Intelligenza Apparente

Quanto aggiunge davvero l'intelligenza artificiale a chi deve prendere una decisione? Per misurarlo proponiamo una nuova unità, il turing: quante volte una persona riconosce meglio la risposta giusta quando lavora con una macchina accanto. L'abbiamo usata su oltre 150.000 giudizi umani e 24 modelli. Con GPT-4o, su domande d'esame, la sensibilità delle persone sale del 65 per cento (dal 58 al 77 per cento di risposte giuste), ma la coppia fa peggio della macchina da sola, che vale più del doppio delle persone da sole: la macchina aiuta le persone, le persone la frenano. Misuriamo anche quanto gli errori delle persone e quelli delle macchine sono allineati, cioè quanto cadono sugli stessi casi, dove un secondo parere non serve: in parte lo sono, più o meno secondo il compito. Una cosa lo strumento non può misurarla: chi ha scelto la domanda.

endjneer · paithon.it · 11 novembre 2026 · versione 2
Articolo: https://paithon.it/intelligenza-apparente/
Codice e dati: github.com/paithon-it/turing-meter · doi.org/10.5281/zenodo.24141588
Testo con licenza CC BY 4.0.

Chi vuole sapere quanta potenza assorbe un motore può fare due letture. Il voltmetro ai morsetti dà la tensione, l'amperometro sul filo dà la corrente, e il prodotto dà un numero esatto, in voltampere. È la *potenza apparente*, e serve a dimensionare i cavi. Il motore, però, di quel numero trasforma in lavoro solo una parte: tensione e corrente non vanno al passo, fra l'una e l'altra c'è uno sfasamento, e per sapere quanta di quella potenza diventa lavoro serve un terzo strumento, il *wattmetro*, che tiene conto dell'angolo. **Nessun conto fatto sulle prime due letture lo sostituisce.**

Nel 2025 METR, un'organizzazione che valuta le capacità dei modelli di frontiera, ha chiesto a sedici sviluppatori esperti di lavorare sui progetti open source che conoscevano meglio, con e senza gli strumenti di AI di inizio anno, su 246 compiti assegnati a sorte all'una o all'altra condizione. Prima di cominciare prevedevano che l'AI avrebbe tagliato i tempi del 24 per cento. Alla fine erano convinti di averli tagliati del 20. Il cronometro diceva il contrario: con l'AI ci avevano messo **il 19 per cento in più**.

Quegli sviluppatori avevano letto il voltmetro, e non sono i soli. Da qualche anno collezioniamo letture eccellenti: quanto è brava una macchina da sola nei test, quanto è brava una persona da sola, quanto si sente più brava dopo. Si legge di rado quanto fanno davvero due intelligenze messe nello stesso circuito. E quasi mai l'angolo fra i loro errori, che dice quanto potrebbero fare.

Questo articolo propone quella lettura. Ha un'unità di misura, il **turing**: quante volte la stessa persona riconosce meglio la risposta giusta con la macchina accanto, rispetto a quando è sola. E ha un angolo, che dice quanto gli errori della persona e quelli della macchina cadono sugli stessi casi: se cadono su casi diversi l'uno può correggere l'altro, se cadono sugli stessi il secondo parere non

serve. Uno strumento nuovo, prima di crederci, va puntato su qualcosa di vero, e va provato dove la risposta si conosce già. Abbiamo fatto tutte e due le cose.

I dati sono di altri, resi pubblici: oltre centocinquantamila giudizi di persone su domande d'esame, ragionamenti di senso comune, giudizi di sarcasmo e immagini, messi accanto alle risposte di 24 macchine, una parte delle quali fatte girare da noi su una GPU. Sotto la misura c'è un'idea filosofica, quella che dà il titolo all'articolo, e la misura la tocca solo in parte: lavora su domande che hanno già una risposta giusta, e chi ha scelto la domanda le resta fuori.

In breve

- **L'unità.** Una mente da sola vale un turing: ogni mente fa da campione a se stessa, e nella definizione non compare nessuna macchina. A 1,5 turing si decide con il 50 per cento di sensibilità in più. È un guadagno, come i decibel di un amplificatore, e due turing si confrontano bene a parità di persone e di domande.
- **La fiducia.** Su 1.350 esperimenti finti lo strumento ritrova, dentro il modello, il turing vero entro l'uno per cento in sei situazioni su sette, con intervalli più larghi del necessario, cioè prudenti.
- **Quanto vale una macchina.** Su domande d'esame, con GPT-4o, un gruppo di persone sale a 1,65 turing, ma la coppia resta sotto la macchina da sola. Con Llama 3.1 8B, più debole, sale a 1,40 e la coppia supera la macchina interrogata da sola, non quella che le ha risposto in chat.
- **L'angolo.** Quello fra gli errori di persona e macchina sta sotto i 90 gradi in mediana su tutti i dati, fra 59 e 81 secondo il compito, e in modo netto su αNLI e sul testo di HAIID. Le macchine si somigliano fra loro più di quanto somiglino a noi, e su un compito l'angolo si chiude man mano che la macchina cresce. Oltre i 90 gradi, dove gli errori di due menti si cancellano, i nostri dati arrivano solo in casi isolati: è la *risonanza cognitiva*, il tema dell'articolo che segue.
- **Il limite.** Lo strumento lavora su domande che hanno già una risposta giusta. Chi ha scelto la domanda resta fuori, ed è proprio la parte dell'intelligenza che dà il titolo all'articolo.

Volontà e tecnica

La parola «apparente» ha tre significati, e chi parla di intelligenza artificiale passa dall'uno all'altro senza accorgersene.

Il primo è quello dell'elettricista. In corrente alternata la potenza ha una parte *attiva*, che scalda e fa girare, e una parte *reattiva*, che va e viene fra la rete e i motori per magnetizzarli e non lascia lavoro netto, però occupa i cavi. Stanno ad angolo retto, la loro composizione è l'apparente, e il coseno dell'angolo, il **fattore di potenza**, dice quanta parte è lavoro. Un'impresa con un fattore troppo basso paga una penale in bolletta, e per rimediare installa condensatori: si chiama *rifasare*.

Il secondo è quello di Kant. Di una cosa conosciamo come ci appare, il *fenomeno*, e mai la cosa in sé. Vale per l'intelligenza, e vale prima di tutto per la nostra: di quella di un altro conosciamo ciò che fa e ciò che dice. Turing arrivò per la sua strada alla stessa conseguenza operativa quando, nel 1950, propose di lasciar perdere la domanda se le macchine pensino, e di chiedersi se sappiano passare per persone.

Il terzo è quello dell'illusionista. L'anatra meccanica che Jacques de Vaucanson espose nella Francia del Settecento mangiava grano e, dopo un po', lo evacuava: dentro non digeriva niente, le scorie erano già pronte in un altro scomparto. ELIZA, il programma che Joseph Weizenbaum scrisse a metà degli anni Sessanta, rimandava indietro le frasi di chi scriveva, voltate in domanda. Weizenbaum raccontò che la sua segretaria, che lo aveva visto scriverlo, gli chiese di uscire dalla stanza per restare a tu per tu con il programma.

Tre sensi, e un punto in cui si toccano: ciò che si presenta allo strumento o all'osservatore è sempre qualcosa di diverso da ciò che lavora. C'è poi una famiglia di grandezze che porta lo stesso aggettivo, le forze apparenti. La spinta che vi schiaccia di lato in curva si misura, fa danni, eppure sparisce se la si guarda da fuori dell'auto. Le forze apparenti sono reali *e* relative all'osservatore.

Da questi sensi viene l'idea che dà il titolo all'articolo. L'intelligenza che si vede in un'opera si può pensare come la composizione di due parti ad angolo retto, come la potenza attiva e quella reattiva: la *volontà*, ciò che origina e dirige, e la *tecnica*, ciò che esegue. La loro composizione è l'*intelligenza apparente*. In questo quadro la macchina sta tutta sull'asse della tecnica, e l'angolo retto non sarebbe un caso: chi progetta qualcosa a propria somiglianza ottiene comunque qualcosa di diverso, anche quando voleva una copia. Da qui verrebbe il fraintendimento dell'epoca, che scambia l'oscillazione per il lavoro: l'intelligenza delle macchine sarebbe vera, e la misureremmo ai morsetti sbagliati.

Il punto più vecchio di questa discussione è del 1843. Traducendo un articolo sulla Macchina Analitica di Babbage, Ada Lovelace aggiunse delle note più lunghe del testo, e in una scrisse: «La Macchina Analitica non ha alcuna pretesa di originare nulla. Può fare qualunque cosa noi sappiamo ordinarle di fare». Due verbi che dividono il lavoro come lo divide l'ipotesi, a noi *originare* e *ordinare*, alla macchina eseguire, con centottant'anni di anticipo.

Turing, nel 1950, dedicò a questa frase una sezione del suo articolo, e la chiamò l'obiezione di Lady Lovelace. Una variante dell'obiezione, annotò, dice che una macchina non può mai prenderci di sorpresa. La sua risposta è una delle frasi più disarmanti della storia dell'informatica: «Le macchine mi sorprendono con grande frequenza. In gran parte perché non faccio abbastanza calcoli per decidere che cosa aspettarmi da loro, o meglio perché, pur facendo un calcolo, lo faccio in modo frettoloso e approssimativo, correndo dei rischi».

Di solito si dà per scontato che qui Turing abbia vinto. A rileggerlo, ha spostato il peso: dalla macchina a chi la guarda. La frase dice che la sorpresa è il residuo di un conto non fatto. Turing non lo diceva per sgonfiarla: poche righe dopo chiama fallacia l'idea che, appena un fatto si presenta a una mente, tutte le sue conseguenze vi sorgano insieme. Il conto non fatto è lavoro vero. Ma è lavoro dell'osservatore, e la sorpresa è una grandezza di chi guarda, come una forza apparente. La sorpresa ha la stessa forma del giudizio dei sedici sviluppatori, che non avevano cronometrato niente.

Le due letture della sorpresa, una macchina che origina qualcosa o un osservatore che non ha fatto il conto, restano in piedi tutte e due, e per scegliere bisogna, appunto, fare dei conti. Su volontà e tecnica, però, nessuno strumento ha ancora una scala. Esiste una traduzione più umile, ed è nostra: due intelligenze sono ortogonali se **sbagliano su casi diversi**. Ha un prezzo, e conviene dirlo subito. Le domande dei dati pubblici hanno una risposta giusta già scritta da qualcuno, e su domande così la persona e la macchina lavorano tutte e due di tecnica: lo strumento legge l'angolo fra due tecniche, e la volontà gli resta fuori. Così ristretta, «intelligenza apparente» vuol dire una cosa sola, che si può misurare: ciò che due intelligenze messe insieme promettono.

Finché resta un'analogia un'idea così non può sbagliare, e un'idea che non può sbagliare non dice niente. Il bello di una metafora elettrica è che porta con sé uno strumento di misura, e bisogna avere il coraggio di usarlo.

Due correttori di bozze

Due correttori rileggono lo stesso libro. Se si lasciano sfuggire tutti e due gli stessi refusi, il secondo non serve a niente: il libro esce con gli errori del primo. Se ciascuno manca refusi diversi, insieme ne trovano più di quanti ne troverebbe il migliore dei due. In una coppia non conta solo quanto è bravo ciascuno: conta se i due sbagliano sugli stessi casi.

La *teoria della rilevazione del segnale*, nata con i radar e diventata il modo corrente di misurare la percezione, dà a questa intuizione una forma esatta. Chi decide fra alternative ha in testa un indizio graduato, disturbato dal rumore, e alla fine ne fa uscire un verdetto: questa, o quella. La sua bravura si misura con un numero, la *sensibilità* d', che qui non è quella dei test medici: dice di quanto il segnale si stacca dal rumore, cioè quanto bene chi decide distingue il giusto dallo sbagliato. Vale zero per chi tira a indovinare e non ha un tetto. Se una persona e una macchina hanno sensibilità d_H e d_M sugli stessi casi, e chiamiamo ρ la correlazione fra i rumori dei loro indizi, ρ dice quanto tendono a sbagliare insieme. L'angolo fra i due è

$$\varphi = \arccos \rho$$

e la coppia ideale, quella che mette in comune tutti gli indizi e li pesa come andrebbero pesati, arriva a una sensibilità che ha una bella interpretazione geometrica. Disegniamo i due assi a un angolo φ. C'è un solo vettore S le cui proiezioni ortogonali sui due assi sono esattamente d_H e d_M, e la sua lunghezza è ciò che la coppia ideale può fare:

$$|S|^2 = \frac{d_H^2 + d_M^2 - 2\, d_H d_M \cos\varphi}{\sin^2 \varphi}$$

Chi ricorda la fisica del liceo si aspetterebbe di sommare le due frecce, una in coda all'altra, e con quella regola due frecce che si avvicinano danno una somma più lunga. Qui succede il contrario, ed è giusto così, perché d_H e d_M non sono due pezzi da mettere in fila. Sono le ombre che uno stesso vettore proietta su due assi, e S è l'oggetto che le getta. Se i due correttori condividono una parte del rumore, una parte della seconda lettura ripete ciò che la prima ha già detto, e non va contata due volte: più i due sbagliano insieme, più si accorcia, fino a un punto preciso, il vettore che spiega tutte e due le ombre.

A 90 gradi il coseno si annulla, resta $|S|^2 = d_H^2 + d_M^2$, ed è Pitagora (figura 1): la legge dell'angolo retto, letta come ideale di una coppia, è il caso particolare in cui i due sbagliano in modo indipendente. Chi pensa al motore e al suo wattmetro deve rovesciare una cosa sola. Per l'elettricista un angolo piccolo vuol dire lavoro; per due correttori vuol dire che il secondo ripete il primo, e più l'angolo si apre, più la coppia promette. La figura usa i numeri veri di un gruppo di persone e di GPT-4o su domande d'esame.

> **I precedenti.** L'ideale di un gruppo di osservatori l'hanno scritto Sorkin e Dai nel 1994. Marc Ernst e Martin Banks hanno mostrato nel 2002 che il cervello combina quasi così la vista e il tatto; Bahador Bahrami e colleghi, nel 2010, che due persone che discutono ci arrivano solo se le loro sensibilità sono simili, e misuravano già il beneficio della coppia come rapporto con il migliore dei due. La correlazione nascosta fra persone e reti l'hanno stimata Steyvers e colleghi nel 2022, la sovrapposizione degli indizi su ChatBench Kong e colleghi quest'anno.

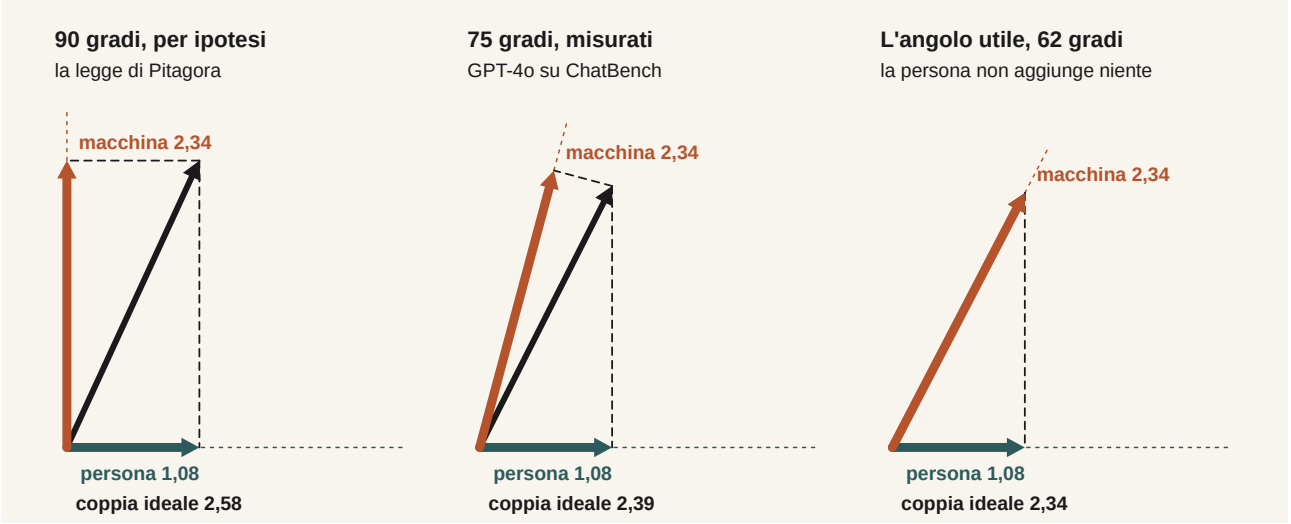


Figura 1. La geometria: le proiezioni di S sui due assi sono le due sensibilità.

Quando l'angolo si stringe il vettore si accorcia, e c'è un punto preciso in cui succede una cosa interessante. Se $\cos\varphi = d_H / d_M$, il vettore S si sovrappone all'asse del più forte: l'ideale della coppia è la macchina da sola. Chiamiamo quel valore **angolo utile**. Sotto quell'angolo la persona può ancora servire, ma in un modo solo: con un peso negativo, cioè usando i propri errori per sottrarre alla macchina il rumore che ha in comune con lei. Una persona non rende questo servizio di proposito, e qui l'ideale non lo conta: si ferma al più forte dei due, per scelta. Con questa scelta l'angolo utile dice *di quanto due intelligenze devono differire perché la più debole aggiunga qualcosa del suo*, e dipende soltanto da quante volte la macchina è più sensibile (tabella 1).

Tabella 1. L'angolo utile
l'angolo minimo fra gli errori perché la persona aggiunga qualcosa all'ideale della coppia, a seconda di quante volte la macchina è più sensibile

La macchina è più sensibile della persona	Angolo utile
2 volte	60 gradi
3 volte	70,5 gradi
5 volte	78,5 gradi
10 volte	84,3 gradi

Più la macchina è brava, più la persona deve restare vicina ai 90 gradi per contare: il suo valore si sposta dalla bravura alla differenza. Tutto questo, va detto, vale dentro un modello: indizi gaussiani, messi insieme in modo lineare, con pesi non negativi. L'ideale è un tetto dentro quel modello. Una persona che sa *quando* la macchina sbaglia, e caso per caso decide se ascoltarla, fa una cosa che il modello non contempla, e può superarlo. Chiamiamo *gating* questa capacità, come il cancello che lascia passare ora un segnale ora l'altro.

> **Che cosa finisce nell'angolo.** ρ è una correlazione fra indizi nascosti, ricavata dalle risposte, a parità di risposta giusta. Ci finiscono dentro due cose che lo strumento non separa: quanto i due ragionano in modo simile, e quanto i casi sono difficili per tutti. Due correttori del tutto indipendenti, davanti a refusi visibilissimi e a refusi invisibili, sbagliano insieme sugli invisibili. Per la coppia fa lo stesso: dove un caso è impossibile per tutti e due, il secondo parere non serve. Per capire *perché* due intelligenze si somigliano, invece, fa differenza.

Il turing

Con tre sensibilità prese sulle stesse decisioni (la persona da sola, la macchina da sola, la persona con la macchina, che chiamiamo d_S) si può dire quanto vale una coppia. Il **turing**, simbolo Tu, è il rapporto fra la sensibilità di una persona con la macchina e quella della stessa persona da sola, sugli stessi casi, $\mathrm{Tu} = d_S / d_H$. Una mente che lavora da sola vale un turing per definizione: ogni mente fa da campione a se stessa. È un numero puro, come i valori «per unità» degli elettrotecnici, che esprimono le grandezze di una macchina in frazione di una base scelta per quella macchina. Ed è un *guadagno*, come quello di un amplificatore, che si dà rispetto al segnale in ingresso: dice di quanto il circuito moltiplica la mente che ci si mette dentro, e non quanto quella mente valga in assoluto. A 1,4 turing una persona, con la macchina accanto, decide con il 40 per cento di sensibilità in più; sotto uno la macchina la appesantisce.

Il quadrato del turing ha un significato suo, e spiega il nome. A Bletchley Park, contro l'Enigma della marina tedesca, con un procedimento che chiamavano *Banburismus*, si sommavano uno dopo l'altro molti indizi piccoli, e Turing aveva dato un nome all'unità con cui pesarli: il *ban*, con il suo decimo, il *deciban*. Lo racconta Irving John Good, che lavorò con lui. Turing trovò anche che, se il *peso dell'evidenza* a favore di un'ipotesi vera ha una distribuzione normale, la sua varianza, misurata in ban naturali, è il doppio della sua media. Con i simboli di questa pagina vuol dire che d'^2 è il doppio del peso che un'osservazione porta in media; i pesi di indizi indipendenti si sommano, e dunque si sommano i quadrati delle sensibilità. Il teorema di Pitagora della coppia è la somma dei deciban di Bletchley, letta in media. Così Tu^2 dice quante copie della persona, con errori indipendenti e indizi messi in comune, servirebbero per fare quanto la coppia. A 1,4 turing ne servirebbero circa due. Sono copie ipotetiche: due persone vere sbagliano insieme più spesso, e ne servirebbero di più.

> **Da dove viene d'.** Good attribuisce la prima pubblicazione del risultato ai teorici del radar dell'Università del Michigan: a Birdsall (1955) in un racconto del 1979, a Peterson, Birdsall e Fox (1954) in uno del 1985. Nel rapporto tecnico di Peterson e Birdsall del 1953 si legge già che il logaritmo del rapporto di verosimiglianza, se normale, ha varianza d'^2 e media $d'^2/2$. Da quella teoria viene d': Tanner e Birdsall, nel 1958, ne fissarono la definizione con l'energia che servirebbe a un ricevitore ideale per ottenere la stessa prestazione.

Nella definizione non compare nessuna macchina, né un test, né un anno: compaiono una mente e ciò che riesce a fare con qualcosa accanto. È ciò che permette alla definizione di non scadere insieme ai modelli. I valori, invece, restano legati alle persone e ai casi: un regolo calcolatore, un classificatore del 2020 e un modello linguistico del 2026 si confrontano solo sulle stesse persone e sugli stessi casi. E c'è una distinzione che pesa su tutti i numeri. La definizione vale per una mente; i turing di queste pagine sono di un gruppo, perché vengono dalle risposte di tutte le persone messe insieme, come negli studi di radiologia che confrontano due modi di leggere le stesse immagini con molti lettori e molti casi. Un gruppo, con una macchina, vale tanti turing; quanto le persone, una per una, differiscano fra loro, i dati non lo dicono.

Un turing, da solo, non dice chi ha fatto il lavoro. Chi ricopia ogni risposta di una macchina k volte più sensibile vale k turing senza averci messo niente, e per questo un turing si legge accanto a $k = d_M / d_H$, quanto vale la macchina da sola nell'unità della persona. Il confronto dà tre zone. Sotto uno, la macchina appesantisce la persona. Fra uno e k, la macchina aiuta la persona e la persona frena la macchina: la coppia decide meglio di lei e peggio della macchina da sola. Sopra

tutti e due, uno e k, la coppia batte entrambi, ed è il solo caso in cui si può parlare di collaborazione. Il rapporto $d_S/d_M = \mathrm{Tu}/k$ dice in quale zona si sta, e lo chiamiamo **resa sulla macchina**: non dice quanto vale la macchina, che è k, ma quanta della sua sensibilità resta alla coppia. A 0,8 la coppia decide con l'80 per cento della sensibilità che la macchina aveva da sola; sopra l'uno, la persona le aggiunge qualcosa. Dove gli intervalli comprendono l'uno la zona non si assegna, e con una macchina più debole delle persone, k sotto uno, le parti si scambiano.

Il secondo numero viene dal quadro elettrico. Le due letture separate, d_H e d_M, composte con il loro angolo danno $|S|$: è l'**intelligenza apparente** della coppia, ciò che le due letture promettono, come i voltampere, con una differenza: qui l'angolo entra già nell'apparente, perché due sensibilità non si compongono senza. La sensibilità che la coppia raggiunge davvero, d_S, è la parte attiva, come i watt. Divise per d_H si leggono tutte e due nell'unità della persona: $|S|/d_H$ sono i *turing apparenti*, d_S/d_H i turing. Il loro rapporto è il *fattore di intelligenza*, $\eta = d_S/|S|$, quanto della promessa diventa decisione. Somiglia al fattore di potenza con una differenza che conta: un motore non dà mai più watt dei voltampere, mentre una coppia può superare l'ideale del modello, e η può passare l'uno.

I due numeri si tengono, $\mathrm{Tu} = \eta \cdot \mathrm{Tu}_{\mathrm{app}}$, e i turing apparenti dipendono soltanto da k e dall'angolo:

$$\mathrm{Tu}_{\mathrm{app}} = \frac{\sqrt{1 + k^2 - 2k\cos\varphi}}{\sin\varphi}$$

A 90 gradi resta $\sqrt{1+k^2}$; sotto l'angolo utile ci si ferma a k. Una macchina più forte alza i turing apparenti anche se le persone non cambiano niente di come lavorano.

> **La corrispondenza elettrica.** Voltmetro e amperometro: la persona da sola e la macchina da sola. Lo sfasamento: l'angolo fra i loro errori. Voltampere: i turing apparenti. Watt: i turing. Fattore di potenza: η.

Un esempio fissa le idee, ed è quello della figura 1: un gruppo di persone e GPT-4o, su domande d'esame a quattro opzioni (tabella 2).

Tabella 2. Una coppia in cifre
ChatBench, tutte le prove; la macchina è quella interrogata da sola

	Valore	Che cosa dice
Persone da sole	58% · d_H = 1,08	Risposte giuste e sensibilità del gruppo
GPT-4o da sola	88% · d_M = 2,34	La macchina vale k = 2,16 volte le persone
Persone con GPT-4o	77% · d_S = 1,78	Ciò che il gruppo fa con la macchina accanto
Angolo fra gli errori	75 gradi	Sopra l'angolo utile, che qui è di 62
Turing apparenti	2,21	Ciò che la coppia promette
Turing	1,65	Quanto sale il gruppo con la macchina
Fattore, η	0,75	Quanto della promessa diventa decisione
Resa sulla macchina	0,76	Sotto l'uno: la coppia resta sotto la macchina da sola

È la zona di mezzo: la macchina aiuta le persone, e le persone frenano la macchina.

I numeri di queste pagine vengono da quattro raccolte pubbliche, che hanno le risposte caso per caso delle persone e delle macchine: *ChatBench*, domande d'esame di matematica, fisica ed etica a cui centinaia di persone hanno risposto da sole e poi parlando con GPT-4o o con Llama 3.1 8B; *ChaosNLI*, nella parte chiamata αNLI, storie a cui manca una frase, con cento voti umani per caso; *HAIID*, cinque compiti brevi in cui le persone ricevevano un consiglio; le immagini degradate di Geirhos e colleghi. Le macchine che mancavano le abbiamo fatte girare noi.

Quanto vale un turing

Cominciamo dal numero che l'unità promette. Su tutte le prove di ChatBench un gruppo di persone, con GPT-4o, vale **1,65 turing** (intervallo al 95 per cento da 1,44 a 1,91); con Llama 3.1 8B, un modello molto più piccolo, 1,40 (da 1,24 a 1,62). Sono gruppi di persone diversi, assegnati a caso all'una o all'altra macchina: il confronto regge per il sorteggio, non perché le persone siano le stesse (tabella 3). E partono da punti diversi: da solo, il gruppo di GPT-4o ha una sensibilità di 1,08, quello di Llama di 1,17. Con la macchina arrivano a 1,78 e a 1,64, l'8 per cento di distanza, che in turing diventa il 18: metà dello scarto viene dai denominatori. Il turing è una definizione, e una definizione si adotta; i numeri, invece, vanno confermati su dati nuovi, perché queste analisi sono venute dopo i primi risultati e il protocollo non le prevedeva.

Tabella 3. Due macchine, due zone
ChatBench, tutte le prove; la macchina è quella interrogata da sola. Intervalli al 95 per cento, ricampionando persone e casi

Coppia	Persone da sole (d_H)	La macchina vale (k)	Angolo	Turing apparenti	Turing	Resa sulla macchina	Zona
Persone con GPT-4o	1,08	2,16	75	2,21	1,65 (1,44-1,91)	0,76 (0,67-0,85)	Le persone frenano la macchina
Persone con Llama 3.1 8B	1,17	1,18	75	1,38	1,40 (1,24-1,62)	1,19 (1,04-1,34)	La coppia batte tutti e due (la macchina interrogata da sola)

Le due righe stanno in zone diverse. Il lato più solido è quello di GPT-4o: da sola valeva 2,16 volte le persone, il gruppo è salito a 1,65, ma **la coppia è rimasta sotto la macchina da sola**, a 0,76 della sua sensibilità, con un intervallo tutto sotto l'uno, e ci resta anche contro la macchina che le ha risposto in chat. Llama valeva poco più delle persone, 1,18 volte, e la coppia ha superato la macchina interrogata da sola, ma non quella che le ha risposto in chat, che era più brava: il sorpasso dipende da dove si attacca lo strumento. L'angolo è lo stesso, circa 75 gradi, e in tutti e due i casi sta sopra l'angolo utile (62 gradi con GPT-4o, 32 con Llama). Cambia il margine: con GPT-4o la coppia ideale promette appena il 2 per cento più della macchina da sola, con Llama il 17. Più la macchina è forte, meno spazio resta alla persona, e basta perdere poco per finire sotto. Agli sviluppatori di METR, sul tempo, era andata peggio: con l'AI erano finiti sotto se stessi. La macchina potenzia davvero le persone, e accanto a una macchina forte una parte di quella potenza le persone la perdono per strada.

Lo stesso modello, poi, non vale gli stessi turing dappertutto. Basta dividere le domande per materia (tabella 4).

Tabella 4. Il turing per materia
ChatBench, tutte le prove; la macchina è quella interrogata da sola, e «la macchina vale» dice quante volte la macchina è più sensibile delle persone. Fra parentesi l'intervallo al 95 per cento, in colore dove esclude l'uno

Coppia	La macchina vale	Turing	Resa sulla macchina	Zona
GPT-4o, matematica	2,33	1,70 (1,44-2,07)	0,73 (0,63-0,83)	Le persone frenano la macchina
GPT-4o, fisica	3,18	2,11 (1,52-3,43)	0,66 (0,44-0,86)	Le persone frenano la macchina
GPT-4o, etica	1,08	1,08 (0,85-1,43)	1,00 (0,66-1,47)	Non si distingue
Llama 3.1 8B, matematica	1,40	1,44 (1,25-1,73)	1,03 (0,90-1,16)	Le persone salgono, la macchina resta pari
Llama 3.1 8B, fisica	1,85	2,39 (1,60-5,17)	1,29 (0,95-1,81)	Le persone salgono, la macchina resta pari
Llama 3.1 8B, etica	0,49	0,91 (0,74-1,11)	1,86 (1,33-2,75)	Le persone salvano la macchina

In matematica e in fisica le persone con GPT-4o guadagnano molto, e frenano la macchina: la resa sulla macchina sta sotto l'uno. Sulle domande di etica, dove le persone da sole sono già sensibili quanto GPT-4o, tutti i numeri stanno intorno all'uno. Con Llama l'etica rovescia la coppia: lì la macchina da sola vale meno delle persone, 0,49, il gruppo non guadagna niente che si possa vedere, e la coppia vale 1,86 volte la macchina, ma solo perché le persone da sole valevano il doppio di lei. **Il turing è della coppia su un compito**: la macchina, da sola, non ne ha. Gli intervalli per materia sono larghi, per una ragione che appartiene all'unità: le persone per materia sono poche, da sole hanno una sensibilità bassa, e un rapporto con un denominatore piccolo e incerto oscilla molto.

La stessa unità legge macchine di altri anni. Nel 2024 una meta-analisi del *Center for Collective Intelligence* del MIT, firmata da Michelle Vaccaro, Abdullah Almaatouq e Thomas Malone, ha raccolto 370 confronti fra persone da sole, AI da sola e persone con l'AI, e ha trovato che in media la coppia faceva peggio del migliore dei due, con perdite nelle decisioni e guadagni nella creazione. Ne abbiamo tenuto le decisioni fra due alternative, le sole per cui si può calcolare una sensibilità: **87 confronti da 27 studi**, usciti fra il 2020 e la metà del 2023, con macchine che l'archivio degli autori etichetta come *deep*, *shallow* o *Wizard of Oz* (queste ultime simulate dagli sperimentatori). Il turing mediano è 1,24 (intervallo da 1,18 a 1,68, ricampionando gli studi), e in 12 confronti su 87 sta sotto l'uno. In 61 di quei confronti, però, le persone da sole e quelle con la macchina sono gruppi diversi, e il turing vuole le stesse persone. Nei 26 in cui lo sono la mediana è 1,34, e sotto l'uno ne sta uno solo.

L'unità funziona anche dove la macchina non c'è. Su HAIID il consiglio era la media di una cinquantina di persone, e chi lo riceveva arrivava a un turing fra 1,16 e 1,59 secondo il compito, con un intervallo che esclude l'uno in tutti e cinque i compiti. L'unità misura ciò che un aiuto aggiunge a una mente, qualunque cosa sia l'aiuto.

La stessa unità attraversa gli anni, se si tiene ferma la mente. Fra le dieci reti per la visione, uscite fra il 2014 e il 2023, prendiamo la più vecchia, la rete VGG16, come mente, e mettiamole accanto le reti venute dopo (tabella 5), una alla volta. Un *combinatore*, una regressione logistica che pesa le loro probabilità, cioè tutti gli indizi e non un verdetto, impara a usarle su immagini che poi non rivede. Una condizione è un tipo e un grado di degrado delle immagini.

Tabella 5. Una mente del 2014 accanto alle macchine venute dopo
la rete VGG16 come mente, nove reti come macchine, sulle immagini date anche alle persone; mediane sulle condizioni valide. «Macchina su mente» dice quante volte la macchina è più sensibile. L'ultima colonna conta le condizioni in cui la coppia fa meglio e quelle in cui fa peggio della macchina da sola

Macchina	Anno	Macchina su mente	Angolo	Turing apparenti	Turing	Meglio / peggio della macchina
ResNet-50	2015	1,09	31	1,14	1,09	34 / 17
ResNet-152	2015	1,18	37	1,19	1,19	26 / 9
EfficientNet-B4	2019	1,44	42	1,44	1,44	15 / 8
ViT-B/16	2021	1,60	48	1,61	1,56	11 / 13
Swin-B	2021	1,63	45	1,64	1,62	8 / 17
ConvNeXt-B	2022	1,58	41	1,63	1,57	6 / 15
ConvNeXt V2-B	2023	1,67	48	1,70	1,59	9 / 16
EVA-02 B	2023	1,87	47	1,93	1,88	9 / 13
EVA-02 L	2023	2,14	44	2,18	2,18	7 / 7

Il turing di una rete del 2014 sale da 1,09 accanto a una rete del 2015 a 2,18 accanto a una del 2023: nove anni di macchine letti sulla stessa scala. E quasi tutto ciò che la scala mostra è sostituzione, perché il turing segue k. I conteggi dell'ultima colonna, su condizioni da 160 immagini e con molti pareggi, non hanno un intervallo e vanno presi come indicazione. Quando poi le immagini peggiorano fino a portare VGG16 vicino al caso, il turing di VGG16 accanto a EVA-02 L sale senza che nessuno collabori di più: togliendo contrasto passa da 1,28 a 6,42, perché è il denominatore a scendere. Per questo un turing si legge accanto a k, e non si legge affatto dove chi fa da campione non vede più niente.

Un numero di turing va dato con ciò che lo rende leggibile: il compito, le persone e la loro sensibilità da sole, la macchina e la sua k, il circuito, cioè il modo in cui la si usa, le prove incluse, e il suo intervallo. La sensibilità da sole serve perché il campione può spostarsi: un turing sale anche quando è la persona, da sola, a scendere.

I morsetti sbagliati

ChatBench registra due macchine per ogni domanda: il modello interrogato da solo, e la risposta che lo stesso modello ha dato a ciascuna persona nella sua conversazione. Non sono la stessa macchina, e la differenza insegna due cose su dove si attacca lo strumento (tabella 6).

Tabella 6. La stessa coppia, misurata a due morsetti
ChatBench, tutte le materie; accuratezza della macchina presa da sola e nella conversazione con la persona. Il confronto vale fra le righe «da sola, stesse conversazioni» e «nella conversazione». Nella conversazione la macchina ha già dentro una parte della persona, quindi le righe «nella conversazione» sono descrizioni, e contano solo le risposte che sono una delle quattro opzioni. Intervalli al 95 per cento per l'angolo e per la resa

	Accuratezza macchina	Angolo	Turing apparenti	Fattore, η	Turing	Resa sulla macchina
GPT-4o, da sola, tutte le prove	88%	75 (59-86)	2,21	0,75	1,65	0,76 (0,67-0,85)
GPT-4o, da sola, stesse conversazioni	91%	79 (61-95)	2,24	0,85	1,90	0,88 (0,78-0,99)
GPT-4o, nella conversazione	92%	81 (60-101)	2,30	0,82	1,90	0,86 (0,77-0,95)
Llama 3.1 8B, da sola, tutte le prove	67%	75 (62-82)	1,38	1,01	1,40	1,19 (1,04-1,34)
Llama 3.1 8B, da sola, stesse conversazioni	73%	74 (58-85)	1,43	1,08	1,54	1,22 (1,08-1,39)
Llama 3.1 8B, nella conversazione	84%	63 (39-79)	1,69	0,91	1,54	0,92 (0,84-1,02)

La prima riguarda le prove che si contano. La chat ha registrato una risposta della macchina riconducibile a una delle quattro opzioni solo in una parte delle conversazioni, e sono quelle andate meglio. Lì il turing di GPT-4o sale a 1,90 e quello di Llama a 1,54; nelle altre, 372 prove con GPT-4o e 432 con Llama, è di 1,03 e 1,11, con intervalli che comprendono l'uno. **Contare solo le conversazioni riuscite gonfia il turing**, e per questo il valore di riferimento è quello su tutte le prove.

Le altre, però, dicono anche dove si perde la potenza. Con GPT-4o sono un quarto delle prove, e la tabella 7 mette tutte le prove in tre righe, secondo che cosa è uscito dalla chat.

Tabella 7. Dove si perde la potenza
ChatBench, GPT-4o, tutte le prove divise secondo la risposta uscita dalla chat; risposte giuste, in percentuale

Dalla chat è uscita	Prove	Persone da sole	GPT-4o da sola	Persone con GPT-4o
Una delle quattro opzioni	1.121	61	91	86
Una risposta fuori dalle opzioni	42	40	56	48
Nessuna risposta	330	50	84	50

Dove la chat ha dato una risposta fuori dalle opzioni la domanda era difficile anche per la macchina. Dove non ne ha data nessuna, invece, la macchina interrogata da sola rispondeva giusto quattro volte su cinque, e le persone con lei accanto sono rimaste dov'erano da sole. Non è solo svogliatezza. ChatBench ha fatto annotare le conversazioni da un modello linguistico: in 222 di quelle 330 prove la persona, secondo l'annotazione, aveva provato a usare la chat, e lì le cifre sono le stesse. La svogliatezza annotata, però, qui è quattro volte più frequente che nell'insieme delle prove. La macchina sapeva la risposta, e **dalla chat non è uscita**.

Prima del gating, decidere se credere alla macchina, c'è dunque l'estrazione: farsi dare la risposta. Nel quadro elettrico è un falso contatto ai morsetti: qualcosa passa, e non fa lavoro. In quelle

prove quasi una persona su tre cambia risposta dopo la chat, ma tante passano dallo sbagliato al giusto quante dal giusto allo sbagliato, 8 su cento per parte; dove la chat ha risposto, i passaggi al giusto sono otto volte quelli allo sbagliato. Perché la risposta non esca non l'abbiamo analizzato: le conversazioni sono nei dati e si possono leggere, e qui non l'abbiamo fatto. La persona può aver chiesto un pezzo della domanda e non tutta, essersi fermata prima, oppure la macchina può aver spiegato senza concludere. E sono prove scelte dopo, da come è andata la conversazione: il confronto descrive, non prova una causa. Con Llama il falso contatto non si vede, perché su quelle domande Llama da sola non ne sapeva più delle persone (54 contro 55).

La seconda riguarda la macchina. Prendiamo Llama 3.1 8B, sulle conversazioni con una risposta in chat. Da sola azzecca il 73 per cento delle domande; nella conversazione, con la persona che chiede, riformula e insiste, l'84. Il salto resta, più corto, anche contando come sbagliate le risposte fuori dalle opzioni. Rispetto alla macchina da sola la coppia sta sopra, a 1,22 volte la sua sensibilità; rispetto alla macchina della conversazione sta a 0,92, con un intervallo che tocca l'uno. Nessuno dei due è il morsetto giusto. La macchina della conversazione ha già dentro una parte della persona: in ChatBench chi rispondeva prima portava la sua risposta nella chat, e la macchina può averla assecondata. Una parte di ciò che si chiama collaborazione avviene dentro la macchina mentre ci si parla, e lo strumento la mette nel conto della persona o della macchina a seconda di dove lo si attacca. Il turing delle persone non cambia, 1,54 in tutti e due i casi; cambiano i turing apparenti e il fattore, che in chat vanno presi come descrizioni e non come la quota di un ideale.

Con GPT-4o la distanza è di un punto, 91 per cento contro 92, e la coppia resta sotto tutte e due le macchine, a 0,88 e a 0,86, con intervalli tutti sotto l'uno.

Resta la mansione che in questo quadro spetta alla persona, quella che uno strumento fuori dal modello potrebbe vedere: il gating, decidere quando affidarsi e quando no. Il solo confronto randomizzato che abbiamo su questo è piccolo. ChatBench assegnava a caso se la persona dovesse rispondere prima da sola o andare subito dalla macchina, e l'accuratezza finale non cambia in modo che si possa vedere: −2,0 punti con GPT-4o, +3,5 con Llama (chi risponde prima, meno chi va subito dalla macchina), con intervalli che comprendono lo zero. Non è una smentita: chi risponde prima sa solo di aver risposto prima, e nessuno gli ha insegnato quando credere alla macchina. **L'ordine delle mosse, da solo, non si vede.** Se il circuito conta, conta con qualcosa di più di un cambio d'ordine: quello che la macchina dice di sé, come si riconosce un suo sbaglio, chi ha l'ultima parola e quante volte può tornare indietro.

L'angolo misurato

Sulle quattro fonti che hanno le risposte caso per caso, persone e macchine sugli stessi casi, l'angolo si misura. La figura 2 mostra dove cade, la tabella 8 quanto è certo.

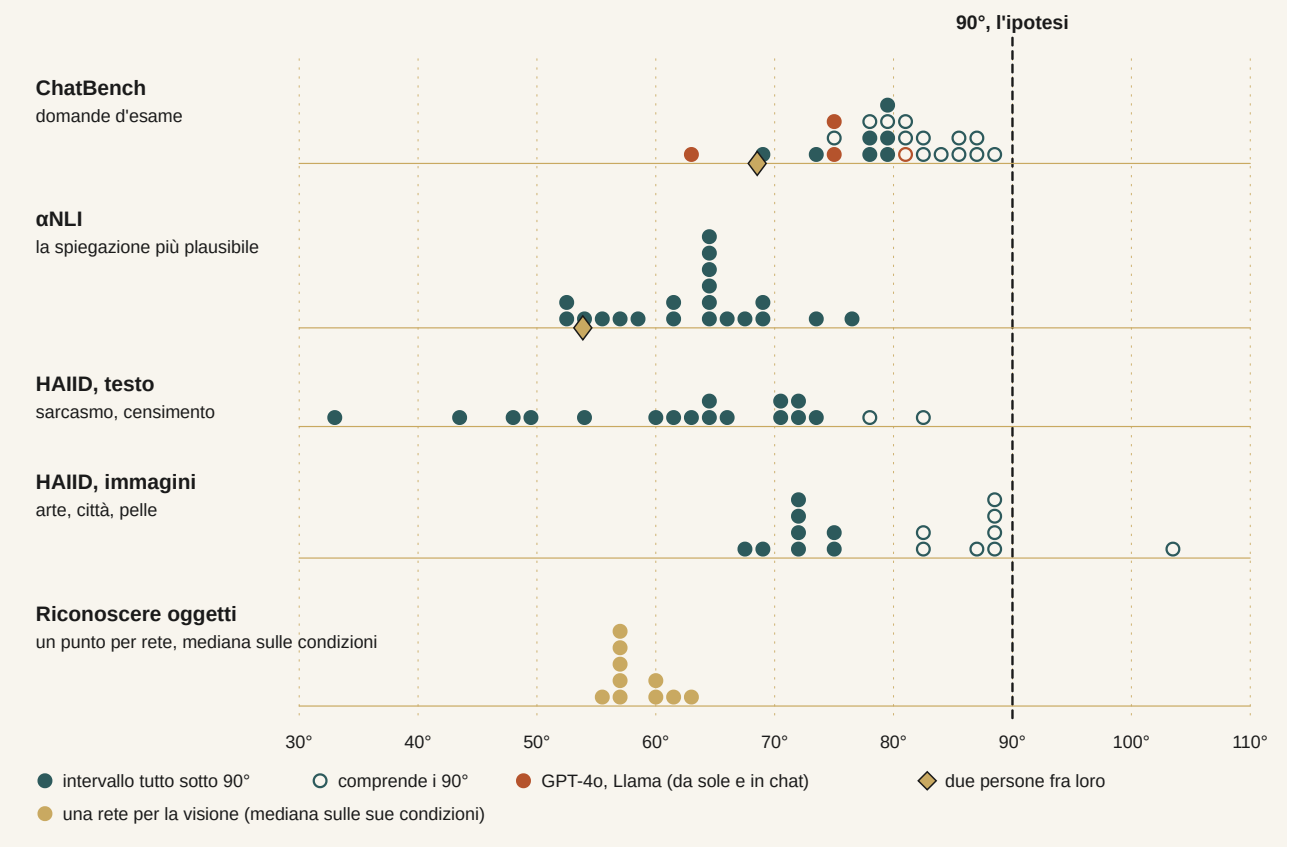


Figura 2. Dove cade l'angolo, per famiglia di dati, rispetto ai 90 gradi.

Tabella 8. L'angolo misurato fra persona e macchina
gradi; mediana ed estremi sui confronti di ogni famiglia di dati; nell'ultima colonna, in quanti l'intervallo al 95 per cento resta tutto sotto i 90

Dati	Confronti	Angolo mediano	Da … a	Intervallo tutto sotto i 90
ChatBench	20	**81**	69-88	7
αNLI	20	**64**	52-76	20
HAIID, testo	18	**65**	32-82	16
HAIID, immagini	16	**78**	68-103	8
Riconoscere oggetti (Geirhos)	689	**59**	56-62 per rete	374

L'ipotesi, decisa prima, era che l'angolo non fosse retto, e il criterio era che l'intervallo al 95 per cento restasse tutto sotto i 90 gradi. **Su αNLI passa in 20 confronti su 20**, sul testo di HAIID in 16 su 18. Su ChatBench (7 su 20) e sulle immagini di HAIID (8 su 16) la mediana sta sotto i 90, ma l'intervallo li comprende in metà dei confronti o più: lì resta una tendenza. Su ChatBench l'angolo è davvero vicino ai 90, e gli intervalli sono più larghi del necessario, quindi il conteggio è prudente; sulle immagini di HAIID i casi sono pochi. Il criterio deciso prima chiedeva l'angolo sotto i 90 gradi su ogni famiglia di dati, e così com'era scritto non passa: in modo netto il risultato vale su due famiglie di dati su quattro, αNLI e il testo di HAIID; gli oggetti di Geirhos fanno storia a sé.

Il punto più importante, però, è che **l'angolo dipende dal compito**, e non solo da quello: le fonti differiscono anche per il tipo di dati, e lo stimatore cambia con loro. Su ChatBench la mediana è 81 gradi, sui ragionamenti di ChaosNLI 64, sul testo di HAIID 65. Un esame di fisica non è una storia di senso comune: nelle storie persone e macchine leggono lo stesso testo prodotto da persone, e

inciampano negli stessi punti. L'angolo misura quanto due intelligenze prendono gli stessi abbagli su quel lavoro, e con il lavoro cambia.

Misurarlo cambia i conti di chi lo dà per scontato. Sui dati della meta-analisi di Vaccaro e colleghi, che non hanno il caso per caso, si può calcolare la *quota del guadagno possibile*, $g = (d_S - d_H)/(|S| - d_H)$, che vale zero per chi ignora la macchina e uno per chi arriva all'ideale. Con l'angolo fissato a 90 gradi le persone ottengono in mediana il 30 per cento del guadagno possibile, circa un terzo: una cifra che si presterebbe a essere citata. Con 60 gradi, come sui testi misurati qui, la stessa quota salirebbe al 43 per cento. **Quel terzo dipende da un angolo che nessuno aveva misurato.** Su ChatBench, dove l'angolo si misura, la quota passa dal 47 al 53 per cento con GPT-4o e dal 73 al 105 con Llama, oltre il cento perché il riferimento è la macchina interrogata da sola, più debole di quella che in chat parlava con le persone. Sale, come il protocollo prevedeva, ma è aritmetica, perché con errori correlati l'ideale si accorcia.

Tabella 9. Chi somiglia a chi
mediana della correlazione fra le risposte giuste, sugli stessi casi. È una correlazione delle risposte, non degli indizi: sta su una scala diversa da quella degli angoli

Compito	Fra due macchine	Fra persona e macchina
ChatBench	0,37	0,12
αNLI	0,63	0,43
HAIID, testo	0,71	0,43
HAIID, immagini	0,67	0,20

Le macchine si somigliano fra loro più di quanto somiglino a noi, su tutte e quattro le famiglie di dati (tabella 9): per chi parte dal parere di una macchina, un'altra macchina aggiunge meno indipendenza di una persona. È come un'azienda che, per farsi controllare i conti, paga consulenti usciti dalle stesse scuole e cresciuti sugli stessi manuali del suo consiglio di amministrazione: il secondo parere inciampa sugli stessi gradini del primo. Steyvers e colleghi, nel 2022, avevano trovato lo stesso ordine su immagini degradate: una correlazione nascosta intorno a 0,4 fra una persona e una rete, a 0,7 fra due persone, fra 0,65 e 0,75 fra due reti. E anche due persone non sono ortogonali. Sulle stesse domande di ChatBench, due persone hanno fra loro un angolo di 66 gradi, più chiuso del loro angolo con la macchina; su αNLI 54, contro 64. **La macchina è più straniera del vicino di banco**, senza essere ortogonale. Sulla stessa geometria si possono mettere anche i pareri della folla, che HAIID mostra come se fossero dell'AI: fra 42 e 84 gradi secondo il compito.

La folla dà anche un termine di paragone. Su ChatBench due persone sbagliano insieme abbastanza (angolo di 66 gradi) da fissare un tetto: nel modello una folla di loro pari, per quanto numerosa e brava a mettere in comune gli indizi, non promette più di $1/\sqrt{\rho}$ volte uno solo: circa 1,57 turing apparenti. Il gruppo, con GPT-4o accanto, ne ha 2,21, più di qualunque folla di suoi pari; con Llama 1,38, meno.

Gli oggetti di Geirhos. Nel riconoscimento di oggetti, sui dati di Geirhos, la mediana è 59 gradi su 689 combinazioni di rete e condizione, con oscillazioni fra le reti: 62 gradi per la più debole, 56 per la più forte (correlazione di rango −0,45). Quei confronti sono 73 condizioni per dieci reti, le reti si somigliano fra loro e in quasi tutte le condizioni le persone sono quattro: pesano meno di quanto dica il numero. Il nostro angolo sembra lontano da ciò che Geirhos aveva trovato nel 2020, errori che coincidono poco più che per caso; ma quella misura guarda i verdetti, e il nostro angolo gli indizi che ci stanno dietro, dove un rumore in comune si vede anche quando i verdetti non coincidono. La tendenza con la forza della rete non si distingue dal caso.

C'è poi la crescita delle macchine. Se volontà e tecnica stanno ad angolo retto, e la macchina sta sulla tecnica, una macchina che cresce non dovrebbe avvicinarsi all'asse umano di un solo grado: nella nostra traduzione, l'angolo fra i suoi errori e i nostri dovrebbe aprirsi, o almeno non chiudersi. Dopo i primi numeri di ChatBench abbiamo aggiunto al protocollo questa prova: sulle stesse persone e le stesse domande, le macchine dalla più piccola alla più grande (figura 3).

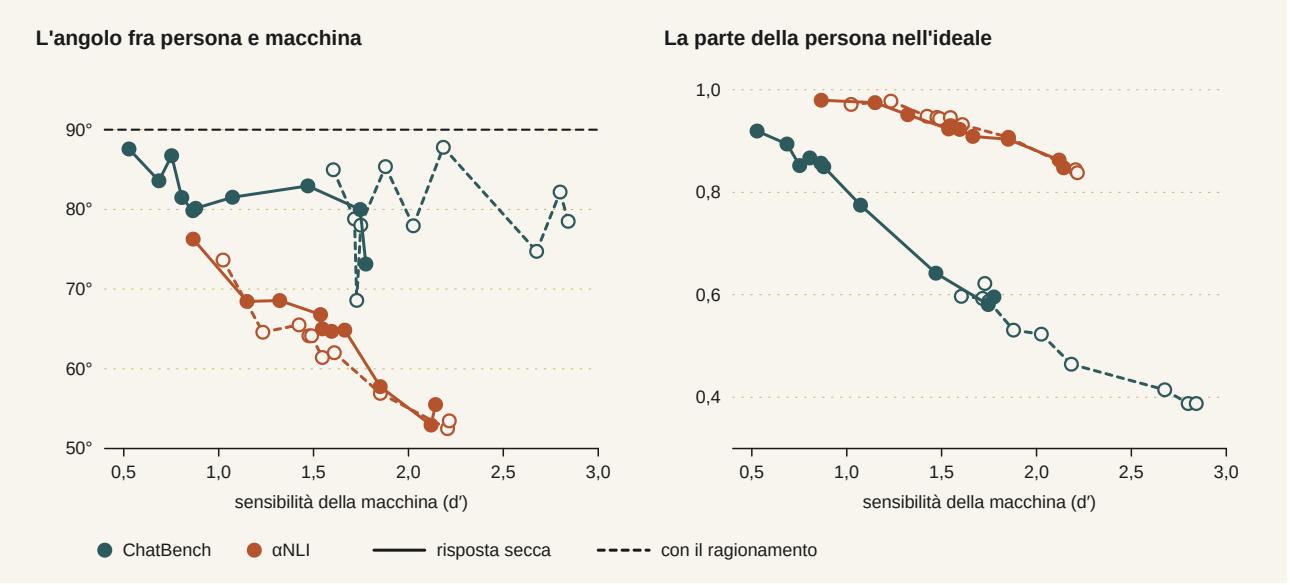


Figura 3. L'angolo e la parte della persona nell'ideale, al crescere della macchina.

La parte destra della figura dà ragione all'ipotesi, ma per aritmetica. La **parte della persona nell'ideale scende** quando la macchina cresce: è la quota $d_H/|S|$, e a parità di angolo è solo questione di lunghezze. In turing apparenti, che ne sono l'inverso, su ChatBench vuol dire da 1,09 a 1,68.

La parte sinistra è più interessante, e va letta con cura. Su ChatBench, in risposta secca, l'angolo va da 88 a 73 gradi, ma quasi tutto il calo viene dal modello più forte: gli altri stanno fra 80 e 88 gradi, con intervalli larghi una ventina di gradi, e senza il modello più forte la tendenza non si distingue più dal caso; con il ragionamento una tendenza non c'è.

Sui ragionamenti di ChaosNLI, che il protocollo non prevedeva per questa prova, il segnale è netto: **l'angolo si chiude quando la macchina cresce**, in risposta secca e con il ragionamento, e l'ordine delle macchine e quello degli angoli vanno insieme. Una spiegazione possibile è poco romantica. Una macchina più forte smette di sbagliare a caso e sbaglia dove il caso è difficile per tutti, comprese le storie in cui la folla dei cento votanti non è d'accordo con la chiave; e l'angolo, che non separa la difficoltà dalla somiglianza, lo legge come vicinanza. Per vederlo abbiamo tolto le 157 storie in cui la maggioranza dei votanti contraddice la chiave. Tutti gli angoli salgono, fra 6

e 17 gradi, e restano sotto i 90; la discesa resta, più corta: le storie contestate ne spiegano fra il 30 e il 43 per cento.

Le chiavi contestate spiegano una parte della discesa, non tutta, e il resto non l'abbiamo spiegato: può essere altra difficoltà condivisa, o somiglianza. Su quel compito, una macchina più capace sbaglia anche dove sbagliamo noi.

Un lavoro di quest'anno, di Kong e colleghi, sugli stessi dati di ChatBench, guarda la cosa da un'altra parte. In un loro modello teorico, con gli indizi della macchina pescati a caso, la sovrapposizione fra ciò che sa la persona e ciò che sa la macchina è, nelle loro parole, «not tunable through AI scale alone»: non si regola con la sola dimensione della macchina. Nei dati, però, cambia con la macchina: la sovrapposizione, nella loro misura, è 0,13 con Llama, la più piccola, contro 0,07 con GPT-4o, che porta più informazione nuova.

L'angolo esiste, sta sotto i 90 gradi in modo netto su due famiglie di dati su quattro, e dipende dal compito e da chi lo misura.

La formula alla prova

Uno strumento convince quando prevede, e la formula dei turing apparenti promette di dire in anticipo fin dove può arrivare una coppia. Per metterla alla prova servirebbero coppie vere, e i dati pubblici non le hanno: su ChatBench la persona non ha mai visto le macchine che abbiamo fatto girare noi. Abbiamo costruito allora una coppia di carta. Un *combinatore a verdetto*, una regressione logistica, prende la risposta che la persona ha dato da sola, che è un verdetto, e la probabilità data dalla macchina, e impara a chi credere quando non vanno d'accordo; lo si misura su casi che non ha visto. Ciò che raggiunge si confronta con ciò che il modello prevede per una coppia così, che della persona ha solo il verdetto: lo stesso combinatore, fatto girare su dati simulati con le sensibilità misurate, una volta con l'angolo misurato e una volta con i 90 gradi. Il test era deciso prima. La formula vera e propria, con tutti gli indizi, si può provare solo fra le reti.

Su ChatBench le due previsioni sbagliano di poco e finiscono pari, 0,06 unità di sensibilità ciascuna; su αNLI quella con l'angolo misurato sbaglia di 0,04, quella a 90 gradi di 0,29. Su HAIID i casi sono troppo pochi per un confronto. Una parte del vantaggio dell'angolo, però, è regalata, perché l'angolo è stimato proprio da quante volte i due rispondono giusto insieme. La prova che non parte avvantaggiata sta nei pochi confronti in cui il combinatore cade fra le due previsioni, e lì l'angolo è più vicino in 8 casi su 10, un risultato che il caso darebbe una volta su dieci. Con numeri così piccoli è un indizio.

Resta l'angolo utile. Se è una soglia vera, sotto quell'angolo la coppia non dovrebbe battere il migliore dei due, e sopra qualche volta sì. La tabella 10 incrocia le due cose.

Tabella 10. L'angolo utile e il guadagno vero

74 confronti, HAIID compresa; «guadagna» vuol dire che il combinatore a verdetto batte il migliore dei due oltre l'intervallo al 95 per cento

	Guadagna	Non guadagna
L'angolo supera l'angolo utile	15	55
L'angolo è sotto l'angolo utile	0	4

Sotto l'angolo utile cadono solo 4 confronti, troppo pochi per dire qualcosa, e in nessuno il guadagno supera il margine d'errore. Sopra, il guadagno si vede in **15 casi su 70**. Il criterio scritto prima chiedeva che le due risposte andassero d'accordo, e non passa: discordano in 55 confronti su 74. Il motivo sta soprattutto nel combinatore, che dalla persona riceve soltanto la risposta scelta: un verdetto, che butta via quasi tutti gli indizi che c'erano dietro. Con un verdetto in mano si può decidere solo a chi credere, e il tetto di una coppia così sta più in basso di quello della figura, disegnato per chi mette in comune tutti gli indizi. Sopra l'angolo utile resta poco da guadagnare, e con l'incertezza di questi dati quel poco, di solito, non si vede.

Fra due reti, invece, gli indizi ci sono tutti: le dieci reti per la visione hanno lasciato, per ogni immagine, la probabilità di ogni categoria. In ogni coppia la più debole fa da mente e la più forte da macchina, e un combinatore, questa volta a probabilità, impara a pesarle su immagini che poi non rivede. Due reti sbagliano sulle stesse immagini molto più di una persona e una rete: l'angolo mediano sta fra 22 e 43 gradi secondo il tipo di immagini, e in più di metà dei 2.615 confronti cade sotto l'angolo utile. I due criteri scritti prima (che l'angolo misurato preveda meglio dei 90 gradi, che il combinatore arrivi all'ideale) passano, ma con poco merito: il combinatore arriva in mediana proprio all'ideale, e una coppia finta che ricopiasse sempre la rete più forte li passerebbe anche lei.

Per sapere se la soglia fosse solo un effetto del vincolo sui pesi abbiamo poi tolto il freno, e lasciato al combinatore anche pesi negativi. In mediana la coppia resta pari alla rete più forte, sotto l'angolo utile e sopra; la differenza sta in quante volte la supera. Sopra l'angolo utile la batte in 533 confronti e le resta sotto in 364; sotto, la batte in 411 e le resta sotto in 455, mentre lì il modello senza fermata prometteva un guadagno mediano del 3,3 per cento. **L'angolo utile separa due tendenze**: sopra la coppia guadagna più spesso, sotto più di rado, e in nessuno dei due casi la regola è netta. Quale coppia guadagni di più, la formula non lo indovina. La formula non sbaglia il tetto, che su questi dati è quasi sempre il più forte dei due; chi guadagna, e quanto, non lo dice.

Simili, non stupidi

Una domanda torna spesso: l'AI ci rende più intelligenti, o ci sta rendendo stupidi? L'angolo utile permette di riformularla in modo che si possa rispondere. Nel modello, accanto a una macchina k volte più sensibile, una persona aggiunge qualcosa del suo solo se l'angolo fra i loro errori supera l'angolo utile, che vale $\arccos(1/k)$ e cresce con la macchina: più la macchina è brava, più la persona deve restare diversa da lei.

Ed è qui che la domanda cambia forma. Nel 2023 Lucía Vicente e Helena Matute hanno mostrato che chi lavora con un'AI che sbaglia in modo sistematico, rimasto da solo, continua a fare gli errori di quella. In uno studio successivo, con 86 studenti di medicina e un compito simulato, gli studenti che avevano usato l'AI, rimasti senza, su casi nuovi del tipo su cui l'AI aveva sempre sbagliato sbagliavano nel **26 per cento** dei casi, quelli del gruppo di controllo nell'1. Succede come a chi impara il tennis da un maestro con un difetto nell'impugnatura: il maestro se ne va, e il difetto resta nel braccio dell'allievo. In questi studi la macchina non ci rende stupidi: ci rende simili a sé. E l'accuratezza non se ne accorge, perché si può diventare più bravi mentre i propri errori si spostano su quelli della macchina. Soglia e angolo si vanno incontro: la soglia sale a ogni macchina nuova, e l'angolo della persona, se l'uso lo chiude, scende. Se così fosse, la macchina rifaserebbe la persona: bene per la rete, male per la coppia.

La domanda si può provare su qualcosa che costa meno di una persona. Sessanta reti piccole, che abbiamo chiamato *persone di silicio*, hanno imparato a riconoscere immagini a tappe, in tre gruppi: uno da solo, uno con il parere di una macchina sana, uno con il parere di una macchina a cui abbiamo iniettato un difetto, che su una parte di gatti e cani scambia le due classi. La copia si vede senza formule: sui 277 casi in cui il difetto cambia la risposta, le reti cresciute accanto alla macchina difettosa danno la sua risposta sbagliata nel 23,2 per cento dei casi, quelle cresciute da sole nell'11,0, e l'intervallo della differenza sta tutto sopra lo zero.

La misura che un nostro protocollo precedente aveva scelto in anticipo, invece, non la vede. Era pensata per separare gli errori presi dalla macchina dal semplice effetto della pratica, che raccoglie gli errori di tutti sui casi difficili e fa salire la somiglianza con la macchina anche se nessuno copia niente. Il trabocchetto si presenta davvero, e la macchina sana non fa suonare l'allarme; ma la causa più probabile del fallimento è che il difetto tocca due classi su dieci, mentre la misura è una media sulle classi, che lo annacqua. I dettagli sono nel repository dell'esperimento. Resta la regola che ci eravamo dati: una rete non è una persona, ma una misura che fallisce su un caso noto **non è pronta per le persone**. Prima di uno studio vero la misura principale va cambiata, mettendo in testa la domanda più semplice: quante volte, sui casi in cui la macchina sbaglia, la persona dà esattamente la sua risposta sbagliata.

Nessuno, per quanto abbiamo trovato, ha ancora seguito per mesi l'angolo fra gli errori delle stesse persone e quelli di una macchina vera, con un gruppo di controllo. Il protocollo per farlo, con la misura corretta, è il pezzo che da soli non possiamo fare: due gruppi tirati a sorte, la stessa pratica settimanale, una macchina congelata, duecento casi mai visti da affrontare da soli prima di cominciare e dopo tre mesi, e un comitato etico che vigili. Lo proponiamo a un gruppo di ricerca che voglia condurlo con noi. Nel conto finale servirà un'accortezza. Se in tre mesi chi lavora con la macchina perde allenamento, la sua sensibilità da sola scende, e il turing sale anche se la coppia non è migliorata, come per VGG16 davanti alle immagini più degradate. Il turing va calcolato allora due volte, con la sensibilità da sola di quel giorno e con quella del primo giorno, su casi della stessa difficoltà. Il rapporto fra i due è proprio quanto la persona, da sola, è cambiata: se il primo supera il secondo, un turing alto nasconde una persona che da sola vale meno di prima.

Oltre i 90 gradi, nei nostri dati, cadono solo confronti isolati, incerti o al limite dello strumento (su HAIID uno arriva a 103, con un intervallo che comprende i 90): nessuna mediana li supera. Lì gli errori andrebbero verso la controfase, una delle due menti tenderebbe ad avere ragione proprio dove l'altra sbaglia, e la formula dei turing apparenti promette più di Pitagora, verso i 180 gradi senza limite. Gli elettricisti conoscono qualcosa di simile. In un circuito con una resistenza, una bobina e un condensatore in serie, le tensioni ai capi della bobina e del condensatore stanno sempre in controfase, e a una frequenza precisa hanno anche la stessa ampiezza e si cancellano: resta la sola resistenza, e la corrente sale al massimo che questa consente. È la *risonanza*. Nel modello, con persona e macchina di pari sensibilità, se si disegna il guadagno della coppia (il quadrato del turing, meno uno) al variare del peso dato alla macchina, a ogni angolo la curva ha esattamente la forma di quella di un circuito risonante, e il peso fa da manopola di sintonia. L'angolo decide quanto il picco è alto e stretto: basso e largo sotto i 90 gradi, sempre più aguzzo verso i 180. Chiamiamo *risonanza cognitiva* questa condizione, in cui gli errori di due menti si cancellano invece di sommarsi. Se una coppia così si possa costruire, per esempio addestrando una macchina

proprio sugli errori di una mente, è la domanda di «Risonanza cognitiva», l'articolo che continua questo.

Perché fidarsi della misura

Una misura, per meritare il nome, deve avere alcune cose, e la tabella 11 le mette in fila. Il turing ne ha una parte, e il resto va detto con la stessa chiarezza.

Tabella 11. Che cosa ha il turing di una misura, e che cosa gli manca
i requisiti di una misura, uno per riga, con la prova che sta in queste pagine

Requisito	Il turing	La prova
Definizione	Sì per una mente; ciò che misuriamo è il rapporto di un gruppo, che non è la media dei turing delle persone	Questa pagina e il protocollo
Zero e scala	Sì, se vale il modello gaussiano a varianza uguale, che sui dati non abbiamo verificato: zero al caso, nessun tetto, rapporti con un senso	Il risultato di Turing, che è un teorema del modello
Indipendenza dalle macchine	Per la definizione sì; i valori dipendono da compito, casi, persone e circuito	I turing per materia
Riferimento	Mobile, come in ogni guadagno: il campione è ogni volta il gruppo da solo	La definizione
Scala assoluta accanto	No: fra compiti diversi, e fra gruppi su prove diverse, manca una prestazione di riferimento	Nessuna: servirebbe una banca di domande tarate
Incertezza	Sui turing e sugli angoli sì, larga per eccesso sul disegno di ChatBench e non tarata altrove; turing apparenti e η senza intervallo	La taratura
Taratura su verità nota	Sì, ma solo nel mondo del modello	1.350 esperimenti finti
Robustezza	Solo al cambio di scala, su ChatBench: meno di un centesimo sui due valori di riferimento, meno di tre per materia; varianze diverse e gating non provati	La scala di Luce
Ripetizione	Solo del calcolo: nessuna misura ripetuta sulle stesse persone	Codice e dati nel repository
Validità	In parte: dice quanto cambia chi decide, non chi ha fatto il lavoro, e comprende il ripensamento	Le tre zone, con k

Due righe vanno lette insieme, quelle del riferimento. Un campione che cambia a ogni misura appartiene alla natura di ogni guadagno. Il guadagno di un amplificatore si dà in decibel rispetto al segnale in ingresso, e nessuno dubita che sia una misura; per dire quanta potenza esce davvero, i tecnici usano un'altra scala, il dBm, riferita a un milliwatt fisso. Il turing è il guadagno. Con una differenza: un buon amplificatore guadagna lo stesso qualunque segnale gli entri, mentre il turing cambia con la mente che entra e con il compito, e per questo si dà insieme alla sensibilità da sola e a k, come la scheda di un amplificatore dà il guadagno insieme al punto di lavoro. Sullo stesso compito il dBm c'è già, ed è la sensibilità della coppia, d_S. Manca fra compiti diversi, ed è la scala che servirebbe per dire se un gruppo con la macchina decide meglio di un altro gruppo, su un'altra materia.

Un'unità vale quanto lo strumento che la legge, e uno strumento si prova dove la verità si conosce. Abbiamo generato dal modello gaussiano 1.350 esperimenti finti, in sette situazioni di cui conoscevamo il turing vero: sei con il disegno di ChatBench, centosessanta persone con sette domande a testa e quattro opzioni, la settima con trenta persone deboli. Su ciascuno ha girato lo stesso programma che ha prodotto i numeri di queste pagine. **La lettura è centrata**: in sei situazioni su sette il turing medio sta entro l'uno per cento dal vero, e una misura sola ha un errore tipico del 4-7 per cento. Gli intervalli, invece, sono più larghi del necessario. Il criterio scritto prima chiedeva che contenessero il vero fra 90 e 98 volte su cento, e non passa in cinque situazioni su sei, dove lo contengono 99 o 100 volte: sbagliano dalla parte prudente, e una parte della ragione è nota, perché sta nel modo di ricampionare insieme persone e domande. Gli intervalli di queste pagine, sul disegno di ChatBench, vanno letti come intervalli al 99 per cento circa. Con trenta persone deboli, come nelle domande di fisica, lo strumento regge peggio: sopravvaluta il turing del 4 per cento, e l'errore tipico di una misura sale al 19.

La taratura ha tre limiti, e il primo pesa. Prova lo strumento nel mondo del modello: indizi gaussiani, coppie che mettono insieme gli indizi, nessun verdetto secco, nessun gating, nessuna chiave sbagliata, nessuna conversazione scelta dopo. Dimostra che lo stimatore è centrato, non che il modello descriva le persone. Il secondo: vale per il disegno di ChatBench, e anche lì somiglia di più alle conversazioni con una risposta in chat, mentre il valore di riferimento usa tutte le prove, 1.493 di 168 persone, circa nove a testa; sugli altri dati, dove i casi o le persone sono pochi, gli intervalli non sono tarati. Il terzo: il valore vero è quello del gruppo, e quanto il turing di un gruppo si discosti dal turing tipico delle sue persone la taratura non lo dice. Una sensibilità calcolata sulle risposte di molte persone messe insieme non è la media delle loro sensibilità, perché dall'accuratezza alla sensibilità non si passa con una retta, e la distanza fra le due è tanto maggiore quanto più le persone sono diverse fra loro (per la rilevazione sì/no l'hanno studiata Macmillan e Kaplan nel 1985). Il turing misurato può spostarsi così in un verso o nell'altro: in su se la macchina rende il gruppo più omogeneo, aiutando di più chi da solo vale meno, in giù se allarga le differenze. Di quanto, sui nostri dati, non l'abbiamo calcolato.

La taratura ha trovato anche un difetto nostro, che dichiariamo senza averlo corretto: la correlazione si calcola su una griglia di soglie troppo larga, e con dati senza fine non arriverebbe al vero. Sui dati di ChatBench sposta l'angolo di meno di un grado; il turing non dipende dalla correlazione.

Il turing regge, su questi dati, al modo di trasformare le accuratezze in sensibilità: con la scala logistica del modello di scelta di Luce al posto della gaussiana, il turing di GPT-4o e quello di Llama cambiano di meno di un centesimo. Smette di funzionare, invece, dove la persona da sola vede poco: un rapporto con un denominatore piccolo e incerto oscilla molto, e quando la sensibilità da sola è vicina a zero il turing sale anche se nessuno collabora di più. Non ha un tetto. Quando una coppia risponde giusto a tutto la sensibilità non si stima più, e serve un compito più difficile: **a saturare sono i test, il turing no.** Affidabile, dunque, in un senso ristretto: centrato, e con un'incertezza dichiarata per eccesso, finché vale il modello. Che cosa racconti un turing, sostituzione o collaborazione, lo dice il confronto con k.

Una misura vale anche per i dati su cui gira, e per la parte delle sue ipotesi che era fissata prima di vederli. Su una GPU da 24 GB abbiamo fatto girare 12 modelli linguistici aperti, dai più piccoli ai più grandi, in due modalità (la risposta secca, oppure il ragionamento prima della risposta), e dieci

reti per la visione uscite fra il 2014 e il 2023, sulle stesse domande e immagini che avevano visto le persone. Il protocollo con le ipotesi è stato scritto prima di guardare i risultati, e ciò che è cambiato dopo è dichiarato nel repository dell'esperimento, in una sezione apposita, «Deviazioni». Erano decisi prima di guardare i dati l'angolo, la somiglianza fra le macchine, la prova della formula e l'esperimento sulle reti che imparano. La taratura e la prova fra le reti hanno criteri scritti prima di farle, ma a risultati già visti. Sono venuti dopo i primi numeri il turing come unità, i due morsetti, l'ordine delle mosse e la crescita delle macchine.

I limiti, e che cosa resta da misurare

La traduzione è la nostra e va dichiarata per prima: **l'ipotesi dell'angolo retto è messa alla prova su una sua traduzione**, gli errori su casi con risposta nota, che non sono volontà e tecnica. Poi ci sono i limiti dei dati. Le persone sono lavoratori reclutati in rete, su compiti brevi; nessuno è specialista, nessuno è seguito nel tempo. Le macchine fatte girare da noi sono modelli aperti piccoli; GPT-4o e Llama 3.1 sono quelle che ChatBench ha reso disponibili: non abbiamo misurato i modelli di frontiera di oggi.

Le prove dicono anche dove lo strumento è ancora corto. Sei aggiunte lo renderebbero più utile, e nessuna chiede una teoria nuova.

- **Una scala assoluta accanto al turing.** Per confrontare gruppi diversi, e prove diverse della stessa materia, serve una prestazione con un riferimento che non cambi a ogni misura: una banca di domande tarate, come quelle che la psicometria costruisce con il modello di Rasch, su cui una persona di riferimento abbia per definizione una sensibilità fissata. Servono anche la stessa misura ripetuta sulle stesse persone, e un gruppo che risponda una seconda volta senza macchina, per separare l'aiuto dal semplice ripensamento.
- **La difficoltà, separata dalla somiglianza.** Nell'angolo finiscono insieme quanto due menti ragionano in modo simile e quanto i casi sono difficili per tutti. Un modello che dia a ogni domanda la sua difficoltà, come fa la *teoria della risposta all'item* nei test standardizzati, e misuri l'angolo su ciò che resta, direbbe quanta parte della vicinanza è somiglianza vera. Conta chi fa la taratura: se la difficoltà si stima sulle persone, una macchina che inciampa sulle stesse domande sta già somigliando a loro, e va detto.
- **Il tempo.** Il turing conta la sensibilità per decisione, non per minuto, e METR ha misurato proprio il tempo. Un turing alto ottenuto controllando a lungo le risposte della macchina può valere meno di quanto sembra. Dividere il peso dell'evidenza per il tempo, deciban al minuto, darebbe la grandezza che il wattmetro misura davvero: lavoro nell'unità di tempo.
- **Ciò che la conversazione tira fuori.** In chat Llama è più accurata che da sola, di una decina di punti: una parte del contributo della persona sta nel far rendere di più la macchina. Il guadagno si può separare in due fattori, quanto la conversazione alza la sensibilità della macchina e quanto bene la persona decide alla fine, a patto di misurare la macchina della conversazione su domande a cui la persona non ha ancora risposto. E le conversazioni da cui non esce una risposta si possono leggere una per una, per sapere perché.
- **La sicurezza di sé della macchina.** Bahrami e colleghi hanno mostrato che due persone battono la migliore delle due solo se sono brave più o meno uguale e possono dirsi, caso per caso, quanto sono sicure; il loro modello suppone che ciascuno comunichi bene la propria sicurezza. Un modello linguistico tende a usare lo stesso tono quando ha ragione e quando ha

torto, e questo rende difficile il gating: misurare il turing con macchine che dichiarano quanto sono sicure, e con macchine che non lo fanno, direbbe quanto pesa.
- **Le domande aperte.** Su compiti senza una risposta giusta d' non esiste. Lì si potrebbe misurare quanto spazio delle idee esplorano la persona, la macchina e la coppia, per esempio dalla dispersione delle loro proposte: un gemello del turing per il gesto che il nostro strumento non vede.

Che cosa abbiamo trovato

Tre cose reggono meglio delle altre: lo strumento è centrato dentro il modello; con GPT-4o le persone salgono e la coppia resta sotto la macchina, da sola e in chat; su αNLI e sul testo di HAIID l'angolo sta sotto i 90 gradi. Il resto è indizio: il sorpasso con Llama, la chiusura dell'angolo con la crescita, la soglia dell'angolo utile. Messi in fila, i risultati hanno solidità diverse, e la tabella 12 le dà insieme ai numeri. «Deciso prima» vuol dire che ipotesi e criterio stavano nel protocollo prima dei conti; «aggiunto dopo», che l'analisi è venuta guardando i dati, e vale come indizio da confermare.

Tabella 12. Che cosa abbiamo trovato
una riga per domanda; «deciso prima» vuol dire nel protocollo prima dei conti

Domanda	Risposta	Cifra	Quanto è solida
Lo strumento ritrova il turing vero?	Sì, nel mondo del modello, con intervalli prudenti	entro l'1% in 6 situazioni su 7; l'intervallo al 95% contiene il vero nel 98-100% dei casi	Criterio scritto prima della prova, a risultati già visti; non passa per eccesso di prudenza
Quanto vale un gruppo di persone con GPT-4o?	Sale, e frena la macchina	1,65 turing; resa sulla macchina 0,76	Aggiunto dopo; intervalli che escludono l'uno
E con Llama 3.1 8B, più debole?	Sale, e la coppia supera la macchina interrogata da sola	1,40 turing; resa sulla macchina 1,19	Aggiunto dopo; dipende dal morsetto: 0,92 contro la macchina della conversazione, con un intervallo che tocca l'uno
Contare solo le conversazioni riuscite cambia il turing?	Sì, lo gonfia; e dove la chat non ha dato una risposta, le persone non guadagnano niente	1,90 e 1,54 dove la chat ha risposto, 1,03 e 1,11 dove no; dove non ha dato nessuna risposta, GPT-4o da sola 84%, le persone con lei 50%, come da sole	Aggiunto dopo; prove scelte da come è andata la conversazione: descrive, non prova una causa
L'angolo fra persona e macchina è retto?	Su αNLI e sul testo di HAIID no; su ChatBench non si distingue dal retto in 13 confronti su 20	mediane fra 59 e 81 gradi; su αNLI 20 intervalli su 20 sotto i 90	Deciso prima; il criterio, su ogni famiglia, non passa
Se la macchina cresce, l'angolo si apre?	No: su αNLI si chiude, su ChatBench quasi niente	su αNLI da 76 a 56 gradi	Aggiunto dopo i primi numeri; αNLI fuori dal protocollo; senza le chiavi contestate la discesa resta, più corta
La formula prevede ciò che fa una coppia?	Il tetto sì, il guadagno no	su αNLI errore 0,04 contro 0,29, su ChatBench pari, con la previsione avvantaggiata; 8 su 10 dove non lo è; a verdetto, guadagno oltre il margine in 15 casi su 70	Deciso prima; prova debole
L'angolo utile è una soglia vera?	Separa due tendenze: sopra la coppia batte più spesso la rete più forte, sotto più di rado	con pesi liberi, sopra 533 volte meglio e 364 peggio, sotto 411 e 455; in mediana pari in tutti e due i casi; sotto, il modello senza fermata prometteva +3,3%	Fra le reti, lettura nata dopo i risultati e pesi liberi aggiunti dopo; fra persone e macchine il criterio deciso prima non passa (55 discordanze su 74)
Chi impara da una macchina difettosa ne prende gli errori?	Sì, nelle reti; la misura scelta prima non lo vede	23,2% contro 11,0%	Esperimento deciso prima; la misura dichiarata prima fallisce; la copia è una misura aggiunta dopo

Ciò che lo strumento non vede

Nella nostra traduzione l'ipotesi da cui siamo partiti prevedeva due cose: un angolo retto fra persona e macchina, e un angolo che non si chiude quando la macchina cresce. La prima non regge dove l'angolo si misura bene. La seconda non regge su αNLI; le chiavi contestate spiegano meno della metà della discesa, e il resto non l'abbiamo spiegato. Dunque l'ipotesi ha torto? La traduzione dice di sì, ed è compatibile con una macchina che, crescendo, sbaglia dove sbagliamo noi. La traduzione, però, è nostra: l'ipotesi parla di volontà e tecnica, e qui si parla di errori su esami e storie. L'errore su una domanda con la risposta già scritta sta tutto sull'asse della tecnica. **Quello che abbiamo misurato con tanta precisione è l'angolo fra due tecniche.**

Tutti i numeri di queste pagine condividono un presupposto. Qualcuno ha scelto la domanda, qualcuno ha fissato le opzioni, qualcuno ha deciso quale fosse la risposta giusta. Ogni caso arriva con la sua chiave di correzione. Il d' si calcola soltanto perché qualcuno, prima di noi, ha già deciso cosa conta come giusto.

Conviene tenere distinti due gesti: *comprimere*, trovare il meglio dentro un mondo dato, ed *espandere*, fare un mondo nuovo. Tutta la nostra misura vive dentro un mondo dato. È come giudicare un esploratore da quanto cammina veloce su un sentiero già tracciato, senza chiedersi chi ha disegnato la mappa. Nei confronti di queste pagine non c'è una sola domanda scelta da una macchina, e non c'è una sola domanda scelta dalla persona che risponde. Nessuno, in questo esperimento, sceglie la domanda: lo strumento comincia a funzionare dopo.

Si potrebbe trarne un sollievo: l'ipotesi originale, quella su volontà e tecnica, è in salvo. Sarebbe un sollievo a buon mercato, perché un'ipotesi al riparo da ogni strumento è fuori misura. In linguaggio kantiano: la volontà, per il nostro strumento, è una cosa in sé. Lo strumento vede soltanto come si proietta sulle domande che qualcuno ha già posto.

La macchina, del resto, può arrivare a risposte che non sapremmo riconoscere, e l'accuratezza non può vederle per costruzione, perché è la proiezione di un'intelligenza sul piano delle domande umane. Il wattmetro legge il lavoro che il circuito riesce a fare su quel carico. Di un carico diverso non sa nulla.

Si può obiettare che una domanda un valore misurabile ce l'ha. Prima ancora della risposta si può calcolare di quanto restringerà, in media, le ipotesi in gioco: è il *guadagno d'informazione atteso*, e si conta negli stessi deciban di Bletchley. Su una diagnosi, o su una ricerca fatta di molte mosse, si potrebbe contare quanti deciban per domanda ottengono una persona e una macchina quando tocca a loro decidere che cosa chiedere, e sarebbe una misura preziosa. Per farlo, però, bisogna avere già l'elenco delle ipotesi. Il guadagno d'informazione misura la domanda migliore dentro un mondo dato, e il mondo resta di qualcun altro.

Quello che manca, e che nessuno ha, è uno strumento che legga l'atto di aprire un dominio: quanto di ciò che vale la pena chiedere viene chiesto, e da chi. Non si misura con un test a risposta chiusa. Forse si misura guardando, a distanza di anni, quali domande sono diventate domande grazie a chi. È un conto che non abbiamo fatto.

Originare, di nuovo

Una postilla di metodo, che riguarda chi ha scritto questo testo. Il codice degli esperimenti, le analisi e una prima stesura delle pagine che avete letto sono stati prodotti da un modello linguistico, sotto la direzione dell'autore e rivisti da lui. È un caso di quelli di cui parla l'articolo, e il più pulito che avessimo: sappiamo chi ha ordinato il compito e chi lo ha eseguito. La domanda, se ci fosse un angolo retto fra due intelligenze e come si potesse leggere, stava da mesi in un quaderno.

Lovelace aveva scritto che la macchina può fare «qualunque cosa noi sappiamo ordinarle di fare». Con gli anni quella frase è stata presa come il limite della macchina. A rileggerla è anche la descrizione esatta di una competenza nostra, e neanche la più coltivata: *sapere ordinare*. Saper dire che cosa si vuole abbastanza bene perché ottenerlo significhi qualcosa. Nei nostri dati ce n'è forse un'ombra: in un quarto delle prove con GPT-4o la chat non ha dato una risposta fra le quattro, su domande che la macchina, interrogata da sola, risolveva quattro volte su cinque. Perché, non lo sappiamo. In tutti i confronti che abbiamo misurato quel sapere era già stato esercitato, da chi ha scritto una domanda d'esame, da chi ha inventato una storia a cui mancava una frase. Le persone e le macchine dei nostri confronti hanno lavorato dentro un mondo che qualcun altro aveva già fatto.

Il wattmetro che cercavamo si può costruire, e lo abbiamo tarato: dice di quanto una macchina cambia chi decide, quanto due intelligenze si sovrappongono, quanto una serve ancora accanto all'altra. Ha detto cose che l'ipotesi da cui siamo partiti non prevedeva, e abbiamo scritto anche quelle. Ciò che non legge è il gesto che lo ha reso utile: qualcuno che, davanti a un motore, si è chiesto quanto lavoro stesse facendo davvero.

I conti

Le sensibilità. Il d′ da scelta forzata fra m opzioni si ricava invertendo $PC = \int \phi(x - d)\,\Phi(x)^{m-1} dx$. Con due classi si usa la convenzione sì/no, la somma degli z delle risposte giuste sulle due classi, che vale $2z(a)$ quando il criterio è neutro e differisce di un fattore $\sqrt{2}$ dalla scelta forzata: turing, η, g e angoli sono rapporti e non ne dipendono, mentre gli errori in unità di sensibilità non si confrontano fra famiglie di dati. Dove la varianza non è uguale la stessa definizione si scrive con l'area sotto la curva ROC, $d_a = \sqrt{2}\,z(\mathrm{AUC})$. Le sensibilità sono di gruppo: vengono dalle risposte di tutte le persone, messe insieme, e lo stesso vale per i turing; la definizione vale per una mente, la misura qui è di un gruppo. Nella convenzione sì/no il peso dell'evidenza atteso per osservazione, in unità naturali, è $d'^2/2$. La scala logistica usata per confronto è $\ln[a(m-1)/(1-a)]$, con a l'accuratezza.

L'ideale della coppia. È la sensibilità dell'osservatore lineare ottimo con pesi non negativi su due indizi gaussiani correlati: $|S|^2 = \mathbf{d}^\top \Sigma^{-1} \mathbf{d}$, con $\mathbf{d} = (d_H, d_M)$ e Σ la matrice di correlazione, sopra l'angolo utile; sotto, l'ottimo libero darebbe al più debole un peso negativo e supererebbe il più forte, e qui ci si ferma al più forte. I turing apparenti sono $|S|/d_H$. L'efficienza η di Tanner e Birdsall si misura sulle energie, rispetto a un osservatore ideale: con l'ideale della coppia al posto del ricevitore ideale, è il quadrato del nostro fattore. Il tetto di una folla è quello di n osservatori di pari sensibilità e pari correlazione, $d'\sqrt{n/(1+(n-1)\rho)}$, per n che cresce. Sul tetto della folla

di ChatBench, 1,57 turing apparenti, l'intervallo va da 1,37 a 1,71: ricampiona solo i casi, ed è più stretto del vero.

L'angolo. ρ è la correlazione tetracorica condizionata alla risposta giusta, stimata per ciascuna persona con la sua soglia, con un restringimento empirico-bayesiano verso la soglia media; con più di due opzioni la correlazione delle risposte giuste si converte in correlazione degli indizi con una tabella calcolata per simulazione. Fra due persone la conversione le suppone entrambe con la sensibilità media del gruppo, ed è un'approssimazione. La tabella va da −0,6 a 0,99, e tronca gli angoli fra 8 e 127 gradi. Sugli oggetti di Geirhos la stima tocca il limite della tabella in 30 confronti su 689, e l'intervallo in 266: lì gli intervalli sono più incerti di quanto si legga, e la tendenza dell'angolo con la forza della rete ha p = 0,19. Su ChatBench, in risposta secca, la correlazione di rango fra la forza delle macchine e l'angolo è −0,73, e senza il modello più forte −0,63 (p = 0,08); con il ragionamento è −0,02. Su αNLI, togliendo le storie con le chiavi contestate, gli angoli cambiano così (tabella 13).

Tabella 13. L'angolo su αNLI, con e senza le chiavi contestate

gradi, dalla macchina più debole alla più forte; correlazione di rango fra la forza delle dieci macchine e l'angolo; «senza» toglie le 157 storie in cui la maggioranza dei cento votanti contraddice la chiave; la discesa spiegata è calcolata sugli angoli prima di arrotondarli

Modo	Tutte le storie	Senza le contestate	Correlazione, tutte	Correlazione, senza	Discesa spiegata dalle contestate
Risposta secca	76,3 → 55,5	82,6 → 68,1	−0,96	−0,89	30%
Con il ragionamento	73,6 → 53,5	80,7 → 69,1	−0,95	−0,90	43%

La taratura e gli intervalli. Gli intervalli, compresi quelli del turing, vengono da un bootstrap a due livelli, persone e casi, con pesi di Poisson; quello del turing mediano della meta-analisi ricampiona gli studi. Per provarli avevamo scritto prima il criterio: un errore sistematico entro il 5 per cento, e un intervallo al 95 per cento che contenga il vero fra 90 e 98 volte su cento (tabella 14). Valeva per sei situazioni su sette; per la settima, trenta persone deboli, ci aspettavamo di peggio e avevamo deciso solo di riportare di quanto.

Tabella 14. La taratura dello strumento

dati simulati dal modello con il disegno di ChatBench (la settima situazione con trenta persone e otto domande a testa); per ogni situazione almeno 150 esperimenti, e su ciascuno lo stesso programma dell'articolo. La copertura è la quota di esperimenti in cui l'intervallo al 95 per cento contiene il turing vero; la larghezza è il rapporto fra l'intervallo e quello che basterebbe. Criterio dichiarato prima: errore entro il 5 per cento e copertura fra 90 e 98

Situazione	Turing vero	Errore sistematico	Copertura	Larghezza	Criterio
Coppia ideale	2,28	−0,6%	99%	1,4	Troppo largo
Coppia a metà strada	1,72	+0,5%	99%	1,6	Troppo largo
Macchina debole, seguita troppo	0,91	−0,1%	99%	1,5	Troppo largo
La persona copia la macchina	2,20	−0,2%	98%	1,3	Passa
Macchina pari alla persona	1,30	−0,8%	100%	1,6	Troppo largo
Macchina che risponde prova per prova	2,27	−0,1%	100%	1,7	Troppo largo
Trenta persone deboli, come in fisica	2,57	+4,1%	99%	2,4	Non soggetta al criterio

L'intervallo del turing è da 1,3 a 1,7 volte più largo del necessario. Una ragione nota spiega quella larghezza: ricampionare persone e casi con pesi indipendenti, moltiplicati fra loro, conta fino a tre volte la parte di variabilità che appartiene alla singola risposta, ed è prudente per costruzione (Owen, 2007; Owen ed Eckles, 2012). Con sette domande a persona quella parte domina, e una larghezza fino a 1,7 volte è ciò che ci si aspetta. Per l'angolo la prudenza è maggiore: l'intervallo contiene il vero 97-100 volte su cento, ed è fino a 2,1 volte più largo. La ragione nota non spiega tutto: per l'angolo, e per il turing con trenta persone deboli, l'intervallo è più largo ancora, fino a 2,4 volte, e lì il meccanismo non l'abbiamo isolato.

La formula alla prova, nei numeri. Il combinatore a verdetto è una regressione logistica standardizzata addestrata sui soli disaccordi in validazione incrociata a gruppi di casi; la previsione simula la stessa regressione sui dati generati dal modello gaussiano con le sensibilità e la correlazione misurate (tabella 15).

Tabella 15. Quanto sbaglia la previsione

errore assoluto medio, in unità di sensibilità, fra ciò che il combinatore a verdetto raggiunge e ciò che la formula prevede, con l'angolo misurato e con i 90 gradi. Poi, i confronti in cui l'angolo misurato si avvicina di più e quelli che cadono fra le due previsioni, i soli in cui il confronto è informativo

Dati	Con l'angolo misurato	A 90 gradi	Meglio con l'angolo	Fra le due previsioni
ChatBench	0,057	0,057	9 su 20	6
αNLI	0,042	0,289	20 su 20	4
HAIID, testo	0,373	0,521	17 su 18	0
HAIID, immagini	0,378	0,378	9 su 16	2
Tutti, senza HAIID	0,050	0,173	29 su 40	10
Tutti	0,199	0,302	55 su 74	12
Riconoscere oggetti	0,044	0,139	497 su 673	318

Su HAIID i casi sono 24-32 per compito, e la sensibilità del combinatore non è confrontabile con la previsione. Le due righe «Tutti» mettono insieme famiglie con convenzioni diverse, la scelta forzata di ChatBench e il sì/no di αNLI, che differiscono di un fattore $\sqrt{2}$: vanno lette come indicazione. Portati tutti alla scelta forzata, senza HAIID, gli errori sarebbero circa 0,04 contro 0,13. I confronti in cui il combinatore cade fra le due previsioni sono 12 contando HAIID e 10 senza, e l'angolo è più vicino in 8 in tutti e due i conti; il caso darebbe un risultato così quattro volte su dieci nel primo, una su dieci nel secondo. Sugli oggetti, dove questi confronti sono centinaia, la stessa prova dà 265 su 318, con l'avvertenza che non sono indipendenti.

Fra le reti i criteri scritti prima erano due: che la previsione con l'angolo misurato sbagliasse meno di quella a 90 gradi, e che il combinatore arrivasse in mediana fra 0,9 e 1,1 volte l'ideale. Passano tutti e due (errore 0,10 contro 0,75, rapporto mediano 1,00); una coppia finta che ricopiasse sempre la rete più forte li passerebbe con un errore anche minore (0,07). I confronti validi sono 2.615, e in 1.383 l'angolo cade sotto l'angolo utile. Il combinatore aveva pesi non negativi, come l'ideale, e quindi sotto l'angolo utile non poteva battere la rete più forte se non per rumore: lì la soglia era confermata per costruzione. Anche così, la coppia supera la rete più forte nell'8 per cento dei confronti sotto l'angolo utile e nel 40 sopra. La prova con pesi liberi, anche negativi, è venuta dopo: sotto l'angolo utile la formula senza fermata prometteva un guadagno mediano del 3,3 per cento, sopra ne prometteva uno dell'1,3 per cento sulla rete più forte. In tutti e due i casi il combinatore in mediana non ne ottiene. Quale coppia guadagni di più la formula non lo indovina (correlazione di rango 0,09).

Da dove vengono i confronti. Le quattro raccolte sono di altri: ChatBench di Chang, Anderson e Hofman (2025); ChaosNLI di Nie, Zhou e Bansal (2020), di cui abbiamo usato la parte αNLI, 1.532 storie in cui la frase mancante si sceglie fra due candidate; HAIID di Vodrahalli e colleghi (2022), con compiti di sarcasmo, censimento, arte, città e dermatologia; le prove, una per una, di Geirhos e colleghi (2021), su persone e reti davanti a immagini degradate (tabella 16). I confronti non pesano tutti uguale. Quelli sulle immagini di Geirhos sono tanti ma hanno poche persone ciascuno, e si somigliano fra loro; le raccolte indipendenti di giudizi umani sono quattro.

Tabella 16. Da dove vengono i confronti
un confronto è un gruppo di persone e una macchina, o due macchine, sugli stessi casi

Fonte	Che cosa c'è	Confronti	Dove si usa
ChatBench	Domande d'esame a quattro opzioni; persone da sole, poi in chat con GPT-4o o con Llama 3.1 8B	20 con i nostri modelli, 4 con quelli della chat (due macchine, due morsetti)	Turing, morsetti, angolo, crescita
ChaosNLI, αNLI	Storie a cui manca una frase; cento voti per caso	20	Angolo, crescita
HAIID	Cinque compiti brevi, con il parere di una folla	34, più 5 con la folla	Angolo, folla
Geirhos e colleghi	Immagini degradate; poche persone per condizione	689	Angolo, formula
Vaccaro e colleghi	Meta-analisi: solo accuratezze, niente caso per caso	87	Turing di altre macchine
Dieci reti fra loro	Probabilità per immagine, cioè tutti gli indizi	2.615	Formula, una mente del 2014
Esperimenti finti	Generati dal modello, con la verità nota	1.350	Taratura
Persone di silicio	Sessanta reti piccole che imparano in tre gruppi	3 gruppi	Copia degli errori

I morsetti. Il confronto fra la macchina da sola e quella della conversazione tiene solo le risposte che sono una delle quattro opzioni. Contando come sbagliate anche le altre, con Llama la macchina della conversazione scende al 79 per cento e quella da sola, sulle stesse prove, al 71; con GPT-4o la distanza di un punto si rovescia.

I limiti dei dati, nel dettaglio. Il combinatore a verdetto è un'approssimazione di una coppia vera e funziona con ciò che le persone hanno dichiarato: una regressione che pesa i verdetti non è una persona che discute. Su HAIID i casi per compito sono 24-32 e gli intervalli sono larghi: in 30 combinazioni di modello e compito su 64 la macchina resta sotto il caso, nel punto o in un ricampionamento, e il confronto non si fa; dove si fa, la sua sensibilità tocca spesso il massimo che così pochi casi consentono. Il «consiglio» che HAIID mostrava come dell'AI era in realtà la media di una cinquantina di persone: lo trattiamo come controllo, non come macchina. Sugli oggetti, nella prova della formula restano 673 confronti su 689, perché negli altri il combinatore non supera il caso. Infine, la scala è di famiglia: dentro il nostro insieme di macchine, «più forte» si mescola con «più recente» e «di un altro laboratorio», e il disegno non separa queste tre cose.

Codice e dati. Protocollo, deviazioni, codice e dati derivati sono nel repository dell'esperimento; i dati di ChaosNLI e Vaccaro non si ridistribuiscono: ChaosNLI lo scarica un programma, la tabella di Vaccaro si prende dal repository OSF degli autori.

Per approfondire

- Serina Chang, Ashton Anderson e Jake M. Hofman, *ChatBench: From Static Benchmarks to Human-AI Evaluation*, 2025.

- Yixin Nie, Xiang Zhou e Mohit Bansal, *What Can We Learn from Collective Human Opinions on Natural Language Inference Data?* (ChaosNLI), EMNLP 2020.
- Kailas Vodrahalli, Roxana Daneshjou, Tobias Gerstenberg e James Zou, *Do Humans Trust Advice More if it Comes from AI? An Analysis of Human-AI Interactions* (HAIID), AIES 2022.
- Robert Geirhos e colleghi, *Partial success in closing the gap between human and machine vision*, NeurIPS 2021.
- Robert Geirhos, Kristof Meding e Felix A. Wichmann, *Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency*, NeurIPS 2020.
- Michelle Vaccaro, Abdullah Almaatouq e Thomas Malone, *When combinations of humans and AI are useful: A systematic review and meta-analysis*, *Nature Human Behaviour* 8, 2024.
- Joel Becker, Nate Rush, Elizabeth Barnes e David Rein (METR), *Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity*, 2025.
- Dingwen Kong, Liang Lyu, Saurabh Amin, Amine Bennouna, Daniel Huttenlocher e Asuman Ozdaglar, *AI-Assisted Classification under Correlation Neglect and Trust*, 2026.
- Robert D. Sorkin e Huanping Dai, *Signal detection analysis of the ideal group*, *Organizational Behavior and Human Decision Processes* 60(1), 1994.
- Irving John Good, *A. M. Turing's Statistical Work in World War II*, *Biometrika* 66(2), 1979, pp. 393-396.
- Irving John Good, *Weight of Evidence: A Brief Survey*, in *Bayesian Statistics 2*, 1985, pp. 249-270.
- W. W. Peterson e T. G. Birdsall, *The Theory of Signal Detectability*, rapporto tecnico n. 13, Electronic Defense Group, University of Michigan, 1953.
- W. W. Peterson, T. G. Birdsall e W. C. Fox, *The theory of signal detectability*, *Transactions of the IRE Professional Group on Information Theory* 4(4), 1954.
- Wilson P. Tanner Jr. e Theodore G. Birdsall, *Definitions of d′ and η as Psychophysical Measures*, *The Journal of the Acoustical Society of America* 30(10), 1958.
- Neil A. Macmillan e Howard L. Kaplan, *Detection theory analysis of group data: Estimating sensitivity from average hit and false-alarm rates*, *Psychological Bulletin* 98(1), 1985.
- Donald D. Dorfman, Kevin S. Berbaum e Charles E. Metz, *Receiver Operating Characteristic Rating Analysis: Generalization to the Population of Readers and Patients with the Jackknife Method*, *Investigative Radiology* 27(9), 1992.
- Art B. Owen, *The pigeonhole bootstrap*, *The Annals of Applied Statistics* 1(2), 2007, pp. 386-411.
- Art B. Owen e Dean Eckles, *Bootstrapping data arrays of arbitrary order*, *The Annals of Applied Statistics* 6(3), 2012.
- Marc O. Ernst e Martin S. Banks, *Humans integrate visual and haptic information in a statistically optimal fashion*, *Nature* 415, 2002.
- Bahador Bahrami e colleghi, *Optimally Interacting Minds*, *Science* 329, 2010.
- Lucía Vicente e Helena Matute, *Humans inherit artificial intelligence biases*, *Scientific Reports* 13, 2023.
- Lucía Vicente, Helena Matute, Caterina Fregosi e Federico Cabitza, *Machine learning systems as mentors in human learning: A user study on machine bias transmission in medical training*, *International Journal of Human-Computer Studies* 198, 2025.
- Mark Steyvers, Heliodoro Tejeda, Gavin Kerrigan e Padhraic Smyth, *Bayesian modeling of human–AI complementarity*, *PNAS* 119(11), 2022.

- Ada Lovelace, *Sketch of the Analytical Engine invented by Charles Babbage* di L. F. Menabrea, *with notes upon the Memoir by the Translator*, nota G, 1843.
- Alan M. Turing, *Computing Machinery and Intelligence*, *Mind* 59(236), 1950.

Come citare. endjneer (2026). *Intelligenza Apparente.* paithon.it, versione 2, 11 novembre 2026. https://doi.org/10.5281/zenodo.24141588

Codice e dati: https://github.com/paithon-it/turing-meter.