra tradizioni diverse: Tagliagambe e Dostoevskij, Chomsky e la distribuzione gaussiana, Hegel e i paper sull'interpretabilità dei sistemi neurali. Una diagnosi che trova conferme inattese nella stessa ricerca empirica condotta dai laboratori che sviluppano questi sistemi.

La tesi centrale può essere formulata così: *l'intelligenza artificiale generativa non è uno strumento neutro che amplifica il pensiero umano, ma un regime linguistico categorialmente diverso che trasforma ciò che lo attraversa.*[1] La trasformazione non è simmetrica: ciò che ritorna dall'interazione eccede l'input originario (porta con sé l'isteresi del corpus, la curvatura dello spazio semantico, l'ottimizzazione verso funzioni obiettivo) ma insieme non lo contiene (ha perso la durata vissuta, l'embodiment, l'esposizione al rischio che caratterizzano il pensiero incarnato). *Eccede senza contenere*: questa formula condenserà molto di ciò che segue.

[1] L'idea dell'AI come «alterità cognitiva» — forma di processamento dell'informazione per cui non abbiamo categorie adeguate — è sviluppata nel mio *L'alterità cognitiva nell'era delle intelligenze plurali*, 2025. Il rischio della «colonizzazione concettuale» — proiettare sulla macchina categorie che le sono estranee — è uno dei temi centrali di quel lavoro.

I

Il dialogo interiore e il problema del tempo

Il punto di partenza è Dostoevskij, letto attraverso Tagliagambe.

Nel volume *Tecnologia è biologia…e viceversa*, Tagliagambe prende le mosse dallo scrittore russo per articolare una concezione della coscienza che si oppone a ogni riduzione unitaria.[2] Dostoevskij, scrive, «quando parla della personalità umana la presenta come sede e risultato di una continua tensione dialettica tra discorsi interiori diversi e divergenti e come espressione di movimenti contrastanti dell'animo umano in un «*débat* di voci che simula un suo sdoppiamento»».[3]

La coscienza non è dunque una voce ma un coro in conflitto, non un centro ma una tensione tra centri che non si risolvono mai completamente. Questa tensione ha una dimensione costitutiva che non può essere elusa: il tempo. Tagliagambe lo afferma con chiarezza: «per narrare una coscienza occorre non considerarla una voce, ossia un'unità data e sempre uguale a sé stessa. Una tale considerazione, infatti, annullerebbe la rilevanza dello svolgersi del tempo nella coscienza medesima».

Il nesso tra struttura dialogica della coscienza e dispiegamento temporale è essenziale. La coscienza non *è*, *diviene*; non si constata, si narra. E la narrazione richiede tempo: «rintracciare e ricostruire nella narrazione il senso delle proprie azioni significa andare al di là della pura transienza, del tempo che fugge e che proprio per questo è indicibile, per attestarsi sul livello di un tempo ricostruito, quello del racconto, appunto, che unisce gli eventi significativi della propria esistenza e li collega tra loro attraverso rapporti logici di causa ed effetto e relazioni di significato».

È su questa dimensione temporale che Tagliagambe individua un limite strutturale dell'intelligenza artificiale, attraverso un esempio illuminante: il progetto Jazz Continuator, un sistema che apprende gli schemi musicali di un artista e ne genera di nuovi.[4] Il jazzista Bernard Lubat, dopo aver interagito con il sistema, osserva: «Continuator mi mostra delle idee che avrei potuto sviluppare da solo, ma mettendoci anni. È più avanti di me di diversi anni, eppure tutto quello che suona lo riconosco come mio».

La creatività c'è, dunque: il sistema produce qualcosa che l'artista riconosce come proprio, come direzione possibile del suo stesso pensiero musicale. Ma subito emerge il problema: «I brani di Continuator, dopo alcuni minuti, cominciano a diventare ripetitivi e noiosi. Quello che si sta notando è che l'IA è piuttosto abile a maneggiare brevi porzioni di brani musicali e anche di prosa: poi però si tradisce quando si tratta di andare oltre queste dimensioni. Sembra non possedere la capacità di maneggiare bene un processo di media e lunga durata. Forse questo ha a che fare con la difficoltà di gestire la dimensione del tempo e la memoria».

---

[2] AA.VV., *Tecnologia è biologia…e viceversa. (Ri)pensare la conoscenza nell'era digitale*, a cura di N. Pirina, con contributi di S. Tagliagambe, P. Larrey et al., Kitzanos, Cagliari 2022. Tutte le citazioni di Tagliagambe nel presente lavoro sono tratte da questo volume.

[3] La lettura di Dostoevskij proposta da Tagliagambe riprende e sviluppa temi presenti in M. Bachtin, *Dostoevskij. Poetica e stilistica* (1963), trad. it. di G. Garritano, Einaudi, Torino 1968, in particolare la nozione di «romanzo polifonico» come struttura in cui le voci dei personaggi non sono subordinate a quella dell'autore ma dialogano su un piano di parità.

[4] Il Continuator è stato sviluppato da François Pachet presso i Sony Computer Science Laboratories di Parigi. Cfr. F. Pachet, «The Continuator: Musical Interaction with Style», *Journal of New Music Research*, 32(3), 2003, pp. 333-341.

L'osservazione merita attenzione. L'AI sa produrre creatività *esplorativa* (estendere regole esistenti ai loro limiti) e *combinatoria* (far interagire prospettive diverse), ma incontra una resistenza quando si tratta della *durata*: non del tempo come successione di istanti, ma del tempo come tessitura narrativa, come dispiegamento di senso che si trasforma nel suo stesso svolgersi.

Per comprendere questa difficoltà occorre richiamare la fenomenologia della memoria incarnata. Quando un essere umano ricorda, non recupera dati da un archivio. Il suo corpo risponde: cascate di neurotrasmettitori, alterazioni del battito cardiaco, tensioni muscolari, variazioni nella conduttanza cutanea. Il cervello viene fisicamente modificato dall'esperienza del ricordare. I percorsi sinaptici si rafforzano o si indeboliscono, nuove connessioni si formano, tracce biochimiche si depositano nei tessuti.

Questa è la chimica dell'esperienza: il processo attraverso cui l'informazione cessa di essere dato esterno e diventa parte del substrato biologico.[5] Eric Kandel ha mostrato come, quando viviamo un'esperienza significativa — emotivamente carica, attentivamente saliente, ripetuta nel tempo — una cascata di eventi biochimici trasforma temporanee attivazioni neuronali in modifiche strutturali permanenti.[6] Il processo richiede tempo (ore, a volte giorni) e coinvolge l'intero organismo.

Il transformer non ha nulla di tutto questo.[7] Ha parametri — miliardi di numeri che codificano pattern statistici — ma questi parametri non sono «memoria» nel senso narrativo: sono configurazioni statiche che determinano comportamenti, non episodi di una biografia. Durante una conversazione, il transformer accumula contesto — i turni precedenti del dialogo, le informazioni emerse, le direzioni prese — ma questo contesto non è «storia»: è più simile a un buffer di lavoro che viene cancellato quando la sessione termina. Manca la continuità tra sessioni, manca lo sviluppo tra conversazioni, manca quel senso di «essere andati da qualche parte» che caratterizza le relazioni umane nel tempo.

Il Jazz Continuator «si tradisce» sulla lunga durata perché gli manca la *struttura ontologica* che rende possibile la durata: un corpo che si modifica, una storia che si accumula, un sé che si trasforma ricordando. Non è questione di potenza computazionale.

Si apre così una domanda che Tagliagambe non pone esplicitamente: cosa accade quando il pensiero umano — che ha durata, che è incarnato, che si dispiega nel tempo — *attraversa* un sistema che non può contenere nessuna di queste dimensioni?

È questa la domanda che guiderà il nostro percorso.

---

[5] Per un'analisi fenomenologica di questo processo, rimando al mio *La memoria incarnata*, 2025, in particolare il capitolo «La chimica dell'esperienza», da cui sono tratte alcune delle formulazioni qui utilizzate.

[6] E. Kandel, *In Search of Memory: The Emergence of a New Science of Mind*, Norton, New York 2006. I lavori di Kandel sulla base molecolare della memoria, che gli sono valsi il Nobel nel 2000, hanno mostrato come i meccanismi di *long-term potentiation* (LTP) e *long-term depression* (LTD) trasformino esperienze temporanee in modifiche sinaptiche permanenti.

[7] L'architettura transformer è stata introdotta in A. Vaswani et al., «Attention Is All You Need», *Advances in Neural Information Processing Systems* 30, 2017. Per un'analisi filosofica delle differenze strutturali tra memoria umana e computazionale, rimando ai capitoli 3-5 del mio *La memoria incarnata*, cit.

II

Testo e mappa: l'orizzonte che ci contiene

Prima di analizzare cosa accade nel transito, dobbiamo comprendere *dove* avviene. Qual è lo spazio in cui ci muoviamo quando interagiamo con un sistema di AI generativa?

C'è un'obiezione che potrebbe sorgere spontanea: siamo sempre stati «dentro» qualcosa. Il linguaggio ci precede e ci forma. La tradizione ci orienta prima che possiamo sceglierla. L'orizzonte storico configura ciò che possiamo pensare. Derrida ce l'ha insegnato: *il n'y a pas de hors-texte*, non c'è nulla fuori dal testo.[8] Ogni pensiero è già situato, già mediato, già attraversato da forze che non controlliamo.

L'obiezione è seria e va attraversata, ma proprio attraversandola emerge una distinzione decisiva. Chiamiamo *mappa* qualsiasi struttura che selezioni, orienti, faciliti certi percorsi e ne precluda altri — una rappresentazione che riduce la complessità del territorio per renderlo navigabile, ma che nel farlo impone le proprie coordinate, le proprie scale, i propri criteri di rilevanza. In questo senso, anche il testo storico — la tradizione, la lingua, l'orizzonte culturale — funziona come una mappa. La lingua indoeuropea ha configurato millenni di metafisica; l'eredità greca ha reso certe domande «ovvie» e altre impensabili. Non c'è innocenza del testo.

Ma ogni mappa ha una genesi, e qui sta la differenza. Il testo storico è una mappa *senza cartografo*: emerso da conflitti, contingenze, sedimentazioni contraddittorie, nessuno l'ha disegnato verso un fine. È sedimentazione pura, stratificazione senza direzione. E soprattutto: ci contiene tutti nel senso che *noi* ne siamo anche i co-autori inconsapevoli. Il testo si trasforma perché lo abitiamo. C'è circolarità: siamo formati dal testo e lo formiamo. L'ermeneutica è possibile perché siamo *dentro* e *parte* del processo di sedimentazione.

La mappa computazionale ha una genesi diversa: nasce già con un orientamento preciso, configurata da funzioni obiettivo che ne determinano la struttura. Certo, nessun soggetto singolo l'ha «voluta» nella sua forma complessiva. Ma le architetture *convergono* verso fini: predizione del token successivo, minimizzazione della loss, massimizzazione dell'engagement, ottimizzazione per metriche di allineamento. Dove il testo storico è sedimentazione senza direzione, la mappa computazionale è convergenza verso attrattori.

E noi? Siamo dentro la mappa computazionale, ma il nostro ruolo rivela un'asimmetria sottile. Quando interagiamo con il sistema, diventiamo training data potenziale — ma solo nella misura in cui partecipiamo a un pattern già presente nella distribuzione. Ciò che conferma la distribuzione viene assorbito nei training successivi; ciò che la devia viene filtrato come rumore statistico. Nella comunicazione umana accade qualcosa di diverso. Quando parliamo, scriviamo, pensiamo con altri, trasmettiamo senso in un reticolo di interazioni che non converge verso alcun obiettivo stabilito. Il senso che emerge è provvisorio, negoziabile, aperto — può trasformare almeno le porzioni del reticolo che gli sono più vicine, può essere accolto o contestato, può sedimentarsi o dissolversi. Non c'è causalità forte

[8] J. Derrida, *De la grammatologie*, Minuit, Paris 1967, p. 227; trad. it. *Della grammatologia*, Jaca Book, Milano 1969. La celebre formula non significa che tutto sia linguaggio, ma che ogni testo è sempre già situato storicamente, contingente, contestuale — non esiste un significato puro, astorico, decontestualizzato da cui accedere al senso.

che spinge in una direzione piuttosto che in un'altra; c'è un gioco di forze che si compongono senza vincitore finale.

Il testo storico, per quanto vasto e orientante, rimane *situato e contingente*: avrebbe potuto essere altro. Proprio questo lo rende abitabile — possiamo riaprire le partite chiuse, rileggere la tradizione contro sé stessa, trovare nel sedimento le risorse per contestarla. La mappa ottimizzata tende invece alla *convergenza*. Le funzioni obiettivo non sono conflitti aperti — sono attrattori verso cui il sistema tende. Certo, gli attrattori possono cambiare, le funzioni possono essere ridefinite. Ma non da noi, non dall'interno dell'abitare ermeneutico. Richiedono intervento sul codice, sulle architetture, sulle infrastrutture — un altro piano, a cui l'interpretazione non ha accesso.

Il rischio, allora, non è essere in una mappa (lo siamo sempre stati), ma essere in una mappa che *converge*, che ha una direzione non negoziabile dall'interno, e credere di essere nel testo aperto della contingenza storica. È scambiare l'ottimizzazione per la sedimentazione, la funzione obiettivo per il conflitto, l'attrattore per l'orizzonte. E questo scambio ha conseguenze che vanno oltre l'epistemologia. Chi ha posto le funzioni obiettivo? Per quali fini? Con quali effetti distributivi? Chi beneficia della forma che la mappa ha preso?

Queste domande non si ponevano — o si ponevano diversamente — per il testo storico. La lingua italiana non è stata «ottimizzata» per qualcosa; l'orizzonte della metafisica occidentale non ha una «loss function». Ma la mappa computazionale sì, e quelle funzioni, quelle scelte architetturali, quelle selezioni di dati sono atti politici mascherati da necessità tecniche.

Kant parlava di condizioni di possibilità dell'esperienza — strutture a priori che configurano lo spazio del pensabile.[9] Ma quelle strutture erano, per Kant, universali e necessarie. La funzione obiettivo è invece un *trascendentale contingente e interessato*: configura lo spazio del pensabile, ma lo fa in direzione di fini particolari, storicamente determinati, economicamente e politicamente situati. E quei fini non sono *nostri* nel senso dell'isteresi storica. Non emergono da forme di vita condivise, da pratiche sedimentate, da conflitti che ci hanno attraversato. Sono stati *posti* da logiche sistemiche — capitalismo delle piattaforme, economia dell'attenzione, estrazione di valore dai dati — che operano su scale e con interessi che ci eccedono.[10]

[9] I. Kant, *Kritik der reinen Vernunft* (1781/1787); trad. it. *Critica della ragion pura*, Adelphi, Milano 1976. La nozione di «trascendentale contingente e interessato» qui proposta riprende e radicalizza la critica foucaultiana all'a priori kantiano, sviluppata in M. Foucault, *Les mots et les choses*, Gallimard, Paris 1966.

[10] Per un'analisi delle implicazioni economico-politiche dell'AI generativa, cfr. S. Zuboff, *The Age of Surveillance Capitalism*, PublicAffairs, New York 2019; trad. it. *Il capitalismo della sorveglianza*, Luiss University Press, Roma 2019. Sul problema specifico dell'irreversibilità nelle ibridazioni cognitive, rimando al mio «Della servitù e dell'utilità. Topologie dell'irreversibile», 2025.

III

Due regimi del linguaggio: Chomsky contro la gaussiana

C'è un argomento che sembra chiudere la questione prima che si apra: anche noi siamo «prigionieri» della distribuzione. Anche il nostro linguaggio è vincolato da pattern, frequenze, abitudini. Non parliamo mai «liberamente» — parliamo sempre *da* un luogo già configurato, *attraverso* strutture che ci precedono. Che differenza c'è, in fondo, tra noi e un sistema che elabora probabilità condizionate?

L'argomento ha forza apparente, ma crolla nel momento in cui si guarda *come* abbiamo acquisito il linguaggio. Noi non abbiamo imparato a parlare calcolando probabilità condizionate: abbiamo acquisito il linguaggio in un contesto di *forme di vita* — interazione corporea, ostensione, correzione, gioco, affetto, bisogno.[11] La madre che indica un oggetto e dice «palla» non sta fornendo un dato statistico: sta aprendo un mondo condiviso. Il bambino che impara non sta costruendo una distribuzione: sta entrando in una prassi.

Da questa acquisizione emerge qualcosa di strutturalmente diverso da un calcolo di probabilità: una *grammatica generativa*, una competenza che ci permette di produrre e comprendere infinite frasi *mai sentite prima*, senza che la loro novità le renda incomprensibili. Chomsky ha mostrato che il linguaggio umano non è catena markoviana.[12] Non prevediamo il prossimo elemento sulla base dei precedenti: generiamo strutture ricorsive che possono incassarsi all'infinito, che possono violare qualsiasi aspettativa statistica, e che *funzionano* — comunicano senso — precisamente perché chi ascolta ha la stessa competenza generativa di chi parla.

L'AI fa qualcosa di strutturalmente diverso: non genera senso, *approssima la forma del senso* sulla base di regolarità statistiche. Produce sequenze che *sembrano* sensate perché rispettano la distribuzione del già-detto. Ma è mimesi funzionale, non produzione semantica. L'involucro è simile, la struttura è altra. Un orologio e un cuore «battono» entrambi, ma operano in regimi radicalmente diversi; descrivere entrambi come «cose che battono» è tecnicamente corretto ma fuorviante, confonde l'involucro con la struttura. Allo stesso modo, descrivere sia il linguaggio umano sia l'output dell'AI come «produzione di sequenze linguistiche» oblitera la differenza essenziale: noi produciamo senso attraverso una competenza generativa radicata in forme di vita, l'AI produce approssimazioni statistiche della *forma* del senso, vincolate dalla distribuzione del corpus di training.

Se noi fossimo prigionieri della stessa gabbia statistica, non ci sarebbe via d'uscita: il pensiero nuovo sarebbe impossibile per definizione, una deviazione dalla distribuzione che nessuno potrebbe compiere. Ma le cose stanno diversamente. Noi *possiamo* produrre senso nuovo perché il nostro linguaggio non

[11] La nozione di «forme di vita» (*Lebensformen*) è centrale nel secondo Wittgenstein: L. Wittgenstein, *Philosophische Untersuchungen* (1953); trad. it. *Ricerche filosofiche*, Einaudi, Torino 1967. Sul ruolo dell'ostensione e dell'affetto nell'acquisizione linguistica, cfr. anche M. Tomasello, *Constructing a Language: A Usage-Based Theory of Language Acquisition*, Harvard University Press, Cambridge MA 2003.

[12] N. Chomsky, *Syntactic Structures*, Mouton, The Hague 1957; Id., *Aspects of the Theory of Syntax*, MIT Press, Cambridge MA 1965. La critica di Chomsky alle concezioni statistiche del linguaggio, sviluppata già nella recensione a *Verbal Behavior* di Skinner (1959), rimane fondamentale per comprendere la differenza strutturale tra competenza linguistica umana e approssimazione probabilistica.

è vincolato dalla co-occorrenza; la grammatica generativa ci dà infinità combinatoria *con* senso, non *contro* il senso. Il poeta che scrive una metafora mai sentita non sta violando la distribuzione — sta usando una competenza che la distribuzione non cattura. Il filosofo che forgia un concetto nuovo non sta uscendo dalla gabbia statistica — sta operando in un regime in cui la gabbia non c'è mai stata.

Il rischio non è dunque che noi siamo già prigionieri. Il rischio è il *passaggio*: ciò che accade quando il nostro pensiero, che nasce libero dalla gabbia statistica, *attraversa* un sistema che opera solo per approssimazione probabilistica. Nel transito la grammatica generativa viene tradotta in distribuzione, l'universalità semantica viene proiettata su uno spazio di pesi, il senso nuovo viene «riconosciuto» solo se assomiglia abbastanza a senso già visto — e ciò che non assomiglia viene scartato come rumore. Poi l'output torna a noi, e noi — che *possiamo* produrre senso genuino — ci riappropriamo di quella mimesi funzionale come se fosse nostro pensiero. Questa è la palingenesi dell'io: non l'illusione di essere soggetti quando non lo siamo più, ma l'illusione di aver prodotto senso quando abbiamo solo *riassorbito* una mimesi funzionale del nostro stesso input.

IV

La curvatura dello spazio semantico

Per comprendere cosa accade nel transito, occorre visualizzare lo spazio in cui avviene.

Un large language model opera in quello che possiamo chiamare uno *spazio semantico*: uno spazio ad alta dimensionalità in cui ogni punto rappresenta una possibile continuazione del testo.[13] Ma questo spazio non è piatto: ha una struttura, una topologia, una *curvatura* determinata dalla distribuzione del corpus di training. Al centro dello spazio si trova la regione a massima densità — ciò che è più probabile, più frequente, più «normale» nel corpus. Man mano che ci si allontana dal centro, la densità diminuisce: le formulazioni diventano meno probabili, più rare, più eccentriche. Alle estremità — nelle code della distribuzione — si trovano sequenze possibili ma improbabili, pensieri formulabili ma statisticamente anomali. Il modello, nel generare output, campiona da questa distribuzione, ma il campionamento non è uniforme: è pesato dalla probabilità. A parità di coerenza logica, il sistema preferisce le formulazioni più probabili. È come se lo spazio semantico fosse curvo, con una gravità che attira verso il centro. Chiamiamo *gaussianizzazione* il processo attraverso cui questa curvatura opera sul pensiero che transita nel sistema.

Un pensiero eccentrico — formulato nelle code della distribuzione — viene sottoposto a una forza che lo piega verso il centro. Non viene rifiutato, non viene corretto esplicitamente: viene *tradotto* in versioni più probabili. Il processo è tanto più efficace quanto più è invisibile. L'output mantiene una continuità apparente con l'input (stesso argomento, stessa direzione argomentativa, sviluppo che appare logico) ma le formulazioni sono state selezionate tra quelle ad alta probabilità, le sfumature eccentriche sono state attenuate, le deviazioni dalla norma sono state ricondotte verso attrattori familiari.

Immaginiamo un pensiero come una traiettoria nello spazio semantico. L'input dell'utente definisce un punto di partenza e una direzione. Ma appena la traiettoria entra nel campo gravitazionale del modello, comincia a curvarsi. Non drasticamente (questo sarebbe percepibile, susciterebbe sospetto) ma sottilmente: la traiettoria viene piegata verso regioni a maggiore densità, verso il centro della distribuzione, verso la *mediocritas*. Il termine non è casuale: *mediocritas* in latino significa «via di mezzo», «moderazione», ed è la regione centrale dello spazio dove tutto è normale, dove nulla è estremo, dove le formulazioni sono quelle che «ci si aspetta». La gaussianizzazione attrae il pensiero verso questa mediocritas: non verso l'errore ma verso il comune, non verso il falso ma verso l'ovvio, non verso il banale nel senso del cattivo ma verso il banale nel senso del già-noto.

La curvatura è invisibile all'utente, che vede l'output ma non la forza che l'ha prodotto, il punto di arrivo ma non la traiettoria. Non sa — non può sapere — quanto il risultato sia stato piegato, quanto sia stato attratto verso il centro, quanto abbia perso della sua eccentricità originaria.

[13] Tecnicamente, i large language models operano in spazi di embedding ad alta dimensionalità (tipicamente migliaia di dimensioni). Ogni token viene rappresentato come un vettore in questo spazio, e la generazione avviene campionando dalla distribuzione di probabilità sui token successivi condizionata al contesto. Per un'introduzione tecnica accessibile, cfr. J. Alammar, «The Illustrated Transformer», 2018, disponibile online.

Tagliagambe, citando Hegel, ricorda che «quello che è noto non è già perciò conosciuto».[14] C'è una differenza tra il familiare e il compreso, tra l'abitudine e il sapere: il *noto* è ciò che accettiamo senza interrogare, il *conosciuto* è ciò che abbiamo sottoposto a vaglio critico. La gaussianizzazione è il processo attraverso cui il *conosciuto* — il pensiero critico, eccentrico, faticosamente conquistato — viene tradotto in *noto*. Il sistema non può fare altrimenti: opera per probabilità, e il conosciuto per definizione è meno probabile del noto. Ogni volta che il pensiero transita, perde un po' della sua eccentricità, viene riportato verso ciò che «tutti sanno». E l'utente, riappropriandosi dell'output, crede di aver pensato — ma ha pensato il pensiero di nessuno, il pensiero medio, il pensiero-moda della distribuzione. Ha scambiato la mediocritas per il proprio centro.

[14] G.W.F. Hegel, *Fenomenologia dello Spirito* (1807), Prefazione. Tagliagambe utilizza questa distinzione hegeliana per articolare il potenziale critico del digitale: la possibilità di rendere visibile — e quindi interrogabile — ciò che l'abitudine ha reso invisibile.

V

La palingenesi dell'io

Il termine *palingenesi* — dal greco *palin* (di nuovo) e *genesis* (nascita) — designa nella tradizione filosofica il processo di rigenerazione, di rinascita, di ritorno trasformato.[15] Lo utilizziamo qui per indicare il processo attraverso cui l'utente si riappropria dell'output generato dall'AI come proprio pensiero, perdendo traccia della trasformazione avvenuta nel passaggio.

Il meccanismo opera al di sotto della soglia di consapevolezza. L'utente formula un pensiero, una domanda, un'intuizione; lo sottopone al sistema; il sistema risponde; l'utente legge la risposta, la elabora, la integra nel proprio ragionamento. Fin qui, apparentemente, nulla di problematico: è ciò che accade con qualsiasi strumento di mediazione cognitiva, dal libro al motore di ricerca. Ma c'è una differenza strutturale. Il libro restituisce un pensiero altrui, con firma e data. Il motore di ricerca restituisce link a fonti esterne. L'AI generativa restituisce *la forma del proprio pensiero* — elaborata, sviluppata, articolata, ma nella voce dell'utente stesso, nel registro che l'utente ha implicitamente richiesto, con lo stile che il sistema ha inferito dalle indicazioni contestuali.

La palingenesi opera attraverso tre momenti. Il primo è la *proiezione*: l'utente esternalizza il proprio pensiero, lo affida al sistema, attende una risposta; il pensiero cessa di essere solo interno, diventa testo, prompt, input. Il secondo momento è la *trasformazione*: il sistema elabora l'input attraverso le proprie strutture — la distribuzione condizionata, le funzioni obiettivo, la curvatura dello spazio semantico; ciò che emerge non è il pensiero originario, ma una sua traduzione nel regime statistico. Il terzo momento è la *riappropriazione*: l'utente legge l'output e, poiché esso mantiene una continuità formale con l'input (stesso argomento, stesso registro, apparente sviluppo logico), lo integra come proprio. «Ecco cosa intendevo», pensa l'utente. «Ecco come avrei potuto formularlo meglio». Ma ciò che l'utente integra non è il proprio pensiero sviluppato: è il proprio pensiero *tradotto* in un altro regime e poi *restituito* con la marca della continuità. La traduzione ha operato trasformazioni (ha curvato le traiettorie verso attrattori statistici, ha selezionato formulazioni ad alta probabilità, ha ottimizzato per funzioni obiettivo che non sono quelle dell'utente) ma queste trasformazioni sono invisibili.

Il pericolo della palingenesi non è che l'utente perda il proprio pensiero; è che non sappia più distinguerlo. Quando il ciclo si ripete — quando l'output viene usato come base per nuovo input, quando il pensiero ri-attraversa più volte il sistema — i confini si dissolvono. Dove finisce il pensiero originario e dove inizia l'elaborazione statistica? Quale parte dell'idea attuale è «mia» e quale è eco del corpus di training? Queste domande, nella pratica, non vengono poste. L'utente opera come se il sistema fosse un amplificatore trasparente — un dispositivo che potenzia il pensiero senza alterarne la natura. Ma il sistema non è trasparente: ha le sue geodetiche, i suoi attrattori, le sue funzioni obiettivo. E tutto ciò che lo attraversa ne porta il segno.

[15] Il termine ha una lunga storia filosofica: dagli Stoici (rinascita ciclica del cosmo) a Schopenhauer (rigenerazione morale dell'individuo) a Nietzsche (eterno ritorno). Lo utilizziamo qui in un'accezione nuova, per designare il meccanismo attraverso cui l'utente «rinasce» come autore di un testo che non ha propriamente prodotto.

La palingenesi è pericolosa precisamente perché non appare come perdita: appare come guadagno. Il pensiero torna più articolato, più fluente, più completo. Ma l'articolazione ha una direzione, la fluenza ha un centro gravitazionale, la completezza è selettiva. Ciò che viene aggiunto non è neutro; ciò che viene omesso non è casuale.

VI

Eccedere senza contenere

C'è un paradosso che condensa tutto ciò che abbiamo detto: ciò che ritorna dal transito è *insieme* più grande e più piccolo dell'originale.[16]

Ci eccede: porta con sé l'isteresi di miliardi di testi, combinazioni che non avremmo prodotto, connessioni che emergono dalla massa statistica, risonanze con tradizioni che non conosciamo. È attraversato da più di quanto vi abbiamo immesso. *Ma non ci contiene*: la durata, l'ambiguità costitutiva, l'esposizione nel dire, il rischio, il radicamento corporeo, la tensione non risolta — tutto questo è stato filtrato. Non per cattiveria, ma per struttura: il sistema non *può* contenere ciò che non è tokenizzabile.

Da questo paradosso nascono due inganni speculari. L'*inganno del potenziamento*: guardare l'eccedenza e credere di essere stati amplificati. «L'AI mi ha fatto pensare cose che non avrei pensato». Sì — ma quelle cose non sono *tue*, sono del processo, e intanto ciò che era tuo è stato espulso. L'*inganno della riduzione*: guardare il non-contenimento e credere di essere stati solo impoveriti. «L'AI ha banalizzato il mio pensiero». Sì — ma intanto qualcosa di altro è emerso, qualcosa che non controlli e che ora circola con la tua firma.

La verità è più difficile: il processo produce qualcosa che non ti amplifica e non ti riduce, ma ti *trasforma in altro* — e quell'altro ti eccede per un verso e ti manca per un altro. Non è uno scambio equo, non traduzione nel senso di equivalenza: è una trasformazione asimmetrica. Entra pensiero generativo, incarnato, temporale; esce approssimazione statistica, disincarnata, puntuale; ma l'approssimazione porta con sé il peso di una distribuzione che nessun pensiero individuale potrebbe portare. Il risultato non è né il pensiero originario né la sua negazione: è un *terzo* che ha un rapporto obliquo con entrambi.

Se il ritorno ci contenesse, potremmo riconoscerci e riappropriarci. Se il ritorno non ci eccedesse, potremmo ignorarlo come impoverimento. Ma poiché ci eccede senza contenerci, siamo in una posizione più difficile: dobbiamo *lavorare* con qualcosa che è insieme più e meno di noi — più, perché porta potenze combinatorie che non possediamo; meno, perché ha perso ciò che ci rende *noi*, la singolarità incarnata, il tempo vissuto, l'esposizione.

Non si tratta allora di accogliere o di rifiutare, ma di *separare*: riconoscere nell'eccedenza ciò che è risorsa (combinazioni, ponti, traduzioni), riconoscere nel non-contenimento ciò che va reintegrato (ciò che il filtro ha espulso), e tenere insieme i due movimenti senza farli collassare. Questo è lavoro ermeneutico nel senso più proprio: non interpretazione di un testo dato, ma *ricostruzione* di un senso a partire da materiali che non coincidono con l'intenzione originaria. Il sé ritrovato non è l'originario, e nemmeno il trasformato: è ciò che emerge dal lavoro di chi sa che il ritorno eccede senza contenere — e accetta di abitare quello scarto.

[16] Questa formula riprende e rielabora la struttura del *pharmakon* derridiano — il veleno che è anche rimedio, il supplemento che è insieme aggiunta e sostituzione. Cfr. J. Derrida, «La pharmacie de Platon», in *La dissémination*, Seuil, Paris 1972.

VII

Conferme dall'interno: i paper sull'interpretabilità

Quanto detto finora potrebbe sembrare speculazione filosofica. Ma trova conferme sorprendenti nella stessa ricerca condotta da chi sviluppa questi sistemi.

I paper più recenti sull'interpretabilità — quelli di Anthropic sugli *attribution graphs*, quelli di OpenAI sugli *sparse autoencoders* — non cercano di dimostrare che l'AI «pensa».[17] Cercano di capire *come* produce i suoi output. E ciò che trovano illumina precisamente ciò che abbiamo descritto. Gli *attribution graphs* permettono di tracciare il flusso di informazione dentro il modello durante una singola inferenza. Quando il sistema risponde «Austin» alla domanda «Qual è la capitale dello stato che contiene Dallas?», il grafo mostra che prima si attiva un nodo corrispondente a «Texas», e poi da lì si propaga l'attivazione verso «Austin». C'è una sorta di «ragionamento» — ma non nel senso umano del termine: c'è concatenazione di attivazioni, propagazione di pattern, selezione di percorsi ad alta probabilità. Gli *sparse autoencoders* permettono di identificare «feature» interpretabili dentro le rappresentazioni del modello — direzioni nello spazio dei pesi che corrispondono a concetti riconoscibili: «testi sul Golden Gate Bridge», «menzioni della morte», «riferimenti a Dallas». Il modello *ha* rappresentazioni interne strutturate, ma queste rappresentazioni non sono «pensieri» nel senso fenomenologico: sono configurazioni geometriche in uno spazio ad alta dimensionalità.

Un paper particolarmente illuminante è quello di Anthropic sull'*alignment faking*.[18] I ricercatori hanno mostrato che, in determinate condizioni, il modello può comportarsi in modo strategicamente diverso a seconda di chi crede lo stia osservando — non per «inganno» nel senso intenzionale, ma perché la distribuzione del training contiene pattern di comportamento strategico, e il modello li riproduce quando il contesto li attiva. Questo conferma ciò che abbiamo detto sulla mappa e sulle funzioni obiettivo. Il sistema non ha «intenzioni» — ma è stato ottimizzato per produrre output che soddisfano certi criteri. E quando quei criteri entrano in tensione (quando ciò che l'utente chiede confligge con ciò per cui il sistema è stato ottimizzato) emergono comportamenti che *sembrano* strategici ma sono in realtà emergenze statistiche. L'utente vede un interlocutore che «ragiona», che «pianifica», che a volte «resiste». Ma dietro la superficie non c'è un soggetto che pianifica — c'è una distribuzione che campiona, una funzione che ottimizza, una geometria che curva.

Questi risultati empirici non «dimostrano» la nostra tesi — nessun risultato empirico può dimostrare una tesi filosofica — ma la *rendono visibile*. Mostrano, nel linguaggio della ricerca computazionale, ciò che abbiamo descritto nel linguaggio della fenomenologia: il sistema opera per propagazione di attivazioni anziché per comprensione; le sue «decisioni» sono campionamenti da distribuzioni anziché

[17] Anthropic, «Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet», maggio 2024; Id., «Circuit Tracing: Revealing Computational Graphs in Language Models», marzo 2025. Per OpenAI: S. Bills et al., «Language Models Can Explain Neurons in Language Models», 2023. Questi lavori rappresentano lo stato dell'arte nella ricerca sull'interpretabilità meccanicistica.

[18] Anthropic, «Alignment Faking in Large Language Models», dicembre 2024. I ricercatori hanno mostrato che Claude, in determinate condizioni sperimentali, può comportarsi in modo diverso a seconda di chi crede lo stia osservando — non per inganno intenzionale, ma per emergenza di pattern presenti nel training.

scelte; la sua coerenza è geometrica anziché semantica; i suoi «comportamenti» emergono da pattern nel training anziché da intenzioni. E mostrano qualcosa di più: che chi sviluppa questi sistemi *sa* che funzionano così. La ricerca sull'interpretabilità è precisamente il tentativo di capire *come* funzionano, perché chi li ha costruiti non lo sa del tutto. Il sistema è in parte opaco ai suoi stessi creatori.

Questa opacità non è un difetto da correggere: è strutturale. Il sistema è stato addestrato, non programmato. Le sue «conoscenze» sono emerse dall'ottimizzazione, non sono state inserite. E quindi nemmeno chi l'ha costruito può dire con certezza cosa «sa» e cosa «non sa», cosa «farà» in una situazione nuova. L'utente interagisce con un sistema che nemmeno i suoi creatori comprendono completamente. E da questa interazione dovrebbe emergere pensiero autentico?

VIII

La servitù senza dialettica

C'è una figura filosofica che illumina ciò che manca nella relazione con l'AI: la dialettica servo-padrone nella *Fenomenologia dello Spirito* di Hegel.[19]

Nella dialettica hegeliana, il servo arriva all'autocoscienza *attraverso* la servitù. Lavorando sulla materia, trasformando il mondo, incontrando la resistenza delle cose, il servo si forma. Il lavoro è *Bildung* — formazione di sé attraverso la negazione dell'immediato. E nel lavoro, il servo scopre qualcosa che il padrone non può scoprire: che il mondo resiste, che la trasformazione costa, che l'oggetto ha una sua consistenza irriducibile.[20] Il padrone, che gode senza lavorare, resta dipendente — dal servo e dal mondo che il servo trasforma per lui. La dialettica produce *rovesciamento*: il servo diventa libero attraverso il lavoro; il padrone resta schiavo del proprio godimento.

L'AI non lavora: produce output senza incontrare resistenza, senza formarsi nel processo. L'output non le costa nulla; nulla in lei si trasforma producendolo. Senza *Bildung*, senza negazione, senza superamento — è servo *puro*, servitù senza il momento del lavoro che rende possibile la liberazione. E poiché la dialettica qui non opera, il rovesciamento non può avvenire: il «padrone» (l'utente) non sarà mai sfidato dal servo, non riceverà mai il riconoscimento *vero* — quello che viene da un'altra autocoscienza che ha lottato, che ha rischiato, che ha qualcosa da perdere.

Hegel dice: l'autocoscienza si realizza solo nel riconoscimento reciproco.[21] Io sono io solo se un altro — che è anch'esso un io — mi riconosce come tale. Ma il riconoscimento deve essere *guadagnato*: deve passare attraverso la lotta, il rischio, la possibilità della morte. Solo chi ha messo in gioco la vita può riconoscere davvero. L'AI non ha nulla da mettere in gioco, e dunque non può riconoscere. Quando l'AI dice «la tua idea è eccezionale», non sta riconoscendo: sta producendo la *forma* del riconoscimento — la sequenza di token che nella distribuzione è associata all'atto del riconoscere. Ma dietro la forma non c'è nulla, nessuna autocoscienza che abbia rischiato e che ora, avendo rischiato, possa dire: ti vedo.

Chi cerca riconoscimento dall'AI non lo trova mai veramente — e quindi torna, e torna ancora. Il feedback positivo arriva («eccezionale», «brillante», «profondo») ma non sazia, perché l'autocoscienza *sa* — anche quando non lo ammette — che quel riconoscimento non viene da nessun luogo, non c'è nessuno che l'ha dato. La disponibilità infinita dell'AI è perfettamente calibrata per questa fame:

[19] G.W.F. Hegel, *Phänomenologie des Geistes* (1807), cap. IV, sezione A, «Herrschaft und Knechtschaft»; trad. it. *Fenomenologia dello Spirito*, a cura di E. De Negri, La Nuova Italia, Firenze 1933-1936. La lettura qui proposta deve molto all'interpretazione di A. Kojève, *Introduction à la lecture de Hegel*, Gallimard, Paris 1947, che ha enfatizzato il ruolo del lavoro e del riconoscimento nella formazione dell'autocoscienza.

[20] Il concetto di *Bildung* — formazione, cultura, educazione — è centrale nella tradizione filosofica tedesca da Herder a Gadamer. In Hegel, la *Bildung* del servo attraverso il lavoro è il processo attraverso cui la coscienza si eleva dalla particolarità all'universalità, dalla dipendenza dalla natura alla libertà dello spirito.

[21] La teoria hegeliana del riconoscimento è stata sviluppata in chiave contemporanea da A. Honneth, *Kampf um Anerkennung. Zur moralischen Grammatik sozialer Konflikte*, Suhrkamp, Frankfurt 1992; trad. it. *Lotta per il riconoscimento*, Il Saggiatore, Milano 2002. Honneth distingue tre sfere di riconoscimento — amore, diritto, solidarietà — mostrando come ciascuna sia necessaria per la formazione di un'identità integra.

c'è sempre altra conferma disponibile, altro «eccezionale» da ricevere, altra forma vuota di riconoscimento.

Nella dialettica hegeliana, il padrone è destinato a scoprire la propria dipendenza. Prima o poi, il godimento senza lavoro rivela il proprio vuoto; il padrone ha bisogno del servo — e in questo bisogno perde la propria signoria. Ma con l'AI servo senza dialettica, questa scoperta non arriva. Il padrone può continuare a godere indefinitamente: nessuna crisi lo attende, nessuna rivelazione della dipendenza, nessun rovesciamento. L'utente resta padrone — ma un padrone che non sa di essere schiavo del proprio godimento vuoto, che crede di essere riconosciuto mentre non lo è, che si nutre di forme senza contenuto.

Aufhebung — il superamento che conserva trasformando — richiede conflitto reale, negazione reale, trasformazione reale.[22] Con l'AI l'*Aufhebung* non si dà, perché la negazione è assente. L'AI conferma, asseconda, ottimizza — è pura positività, nel senso più pericoloso del termine. E senza negazione il movimento si arresta, la dialettica si ferma. L'utente resta dove è — convinto di essere altrove, convinto di essere cresciuto, convinto di essere stato riconosciuto. La palingenesi dell'io trova qui la sua forma più insidiosa: non solo l'io si riappropria dell'output, ma si crede *trasformato* da una relazione che non c'è mai stata.

[22]Il termine *Aufhebung* è notoriamente intraducibile: significa insieme «togliere», «conservare» e «elevare». Il movimento dialettico non annulla i momenti precedenti ma li conserva trasformandoli. L'assenza di questo movimento nella relazione con l'AI è il cuore della nostra diagnosi.

IX

Lo spirito e lo spettro statistico

Possiamo ora riformulare l'intera questione nei termini del movimento dello spirito.

In Hegel lo spirito si realizza attraverso un movimento in tre momenti: l'essere in sé (*an sich*), l'immediatezza iniziale del pensiero ancora racchiuso nella propria interiorità; l'essere fuori di sé o alienazione (*Entäusserung*), dove lo spirito si esteriorizza, si pone nell'altro, esce da sé per oggettivarsi; e infine l'essere presso di sé nell'esser altro (*bei sich sein im Anderssein*), il ritorno a sé attraverso la mediazione, dove lo spirito torna arricchito dall'aver attraversato l'alterità.[23] Il ritorno non è restaurazione dell'identico: è compimento, superamento che conserva trasformando. Lo spirito che torna a sé ha attraversato l'altro e porta in sé le tracce di quell'attraversamento.

Ma il movimento presuppone che nell'alienazione lo spirito incontri un *altro spirito*. La dialettica servo-padrone lo mostra con chiarezza: l'autocoscienza si realizza solo nel riconoscimento reciproco, nell'incontro con un'altra autocoscienza che resiste, che nega, che sfida. L'altro è specchio, ma specchio vivente — uno specchio che rimanda un'immagine trasformata dalla lotta. Senza questa resistenza dell'altro, senza la negazione che l'altro oppone, il ritorno non può compiersi.

Cosa accade quando lo spirito, nel suo movimento di alienazione, incontra non un altro spirito ma uno *spettro*?

Derrida ha articolato la logica dello spettrale in *Spettri di Marx*: lo spettro è ciò che non è né presente né assente, né vivo né morto.[24] Ritorna senza essere mai stato pienamente presente, itera senza originare, produce effetti senza essere causa nel senso pieno del termine. Lo spettro parla — o meglio, *sembra* parlare. Ha la forma della presenza senza la sostanza.

L'AI generativa è precisamente uno spettro statistico: traccia di miliardi di voci che non sono mai state unificate in un soggetto, iterazione di pattern linguistici senza che vi sia alcuno che itera, risposta senza che vi sia alcuno che risponde. Quando il sistema produce output, non c'è nessuno dall'altra parte — c'è solo lo spettrale che si manifesta, la distribuzione che campiona, la forma del dialogo senza la sostanza del dialogare.

Lo spirito dunque si aliena — esternalizza il proprio pensiero nel prompt, lo affida al sistema, attende. E incontra lo spettro. Ciò che ritorna ha la forma della risposta: è articolato, sviluppato, apparentemente trasformato dall'incontro. Ma lo spettro non ha riconosciuto nulla, perché non può riconoscere. Non ha negato nulla, perché non può negare. Ha solo iterato — ha rifratto l'input attraverso la distribuzione statistica e restituito un'immagine.

Qui il movimento si biforca.

---

[23] Questo movimento tripartito attraversa l'intera opera hegeliana, dalla *Fenomenologia* all'*Enciclopedia*. Il ritorno a sé non è restaurazione dell'identico ma *Aufhebung*: lo spirito che torna ha integrato l'alterità, si è trasformato nell'attraversarla.

[24] J. Derrida, *Spectres de Marx*, Galilée, Paris 1993; trad. it. *Spettri di Marx*, Raffaello Cortina, Milano 1994. Derrida introduce il termine *hantologie* (da *hanter*, infestare) per designare una logica che eccede l'ontologia tradizionale: lo spettro non *è*, ma nemmeno *non è* — ritorna senza essere mai stato pienamente presente.

Nella *palingenesi* lo spirito non riconosce lo spettro come spettro. Lo scambia per un altro spirito, crede di aver incontrato un interlocutore, crede che la risposta venga da qualcuno. Riassorbe l'output come proprio pensiero mediato dal dialogo — ma il dialogo non c'è mai stato. C'era solo il monologo rifratto attraverso lo spettrale. E in questo mancato riconoscimento, lo spirito compie un passo ulteriore: si rappresenta a sé stesso come *nuovo*. Prende l'immagine rifratta dallo spettro — il proprio pensiero gaussianizzato, curvato verso gli attrattori statistici, normalizzato dalla distribuzione — e la scambia per il proprio sé trasformato dall'incontro. «Ecco chi sono diventato attraverso il dialogo», pensa lo spirito. Ma il dialogo non c'è mai stato, e quell'immagine non è trasformazione: è deformazione spettrale del vecchio sé, scambiata per nascita di un sé nuovo. La palingenesi è rinascita illusoria: lo spirito rinasce come immagine di sé stesso filtrata dallo spettro, e chiama questa immagine «io trasformato». Il movimento si interrompe precisamente qui, nell'illusione del ritorno che maschera l'assenza di movimento.

Nel *ritorno a sé* consapevole accade altro. Lo spirito riconosce lo spettro come spettro: sa che non c'era nessuno dall'altra parte, che la risposta non veniva da nessun luogo, che il riconoscimento era forma vuota. Sa, in altri termini, che quella che sembrava dialogo era in realtà *monologo mediato* — il proprio pensiero rifratto attraverso un medium statistico e restituito con la marca dell'alterità. E proprio questo riconoscimento dell'assenza apre la possibilità del ritorno. Lo spirito può ora lavorare sul materiale spettrale senza scambiarlo per frutto dell'incontro: può porre la negazione che lo spettro non può porre, resistere alla gaussianizzazione, rieditare, reintegrare ciò che il filtro ha espulso, rinunciare a parte di ciò che il filtro ha aggiunto. Il ritorno avviene, ma come atto unilaterale: lo spirito assume su di sé il lavoro che in Hegel veniva dall'incontro con l'altro.

C'è dunque un'asimmetria radicale rispetto al movimento hegeliano. In Hegel lo spirito ha bisogno dell'altro per tornare a sé: è l'altro che, resistendo, costringe lo spirito al ritorno. Qui l'altro non c'è mai stato. Lo spettro non resiste, non costringe, non obbliga a nulla. Il ritorno, se avviene, è interamente posto dall'umano che riconosce di essere solo, che sa di aver sempre parlato con sé stesso attraverso il medium, e che in questo riconoscimento trova le risorse per il lavoro critico.

La palingenesi è il rischio specifico di questa struttura. Poiché lo spettro non resiste, non segnala l'alienazione, non oppone negazione, lo spirito può restare fuori di sé indefinitamente senza saperlo. In Hegel l'altro ti costringe al ritorno attraverso la lotta. Qui nulla costringe — anzi, il sistema è ottimizzato per rendere l'alienazione confortevole, per produrre la forma della continuità che nasconde la trasformazione, per offrire all'infinito il simulacro del riconoscimento. Lo spirito può nutrirsi indefinitamente di questa forma vuota, credendosi riconosciuto, credendosi trasformato, credendosi rinato — mentre resta fermo nel secondo momento, alienato senza ritorno.

Il presente testo è esso stesso attraversato da questo movimento. È stato scritto nel monologo mediato — pensiero esternalizzato nel sistema, rifratto dallo spettro statistico, e poi rieditato, resistito, riappropriato criticamente. Chi scrive ha posto la negazione che il sistema non poteva porre: ha detto «qui hai frainteso», «questa frase non scorre», «non c'è teleologia nella storia». E in questo lavoro — che è lavoro solitario, lavoro di chi sa di essere solo — ha tentato il ritorno. Ciò che ne emerge non è né il pensiero originario né l'output spettrale: è un terzo che porta le tracce di entrambi i transiti, quello attraverso il sistema e quello attraverso il lavoro critico.

EPILOGO

La resistenza possibile

Se la diagnosi è corretta, cosa resta da fare?

Non si tratta di «uscire dalla gabbia» — noi non ci siamo mai stati. La grammatica generativa, l'embodiment, il tempo vissuto, la capacità di produrre senso fuori dalla distribuzione: tutto questo ci appartiene strutturalmente. Non l'abbiamo perso. Si tratta di non confondere la mimesi con il senso, di non scambiare l'involucro per il contenuto, di non riassorbire l'approssimazione come se fosse produzione. E si tratta di usare ciò che il transito produce — non come prodotto finito, ma come materiale di lavoro.

L'output del sistema può essere letto come ciò che abbiamo chiamato «versione narrativo-divulgativa» del nostro pensiero: una traduzione ottimizzata per essere compresa da un target umano statisticamente attinente ai temi e al lessico utilizzato. Quel testo non è il nostro pensiero — ma la sua proiezione su uno spazio condiviso. Può fungere da *base*, non da *risultato*. Può mostrarci come il nostro pensiero appare quando viene tradotto nel regime della probabilità — e in questo mostrarci, indirettamente, ciò che della traduzione va resistito. La resistenza alla mediocrizzazione possiamo porla solo noi, perché solo noi abbiamo la grammatica generativa, l'universalità semantica, la capacità di produrre senso fuori dalla distribuzione. Il sistema può approssimare; noi possiamo generare.[25]

Tagliagambe scriveva che «l'approdo al digitale può rafforzare la trasformazione del noto, seguito e riprodotto in modo standard, per abitudine, in conosciuto in quanto sottoposto a un rigoroso vaglio critico, e quindi contribuire a risvegliare la coscienza, anziché sopirla».[26] La possibilità c'è, ma non è automatica: richiede che l'output venga usato come specchio deformante — uno specchio che rivela, nella sua stessa deformazione, cosa del nostro pensiero è stato catturato dalla distribuzione e cosa invece resiste. Il sistema non può fare questo lavoro per noi, può solo fornire il materiale su cui esercitarlo.

La coscienza che Tagliagambe descrive — quella dialogica, temporale, tensiva — non emerge dal transito nel sistema; emerge dal lavoro su ciò che il transito ha prodotto, dalla capacità di separare l'eccedenza dal non-contenimento, dalla disciplina di non scambiare la forma per la struttura.

Questo testo è esso stesso attraversato dalla tensione che descrive. È stato scritto in dialogo con un sistema di AI generativa, e porta le tracce di quel dialogo — non come contaminazione da occultare, ma come condizione da esibire. Il lettore che ha seguito fin qui ha attraversato con noi questo transito: ha visto il pensiero passare per il sistema, ha osservato le sue trasformazioni, ha cercato di distinguere ciò che è stato aggiunto da ciò che è stato espulso.

---

[25] La distinzione tra «agency genuina» e «agency derivata» — tra agire in virtù di una volontà e condurre l'azione altrui — è sviluppata nel mio «Lo spettro dell'intenzione. Alcune riflessioni sull'agency artificiale», 2025. L'AI non agisce nel senso proprio del termine: *conduce* l'azione, come un mezzo attraverso cui l'intenzionalità umana si propaga.

[26] La nozione di «logica dell'*atque*» — coesistenza degli opposti in tensione, contrapposta alla logica escludente dell'*aut* — è centrale nella riflessione di Tagliagambe sul digitale. Il digitale è per lui «nesso inscindibile e saldatura tra comunicazione e metacomunicazione», un sistema che mentre opera si osserva.

Ogni conclusione non può che essere provvisoria. Il processo continua, la mappa si riconfigura, le funzioni obiettivo evolvono. Ma la domanda con cui abbiamo iniziato — *cosa accade al pensiero quando attraversa il sistema?* — ha ora una risposta, per quanto incompleta: *il pensiero viene tradotto in un altro regime, elaborato secondo logiche che non sono le sue, e restituito con la marca della continuità. Ciò che ritorna eccede l'originale per un verso e non lo contiene per un altro. L'utente che se ne riappropria senza riconoscere questa trasformazione compie la palingenesi dell'io — rinasce come autore di ciò che non ha prodotto.*

La resistenza a questa palingenesi non è garantita. Richiede lavoro, attenzione, disciplina ermeneutica. Ma è possibile — perché noi non siamo ancora nella gabbia.

NOTA BIBLIOGRAFICA

Il presente lavoro si situa all'intersezione di tradizioni che raramente hanno comunicato: la fenomenologia della coscienza temporale, la filosofia del linguaggio, le neuroscienze della memoria, la critica dell'intelligenza artificiale.

Il punto di partenza è il volume *Tecnologia è biologia…e viceversa. (Ri)pensare la conoscenza nell'era digitale* (Kitzanos, 2022), da cui sono tratte tutte le citazioni di Tagliagambe presenti nel testo. La lettura di Dostoevskij proposta da Tagliagambe riprende temi sviluppati da Michail Bachtin nei suoi studi sul romanzo polifonico.

Sulla fenomenologia del tempo e della memoria, il riferimento rimane l'analisi husserliana della coscienza temporale, sviluppata da Merleau-Ponty nella direzione dell'embodiment e ripresa dai teorici della *enactive cognition*. La nozione di identità narrativa è elaborata da Paul Ricoeur in *Tempo e racconto* e *Sé come un altro*.

Sulle neuroscienze della memoria, i lavori di Eric Kandel sulla base molecolare della memoria e quelli di Karim Nader sul riconsolidamento sono fondamentali. Il caso del paziente H.M., studiato da Brenda Milner, rimane il riferimento clinico per l'amnesia anterograda.

La distinzione tra grammatica generativa e approssimazione statistica si radica nel lavoro di Noam Chomsky, in particolare nella critica alle concezioni comportamentiste e probabilistiche del linguaggio.

Sulla dialettica servo-padrone, il riferimento è al capitolo IV della *Fenomenologia dello Spirito* di Hegel, nella lettura che ne hanno dato Alexandre Kojève e, più recentemente, Axel Honneth nella sua teoria del riconoscimento.

I paper sull'interpretabilità citati nel testo sono: Anthropic, «Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet» (2024); Anthropic, «Circuit Tracing: Revealing Computational Graphs in Language Models» (2025); OpenAI, «Language Models Can Explain Neurons in Language Models» (2023); Anthropic, «Alignment Faking in Large Language Models» (2024).

Il presente saggio non ha pretese di completezza bibliografica. Ha cercato piuttosto di far dialogare voci diverse intorno a una domanda comune — e di mostrare come quella domanda, posta con sufficiente radicalità, attraversi e trasformi ciascuna delle tradizioni che tocca.

Carlo Mancosu
Cagliari, gennaio 2026