



10.69118/2025_L1

Intelligenze artificiali e mondi immersivi: nuovi paradigmi per le Digital Humanities

Fabio Ciotti

Università di Roma Tor Vergata
fabio.ciotti@uniroma2.it

Artificial Intelligence and immersive worlds: new paradigms for Digital Humanities

Abstract (ITA)

L'articolo esplora l'impatto epistemologico e metodologico dell'introduzione di tecnologie emergenti – intelligenza artificiale generativa (IA) e tecnologie immersive (realtà aumentata, mista e virtuale) – nel contesto delle Digital Humanities. Nella prima parte si analizzano le tecnologie immersive, evidenziandone il potenziale nella creazione di esperienze multisensoriali e interattive che ridefiniscono la nozione di immersione narrativa tradizionalmente associata alla letteratura. Viene discussa la correlazione teorica tra immersione testuale e digitale, con riferimento a diversi approcci teorico-letterari. La seconda parte esamina l'IA generativa e i Large Language Models (LLM), descrivendone l'architettura tecnica, i meccanismi di funzionamento e le capacità emergenti. L'articolo affronta criticamente il dibattito teorico sulle reali capacità cognitive degli LLM, esaminando posizioni deflazioniste e possibiliste, e propone come gli LLM possano essere considerati modelli rappresentativi dei sistemi culturali attraverso il concetto lotmaniano di "semiosfera". Infine, vengono esplorate le prospettive future derivanti dalla convergenza tra tecnologie immersive e IA generativa, inclusa la produzione automatizzata di contenuti multimediali (ecfrasi inversa, generazione video, world models). L'articolo conclude sottolineando la necessità di un approccio interdisciplinare rigoroso e aperto, capace di sfruttare pienamente il potenziale innovativo di queste tecnologie nella ricerca umanistica contemporanea.

Abstract (ENG)

This article explores the epistemological and methodological implications of introducing new technologies – generative artificial intelligence (AI) and immersive technologies (augmented, mixed and virtual reality) – in the context of digital humanities. The first part analyses immersive technologies, highlighting their potential to create multisensory and interactive experiences that redefine the notion of narrative immersion traditionally associated with literature. The theoretical relationship between textual and digital immersion is discussed with reference to different approaches to literary theory. The second part examines generative

AI and Large Language Models (LLMs), describing their technical architecture, operational mechanisms and emerging capabilities. The article critically addresses the theoretical debate about the actual cognitive capabilities of LLMs, examining both deflationist and possibilist positions, and proposes how LLMs can be considered representative models of cultural systems through Lotman's concept of the 'semiosphere'. Finally, the article explores future perspectives arising from the convergence of immersive technologies and generative AI, including the automated production of multimedia content (inverse ekphrasis, video generation, world models). The article concludes by emphasising the need for a rigorous and open interdisciplinary approach capable of fully exploiting the innovative potential of these technologies in contemporary humanistic research.

Parole chiave / Keywords

Tecnologie immersive, IA generativa, Innovazione digitale / Immersive Technologies, Generative AI, Digital Innovation

Introduzione

Il vasto e multiforme campo delle Digital Humanities¹ è sempre stato soggetto alla pervasiva influenza dell'innovazione tecnologica, adattandosi alle alterne fortune delle singole tecnologie informatiche nel corso degli anni. La capacità trasformativa delle tecnologie digitali, d'altra parte, è innegabile, come lo è il loro impatto sociale e culturale, ed è naturale che chi si occupa di adottare criticamente sistemi e modelli computazionali nello studio della sfera simbolica e culturale subisca il richiamo dell'hype tecnologico di turno. In questa fase storica il ruolo dominante nei "discorsi" sul digitale è rivestito dall'Intelligenza Artificiale (IA) generativa, con particolare riferimento ai Large Language Models (LLM)², e, in misura minore, dalla rinnovata attenzione verso le tecnologie immersive (Realtà Aumentata, Mista e Virtuale), ricomprese sotto la suggestiva formula di *Metaverso* (il termine fa la sua comparsa nel romanzo di Neal Stephenson *Snow Crash*³, ma deve il suo successo al fatto di essere stato adottato come fulcro della strategia di marketing della casa madre di Facebook e altri social media globali, ribattezzata con l'occasione Meta).

L'introduzione e la rapida diffusione di queste tecnologie stanno innescando una trasformazione radicale nelle metodologie delle Digital

1 Si veda Ciotti 2023a.

2 Si veda Ciotti 2023b; Roncaglia 2023.

3 Si veda Stephenson 1992.

Humanities, che trascende in modo significativo la dimensione della semplice innovazione strumentale, per configurarsi come una vera e propria svolta paradigmatica. La convergenza e la sinergia tra i sistemi capaci di generare esperienze multisensoriali immersive e i modelli neurali avanzati dell'IA generativa promette di aprire nuove possibilità sul piano analitico ed ermeneutico, di esplorare dimensioni precedentemente inaccessibili dell'esperienza testuale, offrendo vie alternative per l'interpretazione e la comprensione dei fenomeni culturali. In questo intervento intendo riflettere su alcuni dei fondamenti e delle implicazioni teoriche e metodologiche determinate da tale convergenza tecnologica, rivolgendo una particolare attenzione alle potenziali applicazioni nell'ambito degli studi letterari.

1. Le tecnologie immersive e realtà virtuale nel contesto umanistico

Le tecnologie immersive e la realtà virtuale (VR) hanno fatto la loro comparsa agli inizi degli anni '90, anticipate da numerose prefigurazioni letterarie, in particolare quelle di William Gibson, soprattutto nei romanzi della *Trilogia dello Sprawl*.⁴ – *Neuromancer*, *Count Zero*, *Mona Lisa Overdrive* – e nei racconti connessi, e del movimento Cyberpunk.⁵ Inizialmente, suscitarono un'ondata di entusiasmo, alimentando la promessa di mondi digitali capaci di espandere i nostri sensi e permettere interazioni inedite con la realtà.⁶ Tuttavia, le significative limitazioni tecnologiche dell'epoca determinarono presto una crisi di attenzione e un lungo periodo di relativa oscurità mediatica. Solo negli ultimi anni si è assistito a una parziale inversione di tendenza, trainata dai recenti sviluppi hardware e dalla crescente disponibilità di interfacce immersive a basso costo: visori standalone a prezzi accessibili (es. Meta Quest 3, Pico 4) con prestazioni precedentemente riservate a sistemi high-end; controller aptici avanzati per interazioni naturali con oggetti virtuali; sistemi di *eye-tracking* integrati che abilitano nuove modalità di interazione. La diminuzione dei costi e il miglioramento delle prestazioni dell'hardware VR moderno contribuiscono a un'esperienza più fluida, intuitiva e coin-

4 Si veda Gibson 2019.

5 Si veda Bould 2005; Caronia-Gallo 1996.

6 Si veda Rheingold 1991.

volgente, fondamentale per il suo utilizzo efficace nelle applicazioni DH, dove l'engagement dell'utente e il senso di presenza sono aspetti importanti, ma devono essere commisurati alle risorse limitate tipiche della ricerca umanistica.

In linea generale le tecnologie digitali immersive si articolano in tre tipologie che si articolano lungo il continuum della virtualità definito da Milgram e Kishino (1994), ciascuna delle quali offre diverse possibilità di integrazione percettiva tra elementi virtuali e ambiente fisico. La *Realtà Aumentata* (AR) sovrappone elementi visivi digitali al mondo reale, integrando ed espandendo digitalmente l'ambiente fisico dell'utente in tempo reale. La sua funzionalità di base consiste nel creare connessioni tra il mondo reale e le informazioni generate elettronicamente, arricchendo la percezione umana. La *Realtà Mista* (MR) consente l'interazione in tempo reale tra oggetti virtuali e reali, permettendo la sovrapposizione e l'interazione tra elementi digitali e l'ambiente reale a vari livelli. Essa si basa su metodi di input avanzati e sulla percezione ambientale, adattandosi all'ambiente fisico dell'utente. La *Realtà Virtuale* (VR) crea ambienti digitali tridimensionali completi, offrendo il massimo grado di immersione. Essa fornisce in modo massimale una «esperienza interattiva e immersiva generata da un computer» (Pimentel e Teixeira 1993). La VR è uno spazio percettivo immersivo e interattivo generato dal computer che simula un ambiente tridimensionale dotato di parte delle proprietà fisiche dell'ambiente 'reale', cui si accede con l'ausilio di hardware specializzato come visori e controller, e offre un senso di presenza e la possibilità di interagire ed esplorare mondi virtuali realistici e reattivi.

Dal punto di vista tecnico, i sistemi immersivi si avvalgono di un ecosistema tecnologico complesso e integrato per creare l'illusione di un mondo virtuale coerente e navigabile. Gli *Head-Mounted Displays* (HMD) rappresentano l'interfaccia primaria per la VR, offrendo un completo isolamento percettivo dal mondo esterno e una visione stereoscopica e tridimensionale di quello virtuale attraverso schermi posizionati davanti agli occhi dell'utente, combinati con sistemi di tracciamento del movimento che adattano continuamente la prospettiva visuale ai movimenti della testa e altoparlanti 3D. Nelle applicazioni AR, dove non è necessario l'isolamento, sono molto più diffusi dispositivi come gli *smart glass*, ma anche i comuni smartphone e tablet possono essere funzionali in alcuni contesti. Per la percezione tattile si usano i *data glove*, che consentono di tradurre i movimenti delle mani in azioni all'interno dell'ambiente virtuale, permettendo manipolazioni e interazioni con oggetti si-

mulati, ma nei sistemi low end essi sono sostituiti da controller wireless simili a quelli delle piattaforme di gioco. Questi apparati hardware di base possono essere completati da sensori di movimento per tracciare la posizione dell'utente nello spazio, sistemi audio spazializzati per creare un ambiente sonoro immersivo. Non vanno poi dimenticati i sistemi hardware e software per il *rendering* in tempo reale delle immagini in risposta ai movimenti e alle azioni dell'utente. A questi strumenti si aggiungono i framework di modellazione tridimensionale e simulazione comportamentale che costituiscono l'infrastruttura su cui si costruiscono gli ambienti virtuali. L'esperienza VR è dunque il risultato di una complessa orchestrazione di tecnologie integrate che collaborano per ingannare i sensi e creare un'esperienza percettiva convincente.

Come detto, i concetti chiave che permettono di inquadrare teoricamente l'esperienza VR sono *immersione* e *interattività*. L'immersione si riferisce alla sensazione di presenza all'interno di un mondo fittizio, percepito come reale e autonomo grazie alla simulazione computazionale. Questo mondo virtuale, popolato da oggetti e personaggi specifici, crea uno spazio con cui l'utente può relazionarsi in modo diretto e intuitivo. L'interattività definisce invece la capacità dell'utente di agire all'interno dell'ambiente virtuale, modificando l'esperienza attraverso azioni quali l'esplorazione, la manipolazione e la trasformazione dell'ambiente stesso.

Questi due principi, apparentemente tecnici, hanno profonde implicazioni teoriche e hanno assunto un ruolo significativo anche nei dibattiti della teoria letteraria degli ultimi decenni. La correlazione concettuale tra l'esperienza dell'immersione nel testo narrativo durante l'atto della lettura e quella fornita dai sistemi immersivi digitali appare evidente quando consideriamo modelli e riflessioni teoriche quali la semantica a mondi possibili applicata ai testi finzionali elaborata da teorici come Pavel (1986), Eco (1979), Doležel (1999), Ryan (1991), che concepisce i testi narrativi come dispositivi costruttori di mondi alternativi dotati di proprietà spaziali e temporali specifiche; la teoria della finzione come *make-believe* proposta da Walton (1993), dove i fatti finzionali sono "veri nel gioco appropriato di make-believe", ossia veri nel mondo finzionale rappresentato; la nozione bachtiniana di cronotopo,⁷ intesa come matrice spazio-temporale che struttura la narrazione letteraria. Parallelamente, l'interattività digitale rappresenta un'evoluzione delle teorie sulla costru-

7 Si veda Bachtin 1997.

zione cooperativa del significato tra testo e lettore, manifestando materialmente le intuizioni teoretiche postmoderne e trasformando il testo da elemento statico a costruzione dinamica, come sostenuto dalle teorie sull'ipertesto di George Landow, Stuart Moulthrop, e altri studiosi negli anni iniziali della rivoluzione digitale.⁸

In particolare Marie-Laure Ryan ha sviluppato un'analisi sistematica dei rapporti tra la nozione di realtà virtuale e la narratività nel suo influente volume *Narrative as Virtual Reality* (Ryan 2003; edizione riveduta 2015), dove ha esaminato i parallelismi tra l'esperienza immersiva nei mondi virtuali, l'esperienza del lettore nei mondi narrativi e la stessa evoluzione dei sistemi letterari. La teorica statunitense, infatti, osserva come la storia del romanzo moderno offra una prospettiva illuminante sulle dinamiche culturali della nozione di immersione, evidenziando un'alternanza tra fasi dominate da poetiche dell'immersione e fasi orientate al distanziamento critico e alla meta-riflessione. Il romanzo borghese del XVIII secolo mostrava una caratteristica ambivalenza, utilizzando tecniche illusionistiche per creare mondi finzionali convincenti, ma richiamando contemporaneamente l'attenzione sul processo narrativo stesso attraverso interventi autoriali e strategie metaletterarie. Il romanzo realista del XIX secolo ha privilegiato l'immersione, caratterizzandosi per un narratore progressivamente meno visibile e una focalizzazione intensificata sulle esperienze emotive dei personaggi, creando l'illusione di un accesso diretto e non mediato alla realtà rappresentata. La letteratura modernista del XX secolo ha invece compiuto un deliberato allontanamento dalle strategie immersive, privilegiando approcci metanarrativi, sperimentali e autoriflessivi che problematizzavano la relazione tra testo, realtà e lettore.

Sul piano più strettamente teorico Ryan afferma che l'immersione narrativa si articola in multiple dimensioni interconnesse che operano simultaneamente nell'esperienza del lettore. L'immersione spaziale crea un senso di presenza nell'ambiente narrativo, consentendo l'esplorazione mentale degli spazi testuali descritti e la costruzione di una geografia immaginaria in cui collocare l'azione narrativa. L'immersione temporale genera coinvolgimento nella progressione degli eventi, creando tensione narrativa e anticipazione del futuro sviluppo della trama. L'immersione emotiva facilita l'identificazione con i personaggi e la partecipazione em-

8 Si veda Landow 1992; Landow 1994.

patica alle loro esperienze, permettendo al lettore di sperimentare vicariamente stati emotivi complessi. Queste modalità non sono semplicemente additive, ma si integrano in un'esperienza ermeneutica complessa in cui ciascuna dimensione rafforza e modifica le altre. Le narrazioni più efficaci sono quelle che riescono a bilanciare questi diversi tipi di immersione, offrendo un'esperienza cognitivamente ed emotivamente ricca al fruitore.

L'interattività caratteristica dei media digitali, d'altro canto, può essere interpretata come un'evoluzione e materializzazione della concezione postmoderna del significato come processo costruttivo: si è passati da una letteratura orientata a creare un mondo immersivo per il lettore, a una letteratura che sollecita il lettore a diventare un partecipante attivo nella costruzione del significato, fino all'avvento dell'ipertesto e delle narrative interattive come forme di testualità che incorporano strutturalmente questa dimensione partecipativa. D'altra parte, è stato più volte osservato come l'interattività sia connessa a una concezione ludica della testualità, in cui il testo non è più visto soltanto come la rappresentazione di un "mondo" in cui immergersi passivamente, ma come un "gioco" con cui interagire attivamente. In questa prospettiva, l'utente assume il ruolo di un giocatore che esplora, manipola e modifica il testo, partecipando alla costruzione del suo significato. I testi digitali che privilegiano la dimensione ludica rispetto a quella mimetica adottano talvolta una disfunzionalità deliberata, mettendo in discussione la funzione rappresentativa del linguaggio e sottolineando invece la sua natura di sistema di regole manipolabili. L'interattività trasforma così il testo in uno spazio di gioco e sperimentazione, stimolando il fruitore a diventare parte attiva del processo di significazione, superando il ruolo tradizionale di osservatore passivo e distaccato.

La Realtà Virtuale, alla luce di queste considerazioni, trascende la sua natura puramente tecnologica per configurarsi come un dispositivo epistemologico con profonde implicazioni per la teoria letteraria e culturale. Essa permette di materializzare le strutture narrative in spazi navigabili, rendere tangibili le relazioni intertestuali attraverso connessioni spaziali, e visualizzare le stratificazioni semantiche del testo in forma di livelli percettivi sovrapposti. Questa trasformazione determina una riconfigurazione profonda delle modalità attraverso cui il significato testuale viene costruito e interpretato, offrendo nuove possibilità per l'analisi e la comprensione dei fenomeni letterari. Ma anche più rilevante è il valore didattico e di public engagement che essa può assumere, rendendo accessibili testi complessi a pubblici diversificati e potenziando le capacità

didattiche attraverso l'esperienza diretta e l'*embodiment*. Un caso emblematico è rappresentato dai numerosi progetti che utilizzano la VR per ricostruire il mondo della *Divina Commedia*, permettendo agli utenti di esplorare fisicamente gli spazi fenzionali descritti da Dante e di materializzare le complesse geografie simboliche del poema rendendo immediata la percezione della sua struttura spaziale; una "tradizione" di cui il progetto D.A.N.T.E. (*Digital Archive and New Technologies for E-content*, <https://dante.dantelimina.it>), coordinato da Ciro Perna e Elisabetta Tonello, rappresenta uno tra i più stimolanti e fondati esempi.

Resta certamente aperta una domanda critica sul valore epistemico di tali applicazioni nel contesto degli studi e non della divulgazione (posto che oggi abbia senso porre tale netta distinzione): la questione se e in che misura queste esperienze immersive apportino conoscenze genuinamente innovative sui testi letterari, o se rappresentino principalmente riformulazioni di comprensioni già acquisite attraverso metodi interpretativi tradizionali. Se da un lato i mondi virtuali possono enfatizzare alcune dimensioni del testo (come la spazialità, il movimento e la materialità degli ambienti descritti), dall'altro rischiano di ridurne la complessità semantica, trasformando in "scene visive" concrete ciò che è primariamente di natura linguistica e concettuale, con tutte le ambiguità e polisemie che questa natura comporta. La visualizzazione, per quanto sofisticata, comporta sempre una selezione e una disambiguazione che può impoverire la ricchezza semantica del testo letterario. Ma a contrastare questo atteggiamento vagamente apocalittico, si può rilevare, da una parte, come la percezione *embodied* e immersiva permetta di evidenziare aspetti spaziali e visuali che lasciano tracce rade nel testo linguistico ma giocano un ruolo importante nei processi cognitivi di creazione e fruizione letteraria; e, dall'altra, come la storia della cultura, e della letteratura al suo interno, sia stata anche una storia di successive rimediazioni e traduzioni semiotiche⁹ in cui questi esperimenti si inseriscono in perfetta continuità.

2. Intelligenza Artificiale generativa e Large Language Models

Passiamo ora a considerare la seconda innovazione tecnologica che sta caratterizzando questi anni recenti: l'Intelligenza Artificiale generativa. Con questa formula ci si riferisce a una classe di sistemi computazio-

9 Si veda Bolter *et al.* 2003.

nali basati sull'impiego di reti neurali artificiali¹⁰ e metodi avanzati di machine learning,¹¹ caratterizzati dalla capacità di produrre contenuti originali a partire dall'analisi di vastissimi corpora con cui sono addestrati. Questa categoria di sistemi include i grandi modelli del linguaggio (Large Language Model, o LLM) come GPT, Claude, Gemini, DeepSeek e LLaMA che hanno rivoluzionato l'elaborazione del linguaggio naturale negli ultimi anni, e sono balzati agli onori delle cronache e al centro del dibattito scientifico. A questi si affiancano i modelli generativi per immagini come DALL-E, Midjourney e Stable Diffusion, che hanno esteso il paradigma generativo al dominio visuale, e ancora più recenti quelli per video come Sora.

Dal punto di vista tecnico, questi sistemi si basano su architetture di reti neurali di grandi dimensioni, addestrate in modo non supervisionato o *self-supervised* su quantità enormi di dati, prevalentemente provenienti dal Web. La loro caratteristica distintiva rispetto ai sistemi di reti neurali artificiali precedenti, essenzialmente dei classificatori, è che essi sono in grado di “generare” contenuti completamente nuovi. Un LLM generativo, infatti, produce testo in risposta a un input testuale (*prompt*), e non si limita (anzi ne è in genere incapace) a replicare o ricombinare gli esempi presenti nei dati di addestramento, bensì sfrutta la competenza linguistica e testuale derivata dall'apprendimento profondo per produrre output originali, spesso con risultati che sorprendono per qualità e coerenza semantica.

A livello astratto, un LLM può essere concettualizzato come un sistema capace di predire, su base probabilistica, la sequenza di token linguistici più appropriata a partire da una sequenza data. Più precisamente, il modello stima la distribuzione di probabilità per tutti gli elementi (detti token) del suo vocabolario, valutando la probabilità che ciascun token possa essere il successore immediato di una sequenza di input $S=t_1, \dots, t_k$. Poi il modello seleziona tra quelli più probabili (in base a metodi stocastici) un singolo token t_{k+1} , generando una nuova sequenza $S_1=S, t_{k+1}$ e reintroduce S_1 come input, in un processo autoregressivo che continua fino al raggiungimento di un punto di terminazione.

Alla base di questo processo computazionale vi è il concetto di *word embedding*, una tecnica che, analizzando un vasto numero di contesti

10 Si veda Mitchell 2022.

11 Si veda Alpaydin 2010.

d'uso, estrapola le proprietà sintattiche e semantiche di una parola, traducendola in un vettore di numeri reali all'interno di uno spazio multidimensionale. Il *word embedding* operationalizza le tesi della linguistica distribuzionale, sintetizzate nell'aforisma del linguista Firth: «You shall know a word by the company it keeps».¹² In questo paradigma, il significato di una parola è determinato dalla totalità dei contesti in cui essa appare, una intuizione che l'adozione di reti neurali ha tradotto in rappresentazioni della semantica lessicale sottoforma di vettori densi di numeri reali. Ma la semantica lessicale non sarebbe sufficiente a sviluppare modelli capaci di comprendere e generare interazioni linguistiche lunghe e complesse. Un aspetto cruciale nel funzionamento degli LLM basati sull'architettura a *transformer* è rappresentato dal meccanismo dell'*attenzione*, una tecnica matematica piuttosto sofisticata utilizzata per determinare il significato di una parola basandosi sul suo contesto linguistico attuale.¹³ In una rete neurale dotata di meccanismo di attenzione, ogni token di input viene elaborato indipendentemente, e il modello utilizza l'attenzione per identificare quali token del contesto sono più rilevanti per la comprensione del significato del token corrente. L'attenzione assegna pesi a ciascun token di input, indicandone l'importanza relativa nel determinare il significato del token in esame. Il risultato è una rappresentazione vettoriale del token che incorpora l'informazione contestuale, utilizzata poi dal modello per generare l'output appropriato.

La generazione dell'output testuale, infine, avviene attraverso un processo di decodifica multifase. Durante questo processo, il modello valuta ogni parola del suo vocabolario in base alla probabilità che essa segua in modo coerente la sequenza di parole esistente nel prompt. Per calcolare questa probabilità, l'LLM moltiplica il vettore di output dei transformer per la matrice delle parole nel suo vocabolario e applica un algoritmo *softmax* al vettore risultante per ottenere la distribuzione di probabilità normalizzata. La scelta finale della parola da aggiungere alla sequenza di input non è deterministica (*greedy decoding*), ma avviene selezionandone una in modo pseudo-casuale a partire da un pool di candidate che superano un certo valore di probabilità. Questo comportamento stocastico è regolato da un iper-parametro denominato "temperatura", che modula il grado di "libertà" e "creatività" linguistica del modello: tem-

12 Firth 1957, p. 179.

13 Si veda Vaswani *et al.* 2017.

perature più alte aumentano la diversità e l'imprevedibilità dell'output, mentre temperature più basse rendono le generazioni più deterministiche e conservative.

Le "capacità" linguistiche degli LLM derivano da un complesso processo di apprendimento articolato in più fasi. Si parte da un massiccio addestramento "auto-supervisionato" di base, in cui il modello apprende a predire la parola finale, opportunamente mascherata, di miliardi di frasi di testo estratte da enormi corpora linguistici. Questo stadio è seguito da fasi successive di "apprendimento supervisionato" su lunghe serie di coppie domanda/risposta prodotte da esperti umani (*Supervised Fine Tuning*), e di "apprendimento per rinforzo" (*Reinforcement Learning from Human Feedback*, RLHF), che costituiscono il *fine-tuning* del modello base. Questa fase è particolarmente importante per allineare il comportamento del modello con le preferenze culturali e i valori preferiti da chi sviluppa il modello, ottimizzando (ma anche censurando) la qualità, utilità e sicurezza delle sue generazioni.

Nonostante siano fondati su principi matematici ben definiti, gli LLM hanno manifestato proprietà emergenti non direttamente riconducibili a tali principi fondativi. Il complesso processo di addestramento su architetture neurali sufficientemente ampie sembra portare all'acquisizione di capacità non esplicitamente previste dall'obiettivo originale di predizione accurata del token successivo in una sequenza linguistica. Tra queste, particolarmente notevole è la capacità di apprendimento contestuale (*in-context learning*) ovvero la capacità di eseguire compiti per cui il modello non è stato specificamente addestrato,¹⁴ solo sulla base di semplici richieste (*zero-shot*), o al massimo di richieste corredate da pochi esempi (*few-shot*); gli LLM, insomma, dimostrano una apparente capacità di generalizzazione del tutto inattesa, anche se ci sono numerosissimi lavori teorici e sperimentali che argomentano a favore o contro queste presunte facoltà emergenti dei modelli (un dibattito che non è possibile approfondire in questa sede).

Gli LLM esibiscono altre notevoli meta-abilità che trascendono le semplici capacità linguistiche. Una capacità significativa è la 'sensibilità ai prompt', ovvero la capacità di essere influenzati dalle richieste specifiche degli utenti, dai suggerimenti impliciti e dalle implicazioni conversazionali. Ad esempio, strategie di *prompting* che chiedono al modello di

14 Si veda Brown *et al.* 2020.

‘agire come’ o ‘impersonare’ un ruolo specifico possono elicitare risposte radicalmente diverse. Analogamente, richiedere al modello di “pensare passo dopo passo” attraverso tecniche di *Chain-of-Thought prompting* può attivare processi di ragionamento strutturato per compiti complessi, migliorando significativamente le performance in ambiti che richiedono deduzioni multiple. Sfruttando questa ultima capacità sono state sviluppate le più recenti evoluzioni di gran parte dei modelli rilasciati negli ultimi mesi sotto classificati come *Large Reasoning (Language) Models*.¹⁵

Questi fenomeni suggeriscono che gli LLM non sono semplici macchine generatrici di testo basate su correlazioni statistiche; durante l’interazione, essi sembrano impegnarsi, in una certa misura, in un dialogo governato da regole semantiche e pragmatiche complesse, mostrando una sensibilità al contesto e una flessibilità adattiva che richiamano, pur con significative differenze, alcuni aspetti della comunicazione linguistica umana.

3. Il dibattito sulla natura degli LLM

Sin dalle prime evidenze sperimentali delle sorprendenti capacità degli LLM (anche di generazioni precedenti come GPT-3, LaMDA, PaLM), si è acceso un intenso dibattito sulla loro natura e sulle loro reali capacità cognitive. Questo confronto, che vede coinvolti filosofi, informatici, linguisti, psicologi e anche letterati, come chi scrive, verte su questioni fondamentali che trascendono gli aspetti puramente ingegneristici di questi sistemi.

Un primo interrogativo concerne la competenza semantica degli LLM: questi modelli possiedono una vera comprensione del linguaggio, o si limitano a manipolare simboli senza accesso ai loro significati? Questa domanda riecheggia l’esperienza mentale della ‘stanza cinese’ proposto da John Searle negli anni ‘80 (Searle 1980), che metteva in discussione la possibilità che un sistema puramente sintattico potesse sviluppare una genuina comprensione semantica. Un secondo ordine di questioni riguarda la possibilità che gli LLM possiedano stati mentali analoghi a quelli umani: hanno credenze (convinzioni sulla verità o falsità delle proposizioni), desideri (stati motivazionali verso obiettivi) e intenzioni (impegni verso piani d’azione)? Collegata a questo punto è la questione dell’intenzionalità, intesa nella tradizione filosofica come la capacità di avere

¹⁵ Si veda Raschka 2023; Xu *et al.* 2025.

stati mentali “diretti verso” o “a proposito di” qualcosa, una proprietà che tradizionalmente distingue la mente umana dai sistemi puramente meccanici. Infine, si discute se gli LLM possiedano una forma di *agency*, ovvero la capacità di agire in modo autonomo e diretto a scopi, piuttosto che essere semplici strumenti passivi guidati dalle istruzioni umane. Le posizioni in questo dibattito complesso e multidisciplinare possono essere schematicamente divise tra approcci deflazionisti, che tendono a minimizzare le capacità cognitive degli LLM interpretandoli come sistemi di manipolazione simbolica privi di genuina comprensione, e approcci possibilisti o ottimisti, che riconoscono in questi modelli forme emergenti di cognizione che, pur diverse da quelle umane, presentano caratteristiche di genuina comprensione semantica e capacità rappresentazionali.

L'articolo più influente della corrente deflazionista è probabilmente *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* di Bender, Gebru, McMillan-Major e Mitchell (2021). Questo *position paper* articola diverse critiche agli LLM, spaziando da considerazioni teoriche a preoccupazioni etiche, politiche e ambientali. Gli aspetti più propriamente teorici riprendono un precedente lavoro di Bender e Koller (2020), e la tesi centrale sostenuta da questi autori è che il testo generato da un LLM non è fondato su un'autentica intenzione comunicativa, né su un modello del mondo o una rappresentazione dello stato mentale dell'interlocutore. Questa impossibilità deriva dal fatto che i dati di addestramento non includono esperienze di condivisione di pensieri con un ascoltatore in un contesto di comunicazione genuina, né l'architettura del modello possiede i requisiti strutturali per sviluppare tali capacità. Secondo questa visione, sebbene il testo generato appaia sempre più fluente e coerente, la percezione di significato che ne deriva è interamente mediata dalla competenza linguistica umana e dalla nostra predisposizione cognitiva a interpretare atti comunicativi come portatori di significato e intenzione coerenti, indipendentemente dalla loro origine. Se una delle parti della comunicazione (il modello) non possiede significato intrinseco, la comprensione di un significato implicito diventa un'illusione derivante dalla nostra specifica comprensione umana del linguaggio e dalla nostra tendenza a proiettare intenzionalità anche dove questa è assente. In questa prospettiva critica, un LLM è essenzialmente un sistema che “cuce insieme in modo casuale sequenze di forme linguistiche” osservate nei dati di addestramento, basandosi su informazioni probabilistiche sulla loro combinazione, ma senza alcun riferimento al significato: un ‘pappagallo stocastico’ che imita senza comprendere.

A questa posizione deflazionista (oltre che politicamente allarmista) si contrappongono diverse visioni alternative che, pur riconoscendo le limitazioni degli attuali LLM, offrono interpretazioni più sfumate e complesse del loro funzionamento e delle loro potenzialità, almeno in prospettiva.¹⁶ Anche questo dibattito esula dagli scopi e dai limiti di questo articolo, ma riassumendo in modo sintetico la mia posizione, vorrei in primo luogo osservare come la strategia argomentativa alla base dell'argomento del 'pappagallo stocastico' soffra di un eccessivo ricorso a quello che chiamerei l'operatore linguistico 'VERO', di solito inserito in questo schema di proposizione: sia x una facoltà cognitiva umana a piacere; se le performance di un sistema artificiale A su x sono paragonabili (in un qualche senso operationalizzabile) a quelle di un sistema naturale N , allora A è solo in grado di mimare o simulare la "VERA" competenza x di N . Si noti che di solito l'argomentazione non si fornisce alcuna seria spiegazione scientifica su che cosa sia questa "VERA" x .

Si tratta con ogni evidenza di una fallacia del bersaglio mobile (*moving the goalposts fallacy*) – alzare continuamente lo standard di una definizione o di una capacità quando questa viene raggiunta o soddisfatta dal sistema che si sta analizzando, senza mai accettare che l'obiettivo iniziale sia stato davvero raggiunto – fondata su una forma di dualismo cartesiano implicito. Dal mio punto di vista, ritengo che un inquadramento teorico più articolato e intellettualmente produttivo ci venga fornito dall'elaborazione filosofica di Daniel Dennett: considerare gli LLM come 'sistemi intenzionali', ovvero sistemi che possono essere interpretati adottando l'*atteggiamento intenzionale*,¹⁷ la strategia di interpretare il comportamento di un'entità trattandola come se fosse un agente razionale che governa la sua scelta di azione mediante una considerazione delle sue 'credenze' e dei suoi 'desideri'. Non intendo ovviamente sostenere che i modelli linguistici attuali siano effettivamente e intrinsecamente dotati di tutte le proprietà che vorremmo attribuire alla mente umana (facoltà di linguaggio, raziocinio, modello del mondo, teoria della mente, coscienza, empatia...), ma che non c'è nulla di misterioso e irriducibile o nascosto in una insondabile interiorità che permetta di spiegare tali proprietà: tutto ciò che abbiamo sono i comportamenti del sistema, e se l'esame di tali

16 Si veda Chalmers 2023; Lederman-Mahowald 2024; Piantadosi 2024; Millière-Buckner 2024.

17 Si vedano Dennett 1997 e Dennet 1989.

comportamenti ci induce ragionevolmente ad attribuire una di tali facoltà a un agente artificiale, ciò è quanto basta per tale ascrizione. Come dice lo stesso Dennett in un passo della sua recente autobiografia,

la scienza è di per sé una specie di behaviorismo: e una volta che, dato un fenomeno, tutti i comportamenti hanno trovato una spiegazione plausibile, quelli interiori come quelli esterni, quelli macro come quelli micro, non rimane più nulla da spiegare – se non forse perché certa gente si trovasse a disagio per quella spiegazione.¹⁸

4. Gli LLM come modelli dei sistemi culturali

La prospettiva che considera i nuovi sistemi di IA generativa come agenti cognitivi individuali (e che dunque cerca di valutare se essi siano dotati o meno delle proprietà cognitive dei sistemi intelligenti naturali), non è la sola adottabile per cercare di comprendere la loro natura. Anzi, nell'ottica di questo lavoro, che ha lo scopo di indagare il loro impatto nelle Digital Humanities, potrebbe essere più utile e produttivo indagare come gli LLM si relazionino con l'insieme della conoscenza culturale multimodale con cui vengono addestrati. In questa direzione, pur rimanendo fortemente ancorata alla fazione deflazionista, si muove la psicologa dello sviluppo Alison Gopnik. Nel suo articolo *Transmission Versus Truth, Imitation Versus Innovation: What Children Can Do That Large Language and Language-and-Vision Models Cannot (Yet)*,¹⁹ propone di considerare gli LLM essenzialmente come tecnologie culturali, paragonabili, seppur con modalità operative diverse, ai sistemi tradizionali di gestione e disseminazione della conoscenza come le biblioteche, la stampa, e i media etc. Secondo questa visione, gli LLM operano campionando dal patrimonio culturale umano, pur senza possedere una vera capacità di comprensione dei contenuti, ma a differenza degli apparati medialti del passato, non si limitano alla replicazione *verbatim* dei contenuti che memorizzano, ma hanno la capacità di ricombinarli in modi nuovi e pertinenti al contesto. Gopnik sottolinea la distinzione fondamentale tra imitazione e innovazione: i modelli di intelligenza artificiale eccellono nell'imitazione, intesa come capacità di estrarre e replicare pattern esistenti nei dati di addestramento, ma mostrano significative limitazioni nell'innovazione genuina,

18 Dennett 2024, p. 201.

19 Yiu-Kosoy-Gopnik 2023.

che richiede una comprensione profonda e la capacità di interagire creativamente con il mondo fisico e sociale. Secondo questa visione, gli LLM non possiedono la capacità di sviluppare autonomamente nuove idee, non producono nuovi significati e il significato delle loro espressioni è totalmente derivativo, dipendente dai significati presenti nei testi su cui sono stati addestrati; ma proprio per questo si rivelano come dei formidabili strumenti di osservazione e analisi degli universi culturali. Una posizione sostenuta anche da Ted Underwood che osserva come

the immediate value of these models is often not to mimic individual language understanding, but to represent specific cultural practices (like styles or expository templates) so they can be studied and creatively remixed [...] Models of culture are exactly what we need.²⁰

Penso che un utile quadro di riferimento teorico in cui collocare questa modalità di considerare i modelli di IA generativa si possa rintracciare nella teoria semiotica della cultura di Jurij Lotman e della scuola di Tartu.²¹ Il concetto chiave del pensiero lotmaniano a questo riguardo è quello di *semiosfera* (1992): lo spazio semiotico globale all'interno del quale i processi di significazione (semiosi) possono avvenire e la cultura può esistere. Lotman traccia un'analogia potente con la biosfera: come quest'ultima è la condizione necessaria per la vita biologica, così la semiosfera costituisce l'ambiente fondamentale senza il quale nessun segno, testo o sistema di significazione potrebbe emergere o avere senso. Non si tratta quindi di un semplice archivio di segni o testi, ma dell'insieme totale delle loro relazioni e del contesto spaziale e funzionale che li rende possibili e intelligibili. La semiosfera è concepita come uno spazio delimitato, con dei confini che separano il suo 'interno' – lo spazio della cultura specifica, dove i segni sono organizzati e comprensibili – dall' 'esterno', che può essere percepito come non-semiotico, caotico, o appartenente a un'altra semiosfera. Tuttavia, questi confini non sono affatto barriere passive o impermeabili; al contrario, sono zone di intensa attività semiotica. È proprio sulla frontiera che avvengono i processi cruciali di traduzione, filtro, scambio e adattamento tra sistemi diversi, rendendola un'area fondamentale per la dinamica e l'innovazione culturale. Anche internamente, la semiosfera non è uniforme; la sua caratteristica princi-

20 Underwood 2021.

21 Si veda Lotman-Uspenskij 2001.

pale è l'eterogeneità. Al suo interno coesistono e interagiscono una molteplicità di linguaggi, codici e sottosistemi semiotici differenti. Questa struttura interna è anche asimmetrica, organizzata tipicamente attorno a un 'centro' – che rappresenta le strutture dominanti, più stabili e normative della cultura – e una 'periferia', più dinamica, instabile, aperta al cambiamento e al contatto con l'esterno e con elementi marginali interni. Il dialogo e la tensione tra centro e periferia, e tra i diversi sistemi che la compongono, sono motori essenziali dello sviluppo culturale.

La pluralità dei linguaggi, l'intertestualità, la memoria culturale, i processi di traduzione del significato tra sistemi semiotici diversi, l'eterogeneità e perfino le "esplosioni culturali" (momenti di imprevedibile innovazione semiotica) descritti da Lotman trovano sorprendenti corrispondenze nel funzionamento e nei comportamenti dei modelli linguistici: un LLM può essere considerato come un sistema che costruisce ed elabora un modello della semiosfera culturale sottoforma di spazio vettoriale multidimensionale, in cui le relazioni tra i segni e concetti sono rappresentate isomorficamente come proprietà algebriche e geometriche dello spazio vettoriale. Questi modelli, infatti, costruiscono uno spazio semantico dinamico e pluralistico che può essere interpretato come una «superposizione di prospettive culturali»,²² in cui i significati non sono fissati a priori, ma vengono generati, trasformati e connessi durante la produzione di atti linguistici in risposta a specifici prompt e contesti.

È particolarmente significativo notare che Lotman stesso, in un articolo del 1979 intitolato *Culture as Collective Intellect and Problems of Artificial Intelligence*, aveva prefigurato la possibilità di concepire la semiosfera come un intelletto collettivo, suggerendo che questo modello di intelligenza distribuita e collettiva potesse rappresentare un paradigma più esplicito e comprensibile dell'intelletto individuale per lo sviluppo dell'intelligenza artificiale. Infatti, le strutture e i processi dell'intelletto collettivo sono manifestate nel linguaggio della cultura e registrate nei testi, mentre i processi del pensiero individuale rimangono in larga parte inaccessibili all'osservazione diretta:

We must emphasize that collective intellect, as a model for artificial intellect, has several advantages over individual intellect. Since collective intellect is a mechanism created by the history of mankind, it is far more explicit, its procedures are manifest in the language

22 Si veda Kovač *et al.* 2023.

of culture and are recorded in numerous texts, unlike the hidden languages of the human brain.²³

Questa intuizione lotmaniana appare sorprendentemente profetica alla luce degli sviluppi recenti degli LLM, che possono essere interpretati come implementazioni computazionali, per quanto parziali e imperfette, di questa idea di intelletto collettivo materializzato nel linguaggio. E da questo deriva la loro innovativa potenzialità epistemica nel dominio dell'analisi dell'ecosistema culturale. Infatti, gli oggetti e i processi che popolano la semiosfera sono tipicamente "resistenti" all'approccio formalista tradizionale nei metodi computazionali; operazionalizzare e formalizzare (o assiomatizzare) i fenomeni culturali in modo rigoroso si è dimostrato estremamente difficile, se non impossibile, all'interno dei paradigmi computazionali classici. Il paradigma di costruzione della conoscenza sottostante ai sistemi di IA generativa, basato sull'apprendimento distribuito e sull'emergenza di capacità di alto livello da processi computazionali di basso livello, si rivela potenzialmente assai più adatto a trattare questa sfera della realtà rispetto al paradigma formalista basato su regole esplicite e rappresentazioni simboliche. Gli LLM, modellizzando interi sistemi semiotico-culturali attraverso la rappresentazione vettoriale dei token linguistici e delle loro relazioni contestuali, permettono di esternalizzare i processi interpretativi degli oggetti semiotici, distaccandoli dalla dimensione soggettiva della mente ermeneutica umana a cui erano tradizionalmente legati. Questo spostamento dell'attività interpretativa dall'interiorità del soggetto umano all'esteriorità di un sistema computazionale osservabile apre nuove possibilità per lo studio empirico dei processi interpretativi stessi, trasformando l'ermeneutica da pratica puramente riflessiva a oggetto di indagine sperimentale.

5. Le frontiere: convergenze tra IA e metaverso

Abbiamo finora trattato separatamente le due linee di innovazione che ci offrono le tecnologie digitali e le loro conseguenze epistemologiche nello studio dei fenomeni culturali e letterari. Ma già oggi si intravedono le potenzialità che possono essere offerte dalla convergenza tra tecnologie dell'immagine, sistemi immersivi e intelligenza artificiale generativa. Potenzialità che si collocano all'interno della dimensione ludica della frui-

23 Lotman 1979, p. 84.

zione testuale, cui già abbiamo fatto riferimento prima, ma che non per questo sono prive di contenuto conoscitivo, come già aveva ampiamente mostrato, da filologo, Jerome McGann sin dagli anni 90, agli albori dell'era del Web, con i suoi numerosi esperimenti di deformazione o *gamification* dei testi letterari presentati poi nel suo libro *Radiant Textuality*.²⁴

Basti pensare alla possibilità di esplorare visualmente i mondi e gli spazi dei testi mediante processi di ecfasi inversa prodotti da modelli generativi multimodali (o *text-to-image*): il gioco esplorativo, in questo caso, diventa strumento operativo di analisi testuale, permettendo di visualizzare interpretazioni multiple dello stesso passaggio descrittivo e rivelando così la pluralità delle letture possibili inscritta nel testo stesso. In modo reciproco, queste metodologie consentono di indagare le diversità delle semiosfere simulate da diversi modelli, di analizzare le divergenze tra le rappresentazioni visive da essi generate a partire dallo stesso input testuale, evidenziando come differenti architetture AI ‘leggano’ e interpretino il testo in modi distinti, rivelando presupposti impliciti, *bias* interpretativi e differenze nelle priorità semantiche dei vari modelli. Ancora più significativamente, l’ecfasi inversa computazionale permette di studiare sistematicamente la relazione tra descrizione verbale e rappresentazione visiva, aprendo nuove prospettive sulla traduzione intersemiotica, ovvero il passaggio da un sistema di segni (linguistico) a un altro (visivo). La generazione multipla di immagini a partire da un medesimo passaggio testuale crea uno “spazio interpretativo visivo” che può essere analizzato per comprendere le ambiguità semantiche del testo originale, gli elementi descrittivi dominanti e recessivi, e le inferenze culturali e contestuali applicate dai modelli generativi. Ciò consente di evidenziare quali elementi testuali vengono privilegiati o trascurati nella trasposizione visiva, rivelando così le gerarchie implicite nella percezione e interpretazione del testo.

Un’evoluzione significativa di questo approccio è rappresentata dal passaggio dalla generazione di immagini statiche alla produzione di sequenze video a partire da descrizioni testuali. L’introduzione di modelli come Sora di OpenAI (<https://openai.com/sora>) rappresenta un salto paradigmatico in questa direzione, aprendo nuove frontiere per lo studio del testo narrativo. Il video generativo introduce la temporalità come dimensione interpretativa fondamentale, permettendo di visualizzare il ritmo narrativo, le transizioni temporali e le rappresentazioni dinamiche

24 Si veda McGann 2001.

di focalizzazione e punto di vista che nel testo rimangono implicite o sono comunicate attraverso complessi artifici retorici. Una difficoltà particolare per i modelli generativi video consiste nella necessità di mantenere una continuità visiva e narrativa coerente, gestendo implicitamente elementi come causalità e consequenzialità che sono essenziali nella narrazione. Questa necessità li costringe a sviluppare forme di “comprensione” narrativa le cui potenziali applicazioni analitiche sono numerose: lo studio di sequenze descrittive in movimento; l’analisi della costruzione progressiva dello spazio narrativo nel tempo; l’esplorazione delle modalità di transizione tra scene e prospettive narrative diverse.

I *world models* rappresentano infine il paradigma più ambizioso nell’evoluzione dell’IA generativa, transcendendo la semplice generazione di contenuti per mirare alla comprensione e modellazione di interi “mondi” con le loro regole, relazioni e dinamiche interne.²⁵ Un world model genuino dovrebbe possedere una comprensione profonda delle proprietà fisiche e causali del mondo rappresentato: un modello con una rappresentazione di base di un ‘mondo’ (ad esempio, un video di una stanza sporca), dato un obiettivo (una stanza pulita), dovrebbe essere in grado di ideare una sequenza di azioni per raggiungerlo (usare l’aspirapolvere, lavare i piatti, svuotare il cestino) non perché ha osservato specificamente quel pattern nei dati di addestramento, ma perché comprende a un livello più profondo la relazione causale tra sporco e pulizia, e le azioni che permettono di passare da uno stato all’altro. Questi modelli sono addestrati a comprendere e ricostruire le proprietà fondamentali dei mondi rappresentati, inferendo regole, relazioni causali e strutture ontologiche che li governano. Hanno il potenziale per generare ambienti coerenti e interattivi in cui le regole fisiche, sociali e narrative sono consistenti e prevedibili, permettendo forme di simulazione e sperimentazione precedentemente impossibili. L’applicazione di questi world models agli universi narrativi potrebbe consentire la creazione di spazi di simulazione in cui esplorare conseguenze controfattuali, sviluppi narrativi alternativi e interpretazioni divergenti delle ambiguità testuali. Il futuro prossimo delle tecnologie convergenti suggerisce la possibilità di generare interi mondi narrativi navigabili: ambienti virtuali generati dinamicamente a partire da testi letterari, interazioni naturali con personaggi dotati di agency basata su LLM, e ricostruzioni coerenti di ambientazioni, atmosfere e rela-

25 Si veda Ha-Schmidhuber 2018; Garrido *et al.* 2024.

zioni causali testuali. Questa possibilità solleva questioni fondamentali per gli studi culturali e letterari, a partire dal problema della validità epistemica di tali ricostruzioni. Questioni la cui soluzione richiede un serio dialogo interdisciplinare in grado di sviluppare approcci che siano sia tecnicamente solidi sia ermeneuticamente produttivi.

6. Limiti e prospettive

Nonostante le potenzialità innovative che abbiamo ampiamente discusso, è fondamentale riconoscere che i modelli e sistemi generativi attuali presentano significativi limiti epistemici che ne condizionano l'applicabilità nel contesto degli studi letterari e culturali. Un primo limite riguarda l'affidabilità epistemica: questi modelli non producono conoscenza epistemicamente certa e sono soggetti a output erronei o inventati, impropriamente definiti 'allucinazioni' nel gergo tecnico. Queste 'allucinazioni' non sono semplici errori casuali, ma derivano dalle modalità stesse di funzionamento dei modelli probabilistici, che possono generare contenuti plausibili ma fattualmente errati o completamente fittizi, un fenomeno particolarmente problematico nel contesto della ricerca accademica, dove l'accuratezza fattuale è un requisito imprescindibile. Ma ancora più critica è una limitazione strutturale di questi sistemi: la loro "opacità" epistemica. Infatti, non è possibile spiegare in modo certo e completo come funzionano i loro processi inferenziali interni, quali rappresentazioni intermedie costruiscono e quali criteri utilizzano per selezionare determinate interpretazioni rispetto ad altre. Questa caratteristica, nota come il problema della "black box" dell'intelligenza artificiale, pone sfide significative alla validazione scientifica dei risultati ottenuti e alla loro integrabilità all'interno di metodologie critiche tradizionali che valorizzano la trasparenza del processo interpretativo. La validazione delle conoscenze generate dagli LLM rappresenta una sfida epistemologica complessa che richiede approcci diversificati. L'Explainable AI²⁶ e l'interpretazione meccanicistica dei modelli neurali²⁷ offrono potenziali vie per rendere più trasparenti i processi interni di questi sistemi, ma incontrano limiti significativi: problemi di calcolabilità dovuti alla dimensione dei modelli, il fenomeno della

26 Si veda Barredo *et al.* 2020.

27 Si veda Millière 2024; Bereska-Gavves 2024.

sovrapposizione (*superposition*) che rende difficile isolare rappresentazioni specifiche all'interno delle reti neurali, e il problema fondamentale del livello di descrizione appropriato.²⁸

Rimane molto da riflettere e sperimentare sulla natura dei sistemi generativi e sul loro ruolo come potenziali modelli della cognizione umana e dei processi culturali. Ma, indipendentemente da quanto si dimostreranno modelli prossimali della mente umana (una questione che rimane aperta e dibattuta), la loro capacità di modellizzare i sistemi astratti apre enormi spazi di sperimentazione sui fenomeni culturali, letterari e artistici. Le potenzialità analitiche ed ermeneutiche emergenti dalla convergenza tra mondi digitali immersivi e intelligenza artificiale generativa aprono un vasto spazio inesplorato per gli studiosi umanisti capaci di attraversare in entrambe le direzioni i ponti dell'interdisciplinarietà, mantenendo al contempo il rigore metodologico delle proprie discipline d'origine e l'apertura verso linguaggi e paradigmi disciplinari diversi. In questa prospettiva, la convergenza non rappresenta semplicemente l'introduzione di nuovi strumenti nell'arsenale metodologico delle discipline umanistiche, ma l'emergere di un nuovo paradigma epistemologico che richiede una profonda riconsiderazione dei fondamenti teorici e metodologici di queste discipline. Questo processo di ripensamento, lungi dall'impovertire la tradizione umanistica, ha il potenziale per rivitalizzarla e arricchirla, aprendo nuove prospettive su questioni fondamentali relative alla natura del significato, dell'interpretazione, della narratività e dell'esperienza estetica.

Bibliografia

- Alpaydin 2010 = Ethem Alpaydin, *Introduction to Machine Learning*. 2nd ed. Adaptive Computation and Machine Learning, Cambridge (Mass), MIT Press, 2010. <http://mitpress.mit.edu/books/introduction-machine-learning>
- Bachtin 1997 = Michail Michailovič Bachtin, *Estetica e romanzo. Teoria e storia del discorso narrativo*, Milano, Einaudi, 1997.
- Barredo *et al.* 2020 = Alejandro Arrieta Barredo, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barba, Salvador García, *Explainable Artificial Intelligence (XAI): Concepts,*

28 Si veda Elhage *et al.* 2022; Scherlis *et al.* 2022.

- taxonomies, opportunities and challenges toward responsible AI*, «Information Fusion», 58 (2020), pp. 82–115.
- Bender *et al.* 2021 = Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, e Shmargaret Shmitchell, *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, New York, Association for Computing Machinery, 2021, pp. 610–623. <https://doi.org/10.1145/3442188.3445922>
- Bender-Koller 2020 = Emily M. Bender, Alexander Koller, *Climbing towards NLU: On meaning, form, and understanding in the age of data*, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5185-5198. Online: Association for Computational Linguistics.
- Bareska-Gavves 2024 = Leonard Bareska, Efstratios Gavves, *Mechanistic Interpretability for AI Safety – A Review*, arXiv preprint arXiv:2404.14082. <https://doi.org/10.48550/ARXIV.2404.14082>
- Bolter-Grusin 2023 = Jay D. Bolter, Richard Grusin, *Remediation. Competizione e Integrazione Tra Media Vecchi e Nuovi*, a cura di A. Marinelli, trad. in it. da B. Gennaro, Milano, Guerini e Associati, 2023.
- Bould 2005 = Mark Bould, *Cyberpunk*, in *A Companion to Science Fiction*, a cura di David Seed, Oxford, UK: Blackwell Publishing Ltd, pp. 217-238. <https://doi.org/10.1002/9780470997055.ch15>
- Brown *et al.* 2020 = Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei, *Language Models are Few-Shot Learners*, in *Advances in Neural Information Processing Systems 33*, a cura di H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, e H. Lin, Curran Associates, pp. 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf
- Caronia-Gallo 1996 = Antonio Caronia, Domenico Gallo, *Houdini e Faust: breve storia del Cyberpunk*, Milano, Baldini & Castoldi, 1996.
- Chalmers 2023 = David J. Chalmers, *Could a Large Language Model Be Conscious?*, «Boston Review», 1 (2023). <https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/>

- Ciotti 2023a = Fabio Ciotti, *Digital humanities. Metodi, strumenti, saperi*, Roma, Carocci editore, 2023.
- Ciotti 2023b = Fabio Ciotti, *Minerva e il pappagallo. IA generativa e modelli linguistici nel laboratorio dell'umanista digitale*, «Testo e Senso» 26 (2023), pp. 289–315. <https://doi.org/10.58015/2036-2293/671>
- Dennet 1989 = Daniel C. Dennet, *The Intentional Stance*, Cambridge (Mass), MIT Press, 1989.
- Dennet 1997 = Daniel C. Dennet, *True Believers: The Intentional Strategy and Why It Works*, in *Mind Design II*, a cura di John Haugeland, Cambridge (Mass), MIT Press, pp. 57-80. <https://doi.org/10.7551/mitpress/4626.003.0003>
- Dennet 2024 = Daniel C. Dennet, *Pensandoci bene: avventure nella filosofia*. Milano, Raffaello, 2024.
- Dolézel 1999 = Lubomír Dolézel, *Heterocosmica: fiction e mondi possibili*, Milano, Bompiani, 2024.
- Eco 1979 = Umberto Eco, *Lector in fabula*, Milano, Bompiani, 1979.
- Ehlagé *et al.* 2022 = Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Kaplan, Dario Amodei, Martin Wattenberg, Christopher Olah, *Toy Models of Superposition Transformer Circuit Thread*, 2022. https://transformer-circuits.pub/2022/toy_model/index.html
- Firth 1957 = John Rupert Firth, *Papers in Linguistics 1934–1951*, London, Oxford University Press, 1957.
- Garrido *et al.* 2024 = Quentin Garrido, Mahmoud Assran, Nicolas Ballas, Adrien Bardes, Laurent Najman, Yann LeCun, *Learning and Leveraging World Models in Visual Representation Learning*, arXiv (<https://doi.org/10.48550/ARXIV.2403.00504>).
- Gibson 2019 = William Gibson, *Trilogia dello Sprawl*, trad. in it. da G. Cossato, Milano, Mondadori, 2019.
- Ha-Schmidhuber 2018 = David Ha, Jürgen Schmidhuber, *World Models*, arXiv 1803.10122. <https://doi.org/10.48550/ARXIV.1803.10122>
- Kovač *et al.* 2023 = Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, Pierre-Yves Oudeyer, *Large Language Models as Superpositions of Cultural Perspectives*, arXiv:2307.07870. <https://doi.org/10.48550/ARXIV.2307.07870>
- Landow 1992 = George P. Landow, *Hypertext: The Convergence of Contemporary Critical Theory and Technology*, Baltimore, Johns Hopkins University Press, 1992.

- Landow 1994 = *Hyper/text/theory*, a cura di George P. Landow, Baltimore, Johns Hopkins University Press, 1994.
- Lederman-Mahowald 2024 = Harvey Lederman, Kyle Mahowald, *Are Language Models More Like Libraries or Like Librarians? Bibliotechnism, the Novel Reference Problem, and the Attitudes of LLM*, arXiv. <https://doi.org/10.48550/arXiv.2401.04854>
- Lotman 1979 = Jurij Michajlovič Lotman, *Culture as collective intellect and the problems of artificial intelligence*, «Russian Poetics in translation», 6 (1979), pp. 84–96.
- Lotman 1992 = Jurij Michajlovič Lotman, *La semiosfera: l'asimmetria e il dialogo nelle strutture pensanti*, a cura di S. Salvestroni, Venezia, Marsilio, 1992.
- Lotman-Uspenskij 2001 = Jurij Michajlovič Lotman, Boris Andreevič Uspenskij, *Tipologia della cultura*, Milano, Bompiani, 2001.
- McGann 2001 = Jerome McGann, *Radiant Textuality: Literature after the World Wide Web*, Palgrave, Macmillan, 2001.
- Milgram, Kishino 1994 = Paul Milgram, Fumio Kishino, *A Taxonomy of Mixed Reality Visual Displays*, «IEICE Trans. Information Systems E77-D», 12 (1994), pp. 1321-1329.
- Millière 2024 = Raphaël Millière, *Philosophy of Cognitive Science in the Age of Deep Learning*, «WIREs Cognitive Science», 15/5 2024, e1684. <https://doi.org/10.1002/wcs.1684>
- Millière-Buckner 2024 = Raphaël Millière, Cameron Buckner, *A Philosophical Introduction to Language Models – Part I: Continuity With Classic Debates*, arXiv. <https://doi.org/10.48550/ARXIV.2401.03910>
- Mitchell 2022 = Melanie Mitchell, *L'intelligenza artificiale. Una guida per esseri umani pensanti*, Torino, Einaudi, 2022.
- Pavel 1986 = Thomas G. Pavel, *Fictional Worlds*, Cambridge (Mass), Harvard University Press, 1986.
- Piantadosi 2024 = Steven T. Piantadosi, *Modern language models refute Chomsky's approach to language*, in *From fieldwork to linguistic theory: A tribute to Dan Everett*, a cura di E. Gibson e M. Poliak, Berlin, Language Science Press, 2024.
- Pimentel-Teixeira 1993 = Ken Pimentel, Kevin Teixeira, *Virtual Reality: Through the New Looking Glass*, New York, Intel/Windcrest, 1993. <https://books.google.it/books?id=ErrumAEACAAJ>

- Raschka 2023 = Sebastian Raschka, *Understanding Reasoning LLMs*, in *Ahead of AI* (blog). 16 aprile 2023. <https://magazine.sebastianraschka.com/p/understanding-reasoning-llms>
- Rheingold 1991 = Howard Rheingold, *Virtual reality*, New York, Summit Books, 1991.
- Roncaglia 2023 = Gino Roncaglia, *L'architetto e l'oracolo. Forme digitali del sapere da Wikipedia a ChatGPT*, Roma-Bari, Laterza, 2023.
- Ryan 1991 = Marie-Laure Ryan, *Possible Worlds, Artificial Intelligence, and Narrative Theory*, Bloomington, Indiana University Press, 1991.
- Ryan 2003 = Marie-Laure Ryan, *Narrative as Virtual Reality: Immersion and Interactivity in Literature and Electronic Media*, Baltimore, Johns Hopkins University Press, 2003.
- Ryan 2015 = Marie-Laure Ryan, *Narrative as Virtual Reality 2: Revisiting Immersion and Interactivity in Literature and Electronic Media*, Baltimore, Johns Hopkins University Press, 2015.
- Scherlis *et al.* 2022 = Adam Scherlis, Kshitij Sachan, Adam S. Jermyn, Joe Benton, Buck Shlegeris, *Polysemanticity and Capacity in Neural Networks*, arXiv. <https://doi.org/10.48550/ARXIV.2210.01892>
- Searle 1980 = John R Searle, *Minds, Brains, and Programs*, «Behavioral and Brain Sciences», 3/3 (1980), pp. 417-424. <https://doi.org/10.1017/S0140525X00005756>).
- Stephenson 1992 = Neal Stephenson, *Snow crash*, New York, Bantam Books, 1992.
- Underwood 2021 = Ted Underwood, *Mapping the Latent Spaces of Culture*, in *The Stone and the Shell* (blog). 21 ottobre 2021. <https://tedunderwood.com/2021/10/21/latent-spaces-of-culture/>
- Vaswani *et al.* 2017 = Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, e Illia Polosukhin, *Attention Is All You Need*, arXiv:1706.03762. <https://doi.org/10.48550/ARXIV.1706.03762>
- Walton 1993 = Kendall L. Walton, *Mimesis as Make-Believe: On the Foundations of the Representational Arts*, Cambridge (Mass), Harvard University Press, 1993.
- Xu *et al.* 2025 = Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, Chenyang Shao, Yuwei Yan, Qinglong Yang, Yiwen Song, Sijian Ren, Xinyuan Hu, Yu Li, Jie Feng, Chen Gao, Yong Li, *Towards Large Reasoning Models: A Survey of Reinforced*

Reasoning with Large Language Models, arXiv. <https://doi.org/10.48550/ARXIV.2501.09686>

Yiu-Kosoy-Gopnik 2023 = Eunice Yiu, Eliza Kosoy, Alison Gopnik, *Transmission versus truth, imitation versus innovation: What children can do that large language and language-and-vision models cannot (yet)*, «Perspectives on Psychological Science», 19/5 (2025). <https://doi.org/10.1177/17456916231201401>