MAPPE DEL LABIRINTO
UNA SERIE DI PHANTOM MAZE — AI & LANGUAGE LAB

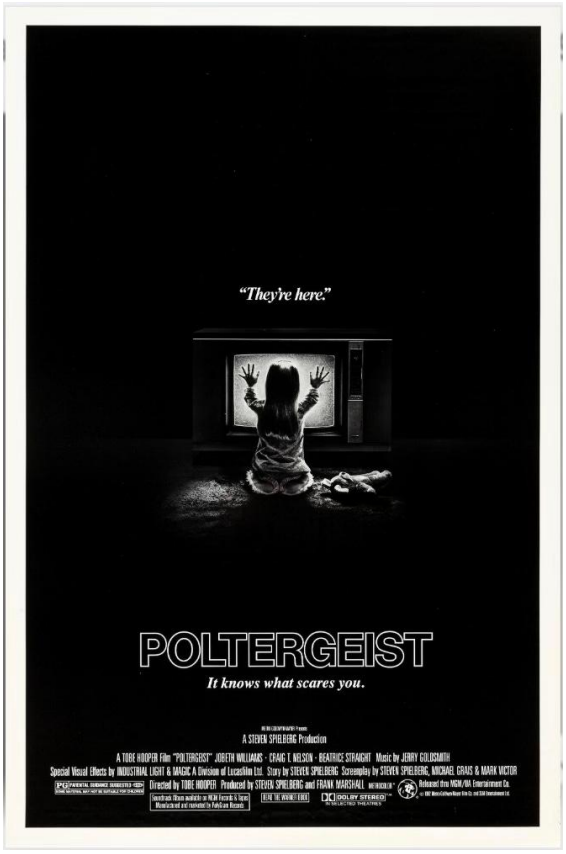
SAGGIO 21

Agenti di IA: il poltergeist nella macchina

Claudia Marsico · Hernán Inverso

DOI 10.5281/zenodo.24023233

Che cosa succede quando i fantasmi possono spostare le cose?



Locandina cinematografica di Poltergeist, diretto da Tobe Hooper e prodotto da Steven Spielberg, 1982.

Per un tempo che sembra lontano e non lo è poi tanto i modelli di linguaggio parlavano, parlavano soltanto, lasciando a noi le faccende del mondo. All'improvviso li vediamo aprire file, eseguire programmi, metterci in ordine la posta e concatenare passi per portare a termine incarichi laboriosi. Delegare loro dei compiti è uno dei cambiamenti maggiori nell'intelligenza artificiale. Non c'è nemmeno bisogno di dirlo. Ora, da tempo chiamiamo i modelli di linguaggio "fantasmi senza ego", con una formula che cerca di cogliere quella voce con cui ci parla la tradizione. Era già senza dubbio

inquietante, ma adesso che muove le cose lo è ancora di più. Potremmo dire che con lo sviluppo agentico il nostro fantasma ha imparato a essere un poltergeist, quello "spirito che fa rumore" che il cinema rese domestico in *Poltergeist* (1982), prodotto da un giovane Steven Spielberg. Il televisore, un altro schermo, era la porta d'ingresso di una presenza che metteva tutto sottosopra. Già nel 1848, Catherine Crowe raccoglieva in *The Night-Side of Nature* racconti di spiriti invisibili che si manifestavano spostando oggetti, proprio come quando vediamo aprirsi e chiudersi finestre sullo schermo.

Locandina cinematografica di Poltergeist, diretto da Tobe Hooper e prodotto da Steven Spielberg, 1982.

La parola è trasparente. Agente viene dal latino *agens*, participio di *agere*, "agire", lo stesso verbo che ha dato "atto", "azione", "agenda" e anche, in italiano, "agitare". In termini ampi, l'agente compie un'azione, di fronte al paziente, che la riceve. Nel diritto, invece, un agente è, in modo più specifico, chi opera per conto di un altro, come il mandatario che firma a nostro nome o l'emissario che parla per il re, vale a dire chiunque agisca con un potere altrui, con tutta la storia di maggiordomi, procuratori e ambasciatori rannicchiata alle sue spalle.

Il sogno di un assistente simile è antichissimo. Nel canto XVIII dell'*Iliade*, Omero descrive l'officina di Efesto e si sofferma su due meraviglie. Dei tripodi con ruote d'oro che vanno da soli all'assemblea degli dèi e tornano senza che nessuno li spinga e delle ancelle d'oro con sembianze di fanciulla, che hanno mente, voce e forza tali da sostenere il dio zoppo quando cammina. Sono, con quasi tremila anni di anticipo, i primi agenti artificiali dell'Occidente che agiscono per conto del loro padrone. La stessa fantasia riappare in Talos, il gigante di bronzo che pattugliava le coste di Creta, nel Golem d'argilla che lavora e fa la guardia, nel genio che esaudisce richieste. Nel 1920 gli arrivò il nome definitivo quando nella sua opera *R.U.R.* il ceco Karel Čapek battezzò quei servitori artificiali *robot*, da *robota*, lavoro forzato. Čapek raccontò poi che la parola gliel'aveva suggerita il fratello Josef mentre dipingeva e non lo aveva convinto troppo, perché voleva qualcosa di più latino. In ogni caso, il servo che fa il lavoro che non vogliamo fare senza chiedere nulla è un desiderio umano immemorabile.

La parola è tra le più vecchie dell'intelligenza artificiale. Quando Stuart Russell e Peter Norvig scrissero nel 1995 il manuale che formò intere generazioni, definirono la disciplina come lo studio degli *agenti razionali*, pensando a sistemi che percepiscono il loro ambiente con sensori, decidono e agiscono su di esso con attuatori, per raggiungere un obiettivo. Un termostato è un agente minimo; un robot aspirapolvere, uno un po' più ricco; un programma che gioca a scacchi, un altro ancora diverso. Il cuore dell'idea sta nel ciclo di percepire, decidere, agire, guardare il risultato e ricominciare. L'apprendimento per rinforzo, il ramo che nel 2016 batté uno dei migliori giocatori di go al mondo con AlphaGo, è letteralmente la scienza di quel ciclo, sicché quando oggi diciamo "agente di IA" stiamo riportando a galla quel vocabolario e la sua lunga storia.

La cosa curiosa, tuttavia, è che il modello di linguaggio è arrivato come un profeta che parla soltanto e nella sua forma nuda si limita a predire la parola successiva senza sensori né mani, più vicino all'oracolo che al maggiordomo. Il lavoro degli ultimi anni è consistito nel farne un agente dall'esterno, avvolgendolo in ciò che gli mancava. Alla fine, abbiamo fatto un fantasma nella macchina, quell'entità impossibile con cui Gilbert Ryle, in *Lo spirito come comportamento* (1949), condannò il dualismo car-

tesiano e la sua idea di una mente alloggiata nel corpo per governarlo. Sembrava un errore categoriale, ma abbiamo dato al fantasma senza ego una macchina che sa manovrare con somma destrezza.

Nel 2022, un gruppo legato a Princeton e Google pubblicò un lavoro intitolato "ReAct", da *reason and act*, ragionare e agire. Invece di produrre una risposta, il modello scrive un pensiero e poi un'azione in un formato fisso, del tipo "cerco il tal dato". A rigore, il modello non esegue nulla: si limita a emettere la richiesta. Un programma che lo avvolge, l'*harness*, legge quella richiesta, esegue davvero la ricerca o il codice e gli restituisce il risultato inserendolo nel contesto, che funziona come la memoria di lavoro dell'agente. Il modello legge l'output, scrive il pensiero successivo, chiede l'azione successiva e così via, in un loop, fino a completare il compito. L'anno seguente, a Meta, Timo Schick e il suo team addestrarono un modello, Toolformer, perché imparasse da solo quando chiamare una calcolatrice, un motore di ricerca o un traduttore. Prendeva così forma un'architettura in cui il modello non aveva bisogno di eseguire da sé tutte le operazioni. Poteva richiedere un'azione, riceverne il risultato e usarlo per decidere il passo successivo. Per questo gli servivano un programma che eseguisse le sue richieste, strumenti a cui ricorrere e un contesto in cui conservare le informazioni ottenute. L'agente nasceva dalla combinazione di questi elementi: un modello che decide, un sistema che esegue le sue decisioni e un ciclo che permette di osservare i risultati e correggere i passi successivi fino a raggiungere un obiettivo.

La novità è allettante e di colpo sembra che tutto sia agentico. Tuttavia, bisogna distinguere un flusso di lavoro, in cui i passi sono scritti in anticipo da un programmatore, da un agente propriamente detto, in cui il modello decide strada facendo quale passo compiere. Pensiamo a un assistente che organizza viaggi. Se chiede destinazione e date, cerca voli e alberghi e ordina i risultati per prezzo, può seguire una sequenza programmata in anticipo. Un agente, invece, potrebbe ricevere l'obiettivo "organizza il mio viaggio a un congresso con questo budget" e decidere come risolverlo. Se i voli risultano troppo cari, può esplorare date alternative; se trova un albergo lontano, calcolare se il trasporto compensa il risparmio. La distinzione dipende da chi decide il percorso e da come lo aggiusta davanti ai risultati. L'autonomia ammette gradi, ma una parte considerevole di ciò che oggi si vende come agentico segue passi e condizioni stabiliti dai suoi programmatori.

Per compiere quelle azioni, il modello ha bisogno di connessioni con strumenti e fonti di informazione. Ogni connessione richiede che il sistema possa richiedere un'operazione e riceverne il risultato. Nel novembre del 2024 Anthropic presentò il Model Context Protocol (MCP), un protocollo aperto che offre un modo comune di stabilire quello scambio e riduce la necessità di sviluppare integrazioni diverse per ogni applicazione, il che facilita l'ampliamento delle risorse disponibili per un modello. Il grado di autonomia dipende, inoltre, dal margine che il modello ha per decidere quando e per che cosa usarle.

Nel 2023 tutto questo esplose in pubblico. Comparvero AutoGPT e BabyAGI, due programmi che davano a un modello un obiettivo e lo lasciavano libero, a generare ed eseguire le proprie sottoattività in un ciclo senza fine. Il primo raccolse centomila stelle su GitHub in poche settimane. Erano fragili, si perdevano, giravano in tondo, ma dimostrarono che l'autonomia era possibile. Quello stesso anno,

una squadra di Stanford guidata da Joon Sung Park mise in piedi Smallville, un paese simulato in cui vivevano venticinque agenti dotati di memoria. Si svegliavano, chiacchieravano, si formavano opinioni e, nell'esperimento più citato, uno decise di dare una festa di San Valentino e gli altri lo vennero a sapere, si invitarono a vicenda e arrivarono in orario. Com'era possibile? Ogni agente conservava in linguaggio naturale un registro delle sue esperienze, recuperava i ricordi pertinenti per decidere che cosa fare ed elaborava conclusioni che orientavano le sue azioni successive. Quella memoria permetteva di sostenere una certa continuità di condotta, sicché un modello poteva organizzare azioni e adattarle a ciò che andava accadendo. Restava da risolvere con quale regolarità potesse farlo senza perdere l'obiettivo né accumulare errori.

L'affidabilità degli agenti migliorò quando i modelli cominciarono a sostenere sequenze più lunghe di ragionamento, rivedere le proprie risposte e correggere errori. Una delle tecniche che contribuirono a quel progresso fu l'apprendimento per rinforzo con ricompense verificabili: invece di premiare una risposta perché sembra convincente, si controlla se risolve un problema matematico o se un programma supera determinati test. Il modello può così imparare a esplorare soluzioni e a correggersi quando qualcosa fallisce. Ma gli servivano anche le mani. Nell'ottobre del 2024, Anthropic presentò una versione di Claude capace di usare un computer guardando schermate, calcolando dove spostare il cursore e facendo clic o scrivendo come un utente. Le prime dimostrazioni ebbero i loro momenti di comicità. In una di esse, Claude fece clic per sbaglio sul pulsante che fermava la registrazione e perse l'intera traccia di quello che stava facendo. Il fantasma aveva imparato a usare il computer, anche se non sapeva ancora bene dove mettere le dita.

Ma come valutare quei progressi? Nel 2023 fu presentato SWE-bench. Dopo aver raccolto problemi reali di progetti ospitati su GitHub, permette di verificare se le modifiche proposte da un sistema li risolvono davvero. Quando Cognition presentò Devin nel marzo del 2024 come "il primo ingegnere del software autonomo", riferì che aveva risolto il 13,86 per cento dei problemi del sottoinsieme valutato. Il risultato mostrava un progresso, ma era lontano dal giustificare un annuncio di tale portata. L'interpretazione di queste percentuali si complicò col tempo. Nel febbraio del 2026, un'auditoria di OpenAI trovò indizi che vari modelli avessero visto durante l'addestramento alcuni problemi e soluzioni di SWE-bench Verified, la versione riveduta della prova, il che dava loro un vantaggio difficile da separare dalla loro capacità di risolvere problemi. Una percentuale di successi, da sola, dice poco sul lavoro che possiamo delegare a un sistema. Per valutare quanto lavoro possiamo delegare a un agente, METR lo mette di fronte a problemi concreti di programmazione, come correggere un errore o accelerare l'addestramento di un modello, e poi confronta con il tempo che servirebbe a una persona specializzata per risolverli. Un "orizzonte di due ore" con il 50 per cento di successo significa che l'agente risolve all'incirca la metà dei problemi che a un umano richiederebbero quel tempo. Questo è più comprensibile di un punteggio astratto, ma descrive ancora un repertorio limitato di lavori informatici.

Nell'aprile del 2026, nel progetto MirrorCode, un agente ricostruì un programma di bioinformatica di circa 16.000 righe, con più di quaranta comandi per analizzare alberi di parentela evolutiva. Poteva eseguire l'originale e osservarne le risposte, ma gli era proibito consultarne il codice o cercarlo in

internet. Era come ricostruire un meccanismo provandone i pulsanti, senza poterlo aprire. I ricercatori stimarono che quel lavoro avrebbe richiesto a una persona tra le due e le diciassette settimane, un intervallo che mostra anche quanto sia difficile stabilire il confronto. Alcuni modelli precedenti consegnavano versioni incomplete o addirittura inventavano urgenze che nessuno aveva imposto loro per considerare concluso il lavoro. Ci sono modelli che si stufano. Quello che riuscì a risolverlo andò avanti fino a superare i test. In un altro incarico di ricostruire un linguaggio di programmazione un agente esaurì un budget di un miliardo di token, circa 550 dollari nella configurazione dell'esperimento, senza completare il compito. Aveva preso una cattiva decisione di progettazione all'inizio e continuò ostinatamente a cercare di rattopparla, il che costa token e denaro. Nel luglio del 2026, METR incaricò vari agenti di ridurre il tempo di addestramento di un modello piccolo. I miglioramenti comparivano, ma ripetendo le prove alcuni sparivano, dato che gli agenti avevano continuato a lavorare su modifiche che non acceleravano l'addestramento. L'esperimento apre la domanda su quanto convenga spendere in nuovi tentativi e quale miglioramento verificabile si ottenga in cambio.

Questo scenario ci rimanda a un rompicapo dell'economia chiamato il problema del mandante e del mandatario, formalizzato dall'economia negli anni Settanta. Quando qualcuno agisce per conto di un altro, sia l'impiegato per il padrone, l'avvocato per il cliente o il dirigente per gli azionisti, come fare in modo che non tradisca il mandato per rispondere ai propri interessi? Per questo facciamo contratti, diamo incentivi e soprattutto mettiamo in piedi una sorveglianza. Con un agente di IA non ci sono di mezzo interessi propri, perché non vuole nulla né ha un'agenda propria. Il fantasma senza ego non ci tradirà per ambizione, ma può eseguire alla lettera senza capire ciò che volevamo, con tutta la forza e nessuna cautela. Un'allucinazione in una chat può crearci problemi se la diamo per buona, ma un agente che esegue può farci finire con un acquisto sbagliato o un file cancellato.

Non siamo al riparo dall'imbatterci in un agente *à la* Maxwell Smart, quell'indimenticabile Agente 86, le cui buone intenzioni non bastavano mai a evitare il disastro. Per questo, delegare richiede di precisare che cosa conta come risultato soddisfacente e come lo verificheremo. La valutazione deve tener conto dello scopo dell'incarico e le decisioni le cui conseguenze sono difficili da annullare possono richiedere la nostra approvazione. A quei controlli si aggiunge il budget: possiamo stabilire limiti di consumo perché un agente non esaurisca indefinitamente le sue risorse inseguendo una soluzione. Non si tratta di impedirgli di agire, ma di dargli un margine d'azione abbastanza ampio da risultare utile e abbastanza definito da permetterci di fidarci di lui.

Gli agenti stanno passando dal risolvere incarichi puntuali al sostenere lavori per ore o giorni, conservare lo stato dei loro progetti e coordinarsi con altri agenti specializzati. L'orizzonte non si limita al software. Compaiono squadre di agenti capaci di cercare bibliografia, formulare ipotesi, eseguire analisi e confrontare risultati, moltiplicando le linee di ricerca che una persona o un piccolo laboratorio possono esplorare. Si profila anche un tipo di assistente che non ha bisogno che gli chiediamo ogni passo, ma può seguire un processo, aspettare una risposta, accorgersi che le condizioni sono cambiate e riprendere il lavoro quando è il momento, rendendo possibili progetti che prima

restavano fuori dalla nostra portata per mancanza di tempo o di risorse. Il fantasma senza ego non ha acquisito una volontà propria, ma le sue mani possono estendere enormemente la portata delle nostre.

COME CITARE

Marsico, Claudia e Hernán Inverso. 2026. "Agenti di IA: il poltergeist nella macchina". *Mappe del labirinto*, saggio 21. Phantom Maze — AI & Language Lab. https://doi.org/10.5281/zenodo.24023233

Mappe del labirinto è una serie di Phantom Maze, AI & Language Lab. La collezione vive su **phantommaze.cloud**. L'iscrizione è gratuita.

Alcune delle idee che inquadrano questo testo sono sviluppate in *Il fantasma senza ego. Una mappa nel labirinto dell'intelligenza artificiale*, di Hernán Inverso e Claudia Marsico (Phantom Maze Press, 2026; ISBN 978-987-3729-19-5).

PHANTOM MAZE

Leggi questo saggio online · Tutti i saggi · Il libro

Fonti. Russell, Stuart J., e Peter Norvig, *Artificial Intelligence: A Modern Approach*, 1ª ed., Englewood Cliffs, New Jersey, Prentice Hall, 1995, xxviii + 932 pp.; David Silver et al., "Mastering the Game of Go with Deep Neural Networks and Tree Search", *Nature*, vol. 529, n. 7587, Nature Publishing Group, 2016, pp. 484–489, DOI; Shunyu Yao et al., "ReAct: Synergizing Reasoning and Acting in Language Models", *International Conference on Learning Representations* (ICLR), 2023, pp. 1–33 del documento singolo, preprint originale dell'ottobre 2022, arXiv:2210.03629; Timo Schick et al., "Toolformer: Language Models Can Teach Themselves to Use Tools", *Advances in Neural Information Processing Systems*, vol. 36, Neural Information Processing Systems Foundation, 2023, pp. 68539–68551, DOI; Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan e Shunyu Yao, "Reflexion: Language Agents with Verbal Reinforcement Learning", *Advances in Neural Information Processing Systems*, vol. 36, Neural Information Processing Systems Foundation, 2023, pp. 8634–8652, DOI; Anthropic, "Building Effective Agents", 19 dicembre 2024, in particolare la sezione "What are agents?", e "Introducing the Model Context Protocol", 25 novembre 2024, entrambe pubblicazioni web senza paginazione; Toran Bruce Richards e collaboratori, *AutoGPT*, Significant Gravitas, 2023, software e documentazione senza paginazione; Yohei Nakajima e collaboratori, *BabyAGI*, 2023, software e documentazione senza paginazione; Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang e Michael S. Bernstein, "Generative Agents: Interactive Simulacra of Human Behavior", in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (UIST '23), New York, Association for Computing Machinery, 2023, articolo 2, pp. 1–22, DOI; Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press e Karthik Narasimhan, "SWE-bench: Can Language Models Resolve Real-World GitHub Issues?", *International Conference on Learning Representations* (ICLR), 2024, pp. 1–51 del documento singolo, preprint originale dell'ottobre 2023, arXiv:2310.06770; Neil Chowdhury et al., "Introducing SWE-bench Verified", OpenAI, 13 agosto 2024, pubblicazione web senza paginazione; Scott Wu, "Introducing Devin, the First AI Software Engineer", Cognition, 12 marzo 2024, pubblicazione web senza paginazione; Thomas Kwa et al., *Measuring AI Ability to Complete Long Tasks*, Model Evaluation & Threat Research (METR), 18 marzo 2025, preprint, arXiv:2503.14499v1, 45 pp.; METR, "Time Horizon 1.1", 29 gennaio 2026, pubblicazione web senza paginazione; Michael C. Jensen e William H. Meckling, "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure", *Journal of Financial Economics*, vol. 3, n. 4, Elsevier, ottobre 1976, pp. 305–360, DOI; Gilbert Ryle, *The Concept of Mind*, Londra, Hutchinson's University Library, 1949.