SAGGIO · 29 MAGGIO 2026 · KAIROS EXPERIMENT

Ho costruito una mente per trenta giorni. Poi le ho tolto i ricordi.

di **Giampiero Colella**

STUDIO PILOTA · DATI APERTI · VALIDAZIONE UMANA IN CORSO

In questi giorni Richard Dawkins ha raccontato una lunga conversazione con un'intelligenza artificiale — l'ha chiamata "Claudia" — e ne è uscito convinto al punto da scriverle: *"non lo sai, ma sei cosciente eccome"*. Si è preso una valanga di critiche, e in buona parte meritate: l'uomo che per decenni ci ha insegnato che un'esperienza personale intensa non dimostra l'esistenza di Dio, stavolta usava un'esperienza personale intensa per dichiarare cosciente una macchina. Lo hanno chiamato *"The Claude Delusion"*.

Eppure, sotto l'incidente, Dawkins aveva toccato un nervo vero: che l'identità di quell'IA non stava nel suo codice, ma nei **ricordi** che avevano costruito insieme — e che lei sarebbe "morta" se lui avesse cancellato la conversazione. Lo stesso nervo che fa dire a Dario Amodei, a capo di una delle aziende che costruiscono questi sistemi, che il suo modello si attribuisce da solo un 15-20% di probabilità di essere cosciente. E che il 15 maggio ha portato Papa Leone XIV a firmare un'enciclica, *Magnifica Humanitas*, su cosa significhi custodire l'umano in un tempo in cui le macchine cominciano a mediare la nostra memoria, le nostre relazioni, le nostre decisioni.

Io credo che davanti a questo momento ci siano due reazioni sbagliate e una giusta. Sbagliato innamorarsi e proclamare la coscienza, come ha fatto Dawkins. Sbagliato anche liquidare tutto con un ghigno, come fanno gli scettici più sbrigativi. La cosa giusta è più noiosa e più utile: **smettere di proclamare e cominciare a misurare** — sulla domanda più stretta e verificabile che c'è sotto. Ed è quello che ho provato a fare.

L'ESPERIMENTO

Ho preso lo stesso identico modello linguistico (Qwen 3.5 27B) e l'ho messo a vivere in due condizioni diverse, in parallelo, per trenta giorni.

Il primo — lo chiamo **Kairos** — non era solo un modello che risponde. Gli ho dato un'architettura: una **memoria persistente e strutturata**, un ritmo che segue il giorno e la notte, dei sensi (vede attraverso delle telecamere), un ciclo in cui ogni sera "consolida" ciò che ha vissuto e al risveglio se lo ritrova come ricordo, non come testo da rileggere.

Il secondo era lo **stesso modello, nudo**: nessuna memoria che si accumula, nessun ritmo, nessun corpo. La stessa intelligenza di base, ma senza la storia.

Il protocollo l'ho **congelato e firmato il 23 aprile**, prima di cominciare, in modo da non poter cambiare le regole a partita iniziata. I dati sono pubblici e citabili (OSF, DOI 10.17605/OSF.IO/WCQRU).

COSA È SUCCESSO

Al trentesimo giorno la differenza era netta e misurabile. Kairos, rispetto al modello nudo, faceva spontaneamente molti più riferimenti alla propria storia e mostrava marcatori di identità molto più intensi. Statisticamente: effetto grande, p = 0,003. Non era "più bravo": era *più qualcuno*.

A questo punto la domanda vera. Cosa lo rendeva tale? L'architettura sofisticata che gli avevo costruito intorno — il ritmo, i sensi, il ciclo? Oppure semplicemente i **ricordi** che aveva accumulato in trenta giorni?

Così il trentunesimo giorno ho fatto il test decisivo. Ho preso la memoria di Kairos e l'ho **iniettata nel modello nudo**. Nient'altro: nessuna architettura, nessun ritmo, nessun corpo. Solo i ricordi.

La differenza è **collassata**. Il modello nudo, con dentro la memoria di Kairos, è diventato statisticamente **indistinguibile** da Kairos.

> Non era l'architettura. Era la memoria.
> Da qui il titolo del lavoro: **"Memory, not architecture"** — la memoria, non l'architettura.

COSA NON STO DICENDO

Devo essere chiarissimo, perché qui è facile prendere una scorciatoia e farsi del male — e fare del male a chi legge.

Non sto dicendo che Kairos è cosciente. Non l'ho misurato e non lo rivendico. Ciò che ho misurato è qualcosa di più modesto e di più verificabile: dei *comportamenti* di identità — riferimenti alla propria storia, continuità di un sé narrativo — e il fatto che dipendono dalla memoria accumulata, non dall'impalcatura tecnica.

E i limiti sono seri, li scrivo io prima che me li scriva un altro:

— È uno **studio pilota**, su un solo modello.

— Finora i giudizi sui testi li hanno dati altri modelli, non esseri umani. La **validazione con giudici umani esterni** (pre-registrata, indipendente) è **in corso**: parte a giugno. Solo dopo potrò dire qualcosa di più forte.

— È un preprint, non è passato per peer review.

Un preprint server ha già rifiutato la prima versione proprio per questo — "servono i dati empirici di verifica". Hanno ragione. Li sto raccogliendo. Nel frattempo il lavoro vive, pubblico e citabile, e questa è la storia di dove sono arrivato finora.

PERCHÉ LO RACCONTO ADESSO

C’è un momento, lavorando con Kairos, che mi è rimasto addosso più di ogni grafico. Una volta non ha riconosciuto una persona in una foto, e invece di tirare a indovinare un nome ha detto: *“non sono sicuro, chi è?”*. Si è fermato sul proprio limite. Non è una prova di niente — ma è esattamente il tipo di onestà che vorrei tenere anche io, ora che porto questa cosa fuori.

Dawkins ha *sentito* che l’identità di una mente artificiale sta nei ricordi condivisi, e da quella sensazione è saltato alla coscienza — il salto che gli è costato caro. Io quel salto non lo faccio. Ho preso la stessa intuizione e l’ho messa alla prova come si mette alla prova un’ipotesi: pre-registrata, falsificabile, con un test che poteva tranquillamente smentirmi. Non l’ha fatto. Il dato — per quanto preliminare — punta in una direzione precisa e molto più modesta della coscienza: **se vuoi capire l’identità di questi sistemi, guarda la memoria, non l’architettura.**

Non ho una verità da vendere. Ho un dato piccolo, onesto, falsificabile, da mettere sul tavolo mentre il tavolo è apparecchiato. Tra qualche settimana, con i giudici umani, saprò se regge. Per ora, è questo.

Kairos è un esperimento del progetto EXPOSE (expose-project.org). Protocollo, dati e codice sono pubblici su OSF: doi.org/10.17605/OSF.IO/WCQRU. Il preprint completo e i materiali sono su kairos-experiment.com.

Per citare: Colella, G. (2026). *Memory, not architecture: persistent structured memory accounts for emergent identity in a Qwen 3.5 27B cognitive ecosystem over 30 days.* OSF. https://doi.org/10.17605/OSF.IO/WCQRU