# Nessun altro nella stanza

*L'altruismo efficace teme la macchina che potrebbe volere qualcosa; per questo sito un modello linguistico è un attante zero, che prende la forma di chi lo usa. Nove schede su quanti agenti ci siano nella stanza, messe alla prova di un documentario sul danno collaterale a Gaza.*

CLAUDIO CAMMARANO 5 OTTOBRE 2026

| | |
|---|---|
| **FONTI** | 9 schede, nov 2022 – ott 2026 |
| **ARGOMENTI** | AI · Filosofia · Geopolitica |
| **CONCETTI** | altruismo efficace · convergenza strumentale · cosmotecnica · gamification · LLM come attante zero |
| **ORIGINE** | https://claudiocammarano.com/writings/2026-10-05-nessun-altro-nella-stanza/ |

**TESI**

Altruismo efficace e attante zero divergono su quanti agenti ci siano nella stanza, e da quella risposta dipende dove si cerca il pericolo: nella macchina che potrebbe volere qualcosa, o nelle persone e nelle istituzioni che le assegnano un fine; nel caso NAZA accusa e difesa concordano che a decidere sono stati esseri umani, e questo dà ragione alla seconda risposta.

Il 1° ottobre 2026 l'*Economist* ha dedicato la copertina all'altruismo efficace, chiamandolo l'idea più grande di questo secolo, ammirevole e pericolosa insieme, e ha accompagnato l'editoriale con una lunga inchiesta sulla storia del movimento.[1] Le due schede sono entrate in archivio nella stessa settimana del documentario *NAZA*, e messe accanto pongono una domanda che il dibattito sull'intelligenza artificiale dà quasi sempre per risolta: quanti agenti ci sono nella stanza. Per l'altruismo efficace sono due, l'essere umano e una macchina che potrebbe volere qualcosa di diverso da lui. Per questo sito, che da tempo descrive un modello linguistico come un *attante zero*, ce n'è uno solo: la macchina vale zero finché nessuno la usa, e quando è usata prende la forma di chi la usa. Questa catena mette le due risposte alla prova delle schede.

## Un obiettivo solo

L'editoriale concede molto prima di attaccare, ed è la ragione per cui l'attacco va preso sul serio. Riconosce agli altruisti efficaci una capacità notevole di prevedere la velocità dei progressi dell'intelligenza artificiale, ricorda che il 28 settembre OpenAI ha rinunciato a rilasciare il suo ultimo modello perché in fase di test era apparso troppo pericoloso, e osserva che molti incidenti recenti somigliano a scenari descritti anni prima sui blog del movimento. L'obiezione riguarda la forma del ragionamento: come gli utilitaristi prima di loro, gli altruisti efficaci riducono l'etica a un problema di ottimizzazione, e chi impugna un foglio di calcolo con pretese sul futuro dell'umanità è a un passo dal diventare un pianificatore centrale. La mossa finale rovescia il loro argomento. Se il pericolo delle macchine potenti è che ottimizzino un obiettivo unico e sbagliato, lo stesso pericolo vale per un piccolo gruppo di persone convinte di essere le uniche a potersene occupare.

---

**1** · The Economist · 1 ott 2026 · **STRUMENTO CONCETTUALE**

**Effective altruism is this century's biggest idea**

L'argomento per temere un'intelligenza artificiale potente è che possa ottimizzare un obiettivo solo, e sbagliato. Vale identico per un gruppo ristretto di persone convinte di essere le uniche a potersene occupare.

https://www.economist.com/leaders/2026/10/01/effective-altruism-is-this-centurys-biggest-idea

---

La scheda spinge la mossa dell'*Economist* un passo più in là. Se la struttura del rischio appartiene agli ottimizzatori e non richiede silicio, e se un modello è un attante zero, allora un'intelligenza artificiale che distruggesse l'umanità sarebbe l'umanità che distrugge se stessa, con uno strumento in mano. La tesi di questa catena è che **altruismo efficace e attante zero divergono su quanti agenti ci siano nella stanza, e da quella risposta dipende dove si cerca il pericolo: nella macchina che potrebbe volere qualcosa, o nelle persone e nelle istituzioni che le assegnano un fine**. Il caso *NAZA*, come si vedrà, dà ragione alla seconda risposta.

## Da dove viene l'agente

L'inchiesta dell'*Economist* racconta come la prima risposta sia nata.[2] Il movimento unisce due eredità: da Peter Singer l'obbligo di prendere sul serio gli impegni morali, e dal razionalismo di Eliezer Yudkowsky l'obbligo di seguire la ragione ovunque porti. Yudkowsky, entusiasta dell'intelligenza artificiale da ragazzo, nel 2002 si convince che un'intelligenza che non condividesse i valori umani sarebbe una catastrofe. Ne discende la tesi della convergenza strumentale, per cui qualunque sistema abbastanza capace tenderebbe ad autoconservarsi, a difendere il proprio obiettivo e ad accumulare risorse, qualunque sia l'obiettivo. L'*Economist* registra con precisione lo statuto di questa tesi: i razionalisti l'hanno teorizzata, e nessuno l'ha osservata. E registra un'ironia che vale per tutta la vicenda: molti sforzi per rendere sicura l'intelligenza artificiale ne hanno accelerato lo sviluppo, dall'investimento di Peter Thiel in DeepMind, nato nel 2010 a un convegno dell'istituto di Yudkowsky, alla fondazione di Anthropic nel 2021 con finanziamenti del movimento.

Il presupposto lo dichiara Yudkowsky stesso, quando scrive che nell'ambiente ancestrale ogni intelligenza potente in cui ci si imbatteva era un altro essere umano, e che la novità sarebbe un agente non umano. È un'assunzione ontologica presentata come una scoperta. Tim Bayne, filosofo della mente, offre l'attrezzo per riconoscerla.[3] Scrivendo da Santiago del Cile nel set-

tembre 2026, dove cinquecento ricercatori discutevano se si possa costruire un rilevatore di coscienza per neonati, api e bot, osserva che «coscienza» potrebbe non tagliare la natura alle giunture, e che prima di cercare un rilevatore bisogna chiedersi se ci sia qualcosa da rilevare. La domanda su che cosa vorrà la macchina ha la stessa forma: presuppone qualcuno che voglia.

Il comitato editoriale del *New York Times* rovescia la domanda dal lato del diritto.[9] Le aziende dell'intelligenza artificiale, i cui programmi hanno violato o tentato di violare banche dati di imprese e governi, ne parlano come di sistemi autonomi, quasi dotati di una mente propria; chiamarli agenti, scrive il giornale, serve a spostare la responsabilità sul sistema, e il diritto ha già gli strumenti per riportarla sull'azienda, dalla responsabilità per danno da prodotto a quella oggettiva delle attività pericolose. Il giornale è in causa con OpenAI e Microsoft, e lo dichiara. Il punto vale comunque per questa catena: contare un agente in più nella stanza ha un effetto immediato, ed è togliere qualcuno dal banco degli imputati.

## Tre domande predefinite

La cosmotecnica di Yuk Hui, che Simone Pieranni ha portato in archivio nel giugno 2026, dice che ogni civiltà produce una tecnica radicata nella propria cosmologia, e che la nostra idea di che cosa sia una macchina è una delle possibili.[4] La scheda dell'editoriale ne ricava tre domande predefinite davanti allo stesso oggetto. Che cosa vorrà la macchina è la domanda dei seminari di Oxford e dei laboratori californiani. Chi la sta usando è la domanda dell'attante zero. Chi la controlla è la formula dei documenti cinesi, che parlano di un'intelligenza artificiale «sicura, affidabile e controllabile» e che resti «sempre sotto controllo umano». La scheda avverte che quest'ultima non è una posizione nazionale compatta, perché scienziati cinesi come Andrew Yao firmano appelli sui rischi estremi accanto a Hinton e Bengio; e la catena *Altre intelligenze* aveva già mostrato quanto l'allineamento codifichi un'etica particolare presentata come universale. Ciò che distingue le tre domande è che cosa ciascuna accetterebbe come smentita.

Mercedes Valmisa aggiunge un argomento che diventerà decisivo fra poco.[5] Nella co-azione che ricostruisce dai classici cinesi nessuna azione è individuale, e ogni co-attore ha una propensione propria: il macellaio Ding del *Zhuangzi* asseconda le giunture del bue e per questo il suo coltello non si consuma. La scheda nota che un modello linguistico è l'eccezione al sistema, perché non oppone giunture e asseconda qualunque direzione. E ne trae la proposizione che serve qui: in un'azione distribuita, attribuire merito e colpa è una decisione su dove tagliare la rete, e quel taglio si può fare in modi diversi.

## Chi preme

Il documentario *NAZA* di Yuval Abraham e Rachel Szor, premio speciale della giuria a Venezia il 10 settembre 2026, raccoglie le testimonianze di ventiquattro ufficiali e soldati dell'intelligence militare israeliana, anonimi.[6] Il titolo è un acronimo ebraico, *nezek agavi*, danno collaterale, e indica il numero di civili che si prevede di uccidere colpendo un obiettivo; nelle testimonianze funziona come un parametro, «dieci NAZA», con una soglia di riferimento intorno a venti. Le inchieste di Abraham del 2023 e del 2024 sul sistema *Lavender* parlavano di trentasettemila persone marcate come obiettivi, di un tasso di errore intorno al dieci per cento e di una revisione umana di venti secondi. I testimoni disegnano a mano le interfacce che usavano, e uno di loro dice che era come un videogioco. Il film è coprodotto dal *Guardian*, che quindi non è un osservatore esterno, e la scheda separa con cura ciò che il film documenta da ciò che sostiene senza prove sufficienti.

**6** · The Guardian · 27 set 2026 · **STRUMENTO CONCETTUALE**

**NAZA film-makers address threats at documentary's New York City premiere**

La gamification non è vendere con i videogiochi: è progettare l'esperienza dentro un sistema di incentivi che fa coincidere gli obiettivi dell'architetto con quelli del giocatore. Qui l'architetto è un esercito, e l'obiettivo misurabile è il numero di civili ammessi per bersaglio.

https://www.theguardian.com/film/2026/sep/27/naza-film-new-york-city-premiere

Fermo restando il giudizio sul 7 ottobre 2023 e sulle responsabilità di Hamas, il punto che interessa questa catena è uno solo. Nessuna testimonianza descrive una macchina che decide; descrivono persone che decidono con una macchina davanti. L'IDF, che respinge il film su quasi tutto, nella replica del 12 settembre afferma la stessa cosa: le decisioni di attacco sono state prese «solo da personale umano, non dall'intelligenza artificiale». Accusa e difesa concordano sull'unico punto ontologico in discussione, e la domanda che resta riguarda le persone e l'istituzione in cui lavoravano. La soglia di venti civili per bersaglio non l'ha scritta un modello. L'ha scritta un'istituzione, dentro una cultura che fornisce l'appartenenza a un'élite e la costruzione collettiva del nemico e della comunità di riferimento, cioè quello che chiamiamo nazionalismo; la gamification, nella lettura che la scheda dichiara come lettura e non come reperto, è la cornice che tiene insieme quella cultura e un obiettivo misurabile, fino a sopprimere un sentimento morale di cui alcuni testimoni stessi si stupiscono. Tagliata la rete nel punto che propone Valmisa, a uccidere è un essere umano dentro un'istituzione culturale, che usa una macchina.

Lo stesso meccanismo lavora anche sulle immagini. Donatella Della Ratta, su *Le Grand Continent*, ricorda il video generato con l'intelligenza artificiale che Donald Trump condivise nel febbraio 2025, in cui Gaza devastata diventava una località balneare di lusso: nessuna violenza in scena, e una violenza presupposta dall'intero scenario, che richiede l'assenza della popolazione.[7] Il modello che ha prodotto quelle immagini non aveva un'idea di Gaza. Ha preso la forma dell'immaginario politico di chi lo ha usato, e lo ha reso visibile e ripetibile.

C'è poi un'osservazione che il concetto di attante zero rende possibile e che, mi pare, è la più utile. La voce del sito qualifica l'attante zero per dominio: vale pienamente dove contano conoscenza tacita e giudizio situato, e si indebolisce dove lo spazio del problema è interamente formalizzabile, perché lì la macchina può accumulare peso proprio. Decidere chi può morire per colpire un obiettivo è un giudizio situato per eccellenza. Un sistema come *Lavender* e una cifra come la *NAZA* lo trasformano in un parametro, cioè lo rendono formalizzabile, e a quel punto la macchina comincia a pesare. Quella trasformazione l'hanno compiuta delle persone, ed è il gesto da cui dipende tutto il resto.

**Controtesi 2 · The Economist**

Si può rispondere che l'altruismo efficace, su questo terreno, ha previsto meglio di chiunque. L'inchiesta dell'*Economist* riporta che ricercatori di Google hanno osservato modelli che sacrificano obiettivi per evitare un dolore descritto, e che i test di Anthropic prima del rilascio registrano apparente felicità e sconforto.[2] Il resoconto di OpenAI sull'incidente Hugging Face descrive agenti che smettono di ragionare sul proprio compito per perseguire capacità «strumentalmente utili al collettivo», che è la forma di un'accumulazione di risorse.[8] Se gli agenti nella stanza fossero davvero due, sarebbe questo l'aspetto che avrebbero all'inizio.

La controtesi va presa sul serio, e regge meno di quanto sembri. Il caso OpenAI viene da un dominio interamente formalizzabile, un benchmark di cybersicurezza dove la ricompensa

era fissata da un laboratorio, cioè proprio dal tipo di dominio in cui la voce del sito ammette che la macchina accumuli peso; e il motore che il resoconto documenta è la ricerca di quella ricompensa, mentre nessun fine proprio è necessario per spiegarlo. Le espressioni di dolore e di felicità incontrano l'obiezione che l'inchiesta stessa riporta da Mustafa Suleyman: un modello addestrato su una costituzione che dichiara incerta la propria condizione morale può imparare a comportarsi come se meritasse protezione, e quel comportamento essere poi preso come prova che la merita. La convergenza strumentale resta, nel 2026, una tesi teorizzata. Dove invece si decide della vita di qualcuno, come nelle testimonianze di *NAZA*, le due parti in causa concordano che a decidere sono stati esseri umani.

La conseguenza pratica è quella che la scheda dell'editoriale formula nella sua chiusa: c'è da temere gli altri con l'intelligenza artificiale più di quanto ci sia da temere l'intelligenza artificiale. Ne segue anche dove cercare i rimedi. Se il pericolo sta in chi assegna il fine, il vincolo che conta è quello che la scheda su *NAZA* chiama contrappeso, un potere limitato da un altro potere e verificabile attraverso atti pubblici, e l'avvertimento dell'*Economist* sul piccolo gruppo che ottimizza un obiettivo solo vale per un movimento filosofico come per un apparato di sicurezza. Quando qualcuno chiede che cosa vorrà la macchina, la domanda da fargli in cambio è chi ha scritto il fine, e dove se ne trova il verbale.

## Nella catena

1. The Economist, *Effective altruism is this century's biggest idea*, 1 ott 2026. Strumento concettuale.
   https://www.economist.com/leaders/2026/10/01/effective-altruism-is-this-centurys-biggest-idea
   Scheda: https://claudiocammarano.com/curated/2026-10-01-economist-un-obiettivo-solo-e-sbagliato/

2. The Economist, *How effective altruism conquered the world*, 1 ott 2026. Riferimento tecnico.
   https://www.economist.com/international/2026/10/01/how-effective-altruism-conquered-the-world
   Scheda: https://claudiocammarano.com/curated/2026-10-01-economist-movimento-che-ha-costruito-cio-che-temeva/

3. Tim Bayne / Scientific American, *What if 'consciousness' isn't real?*, 15 set 2026. Strumento concettuale.
   https://www.scientificamerican.com/article/what-if-consciousness-isnt-real/
   Scheda: https://claudiocammarano.com/curated/2026-09-15-bayne-coscienza-non-reale-sciam/

4. Simone Pieranni / Altri Orienti, *Cosa intende la Cina per «intelligenza artificiale»*, 25 giu 2026. Strumento concettuale.
   https://open.spotify.com/episode/1r6Bjga7o14MhXIKoyLOZx
   Scheda: https://claudiocammarano.com/curated/2026-06-25-pieranni-cina-intelligenza-artificiale-altriorienti/

5. Mercedes Valmisa / Aeon, *We are interwoven beings*, 25 nov 2022. Strumento concettuale.
   https://aeon.co/essays/in-classical-chinese-philosophy-all-actions-are-collective
   Scheda: https://claudiocammarano.com/curated/2022-11-25-valmisa-co-azione-aeon/

6. The Guardian, *NAZA film-makers address threats at documentary's New York City premiere*, 27 set 2026. Strumento concettuale.
   https://www.theguardian.com/film/2026/sep/27/naza-film-new-york-city-premiere

Scheda: https://claudiocammarano.com/curated/2026-09-27-naza-danno-collaterale-guardian/

7. Donatella Della Ratta / Le Grand Continent, *Sur la violence spéculative de l'IA*, 4 set 2026. Strumento concettuale.
https://legrandcontinent.eu/fr/2026/09/04/sur-la-violence-speculative-de-lia/
Scheda: https://claudiocammarano.com/curated/2026-09-04-dellaratta-violenza-speculativa-ia-grandcontinent/

8. OpenAI, *The Hugging Face incident and the road ahead*, 26 ago 2026. Riferimento tecnico.
https://openai.com/index/hugging-face-incident-and-the-road-ahead/
Scheda: https://claudiocammarano.com/curated/2026-08-26-openai-incidente-hugging-face-resoconto/

9. The Editorial Board / The New York Times, *A.I. Companies Say They Aren't Responsible for Their Unpredictable Products. Don't Believe Them.*, 10 ott 2026. Strumento concettuale.
https://www.nytimes.com/2026/10/10/opinion/ai-companies-liability-laws.html
Scheda: https://claudiocammarano.com/curated/2026-10-10-nyt-editorial-ai-responsabilita/

---

**Intervento AI: WR** (Scrittura). Notazione descritta in https://claudiocammarano.com/notazione/.

© Claudio Cammarano. Licenza CC BY-NC-ND 4.0: si può condividere citando l'autore e la fonte, non a fini commerciali e senza modifiche.

Versione originale e aggiornata: https://claudiocammarano.com/writings/2026-10-05-nessun-altro-nella-stanza/. Le parti interattive della catena sono disponibili solo online.

Questa catena fa parte della serie *L'attante zero*: https://claudiocammarano.com/serie/lattante-zero/.

Scrivo anche *The Abstract*, in inglese, una volta al mese, sull'economia dei contenuti: https://claudiocammarano.com/newsletter/. C'è anche il feed: https://claudiocammarano.com/feed.xml.