
La previsione che misura la cosa sbagliata

Cinque paper di Taleb, i benchmark dei modelli, i ricavi annualizzati dei laboratori, i credit default swap sui data center, la scrittura dei paper e il registro delle previsioni sull'AGI: undici schede su numeri che si lasciano verificare e misurano un'altra cosa.

Claudio Cammarano 27 ottobre 2026

FONTI	11 schede, mar 2016 – ott 2026
ARGOMENTI	Epistemologia · Scienza · Economia
CONCETTI	cigni neri · legge di Goodhart · Statistical Consequences of Fat Tails · Taleb, Nassim Nicholas · test severo
ORIGINE	https://claudiocammarano.com/writings/2026-10-27-la-previsione-che-misura-la-cosa-sbagliata/

TESI

Una previsione o un indicatore possono essere perfettamente verificabili e misurare comunque la cosa sbagliata: la frequenza con cui si indovina al posto di quanto costa sbagliare, la posizione in una gara al posto del valore, il ricavo di un mese moltiplicato per dodici al posto del ricavo. Il controllo che conta va fatto prima dell'esito, e chiede che cosa il numero misuri davvero.

Nel 2016 diversi gruppi che prevedevano le elezioni americane pubblicavano probabilità che salivano e scendevano a ogni sondaggio. Nassim Nicholas Taleb, in un paper uscito l'anno dopo, mostrò che quelle previsioni si potevano giudicare prima del voto, senza aspettare chi avrebbe vinto: bastava guardare come si muovevano. La catena che si apre qui riprende il filo di «Non è nemmeno sbagliato», dove il difetto delle previsioni sull'intelligenza artificiale era di non lasciarsi smentire, e lo porta un passo più in là. Undici schede descrivono numeri che si lasciano verificare benissimo e che misurano un'altra cosa rispetto a quella per cui vengono usati. Il test severo, che in «Il test severo» chiedeva quanta probabilità avesse una prova di smentire, qui deve rispondere a una domanda che viene prima: che cosa misura il numero che si sta controllando.

La previsione che si contraddice

L'argomento di Taleb sulle elezioni è tecnico e ha una conseguenza semplice.¹ Una stima binaria aggiornata di continuo deve comportarsi come una martingala, cioè non deve contenere al suo interno una tendenza prevedibile, e allora si può prezzare come un'opzione binaria. Ne viene un vincolo che funziona senza l'esito: quando l'incertezza è alta il valore sta vicino al cinquanta per cento e non dovrebbe muoversi molto, e una previsione che oscilla mentre la corsa resta incerta si sta contraddicendo da sola. Taleb mostra che diversi gruppi, nel 2016, violavano queste relazioni.

1 · Nassim Nicholas Taleb / arXiv · 18 mar 2017 · [STRUMENTO CONCETTUALE](#)

Election Predictions as Martingales: An Arbitrage Approach

Una previsione si può giudicare prima che l'esito arrivi: se salta a ogni sondaggio mentre la corsa resta incerta, sta già dicendo di non essere una stima di probabilità.

<https://arxiv.org/abs/1703.06351>

Per la serie in cui questa catena entra è un acquisto raro, perché quasi ogni controllo di una previsione arriva dopo, quando ormai si sa com'è andata, e a quel punto un risultato azzeccato copre qualunque metodo. Qui il controllo arriva prima, e guarda la forma della previsione invece del suo contenuto.

Indovinare e incassare

Due anni dopo Taleb sposta la domanda dalla coerenza alla cosa misurata.² Una previsione binaria dice se un evento accadrà; chi la usa per decidere incassa un esito continuo, che dipende da quanto l'evento pesa quando accade. Fra le due cose c'è una distribuzione intera, e sotto certe distribuzioni un buon previsore binario non è un buon indicatore: può esserlo all'inverso. Da qui Taleb attacca la letteratura psicologica sulla sopravvalutazione degli eventi rari, che a suo dire misura la confusione fra i due oggetti più che un errore dei soggetti.

2 · Nassim Nicholas Taleb / arXiv · 24 lug 2019 · [STRUMENTO CONCETTUALE](#)

On the Statistical Differences between Binary Forecasts and Real World Payoffs

Un record di previsioni azzeccate non dice come è andata a chi quelle previsioni le ha usate per decidere: fra indovinare la direzione e incassare il risultato c'è una distribuzione di mezzo.

<https://arxiv.org/abs/1907.11162>

Il paper del 2023, scritto con Ron Richman, Marcos Carreira e James Sharpe come replica a Philip Tetlock, porta la distinzione sui tornei di previsione, quelli che premiano chi stima meglio la probabilità di un evento.³ Sapere quanto è probabile un evento non dice niente su quanto costa, e il torneo misura soltanto la prima delle due cose; il giudizio esperto, sostengono gli autori, non cattura il rischio di coda, e il formato del torneo è incompatibile con gli

strumenti che lo misurano. La distinzione si porta su qualunque frase della forma «c'è il tot per cento di probabilità che», ed è una frase che i giornali scrivono ogni giorno.

Un numero solo

Su una variabile a coda spessa, un numero solo è una categoria sbagliata prima che una stima imprecisa. Taleb lo scrive nel luglio 2020 con Yaneer Bar-Yam e Pasquale Cirillo, rispondendo a John Ioannidis sulla pandemia: la media non è il valore che si incontra, e sotto code spesse aspettare l'evidenza prima di agire è una scommessa rischiosa quanto agire senza.⁴ La posizione che sembra prudente, chiedere più dati, è anch'essa una decisione con un costo, e nessuno la dichiara come tale.

4 · Nassim Nicholas Taleb, Yaneer Bar-Yam e Pasquale Cirillo / arXiv · 31 lug 2020 · [DATO CHE CORREGGE](#)

On Single Point Forecasts for Fat-Tailed Variables

Richiedere evidenza prima di agire sembra sempre la posizione prudente, ma su una variabile a coda spessa è una scommessa come un'altra, e nessuno la dichiara come tale.

<https://arxiv.org/abs/2007.16096>

Lo stesso vale per il numero più citato della scienza sperimentale. Nel paper del 2016 sul p-hacking Taleb mostra che la distribuzione del p-value è così asimmetrica che un singolo valore resta volatile a prescindere dalla dimensione del campione, anche quando nessuno ha barato.⁵ La prima catena di questa serie aveva ricostruito che cosa il p-value non dice; questo paper aggiunge che quel poco che dice lo dice in modo instabile.

La gara al posto del termometro

Fuori dalla statistica il meccanismo si vede meglio. Moritz Hardt, intervistato da John Pavlus su *Quanta*, chiama i benchmark «l'errore che ha fatto l'apprendimento automatico».⁶ Dovevano funzionare come termometri, e il campo li ha usati come gare: li ha messi in rete e ci ha iterato sopra, contro la regola di usarli una volta sola, finché basta cambiare di poco il test per passare dal 90 al 73 per cento. I punteggi assoluti valgono poco. Le classifiche invece sono rimaste stabili fra un dataset e l'altro, «mostly by sheer luck», e solo finché tutti si preparavano al test nello stesso modo; con i modelli linguistici, di cui non si sa su che cosa siano addestrati, quella condizione è saltata.

6 · John Pavlus / Quanta Magazine · 16 ott 2026 · [STRUMENTO CONCETTUALE](#)

He Wants AI Benchmarking To Be a Science, Not a Dark Art

I numeri dei benchmark non misurano quanto vale un modello, dicono solo chi ha battuto chi, e oggi nemmeno quello, perché ognuno si prepara al test in misura diversa.

<https://www.quantamagazine.org/he-wants-ai-benchmarking-to-be-a-science-not-a-dark-art-20261016/>

Un benchmark verificato misura con precisione la posizione in una gara, e viene letto come una misura del valore. La proposta di Hardt, presa dalle Olimpiadi di Città del Messico del 1968, è estendere a tutti l'allenamento in quota invece di vietarlo; lui stesso riconosce che così si ottengono indicatori di progresso, e che misure vere e proprie non ce ne sono.

Numeri che non si confrontano

La colonna Schumpeter dell'*Economist* trova lo stesso difetto nei conti dei laboratori.⁷ In una settimana i ricavi di OpenAI sono stati stimati in 70 miliardi di dollari da Axios e in 50 dal *Financial Times*, e i titoli tecnologici sono saliti e scesi a ogni versione. I ricavi di cui si parla sono ricavi ricorrenti annualizzati, le vendite di un mese moltiplicate per dodici, e ogni laboratorio li calcola a modo suo: Anthropic conta le vendite fatte attraverso i fornitori cloud, OpenAI solo dopo averne tolto la quota. C'è chi propone di valutarli sul margine prima dei costi di addestramento, cioè prima della spesa più grande, e la colonna ricorda l'«EBITDA rettificato per la comunità» di WeWork.

7 · The Economist (Schumpeter) · 14 ott 2026 · LETTURA DI MERCATO

The AI boom is threatened by an epidemic of unreliable numbers

Il boom dell'AI si misura con numeri che non si possono confrontare: ricavi annualizzati costruiti in modi diversi, margini calcolati prima della spesa principale, impegni fuori bilancio. Un mercato che vola alla cieca.

<https://www.economist.com/business/2026/10/14/the-ai-boom-is-threatened-by-an-epidemic-of-unreliable-numbers>

Matt Levine, su *Bloomberg*, mostra che cosa succede quando il mercato prova a coprirsi contro quei numeri.⁸ Gli scambi di credit default swap legati ai bond di Oracle, Alphabet, Amazon, Meta, Nvidia e degli altri grandi costruttori sono passati, secondo Morgan Stanley, da meno di due a più di nove miliardi di dollari a settimana fra luglio e settembre, e assicurare il debito di Oracle costa il 3,3 per cento l'anno contro lo 0,33 di cinque anni fa. Il prezzo della protezione misura un rischio specifico, quello che un debitore fallisca; il rischio dei data center, scrive Levine, è quasi tutto sistematico, e chi dovrebbe vendere l'assicurazione ha già in portafoglio gli stessi bond. Lo strumento funziona e misura un rischio diverso da quello che c'è.

Il segnale rovesciato

A volte un indicatore smette di misurare senza che nessuno lo tocchi. L'*Economist*, sulla burocrazia della ricerca, registra un rovesciamento: prima dei modelli linguistici la qualità della scrittura di un paper era un buon indizio della sua qualità scientifica, e oggi non lo è più.⁹ Le sottomissioni a ICML sono raddoppiate in un anno, a 24.000; in 2,5 milioni di articoli biomedici le citazioni inventate sono cresciute di dodici volte fra il 2023 e l'inizio del 2026; due ricercatori svedesi hanno messo online due preprint su una malattia degli occhi inesistente, la «bixonimania», e in poche settimane i chatbot ne parlavano come di una cosa reale. Il revisore che leggeva la prosa come segno del lavoro continua a leggere la stessa prosa, e il segno ha cambiato significato sotto i suoi occhi.

Le previsioni come fatti

Luciano Floridi, Jessica Morley e Claudio Novelli, nel secondo capitolo di *Not Even Wrong*, hanno verificato una per una ventiquattro previsioni pubbliche sull'AGI, la Singolarità e l'estinzione, emesse fra il 1950 e il 2026.¹⁰ La parte che conta per questa catena è il precedente da cui partono. Nel 1973 Sir James Lighthill misurò il registro dell'intelligenza artificiale contro le promesse della generazione fondatrice e concluse che aveva promesso troppo, e il rapporto contribuì al taglio dei finanziamenti britannici negli anni Settanta. Le previsioni, ne ricavano gli autori, sono fatti istituzionali: conferiscono legittimità, orientano le politiche e gli investimenti, e così contribuiscono a determinare il futuro che dicono di descrivere. Un numero sbagliato sulla cosa giusta fa danni; un numero preciso sulla cosa sbagliata li fa con l'autorità della precisione.

Apparato · Che cosa misura il numero

Numero

Viene letto come

Misura

Scheda

Probabilità elettorale aggiornata

Stima di chi vincerà

Se oscilla mentre la corsa è incerta, la propria incoerenza

Taleb 2017

Record di previsioni azzeccate

Bravura del previsore

La direzione, non quanto si incassa

Taleb 2019

Punteggio in un torneo di previsione

Capacità di valutare il rischio

La probabilità, non il costo dell'evento

Taleb e altri 2023

Previsione puntuale su una variabile a coda spessa

Stima prudente

Una media che non si incontra

Taleb, Bar-Yam, Cirillo 2020

Punteggio di un benchmark

Valore del modello

La posizione in una gara

Hardt

Ricavi annualizzati

Ricavi

Un mese per dodici, calcolato in modi diversi

Economist

Prezzo dei credit default swap

Rischio dei data center

Un rischio specifico, dove il rischio è sistematico

Levine

Qualità della scrittura di un paper

Qualità della ricerca

Dopo i modelli linguistici, quasi nulla

Economist

Controtesi 11 • Sarah Mervosh, Francesca Paris e Claire Cain Miller / The New York Times

Si può rispondere che il difetto riguarda numeri costruiti male, e che un numero costruito bene misura quello che deve. Sarah Mervosh, Francesca Paris e Claire Cain Miller, sull'*Upshot* del *New York Times*, hanno chiesto a 37 studiosi dell'istruzione di valutare trenta interventi scolastici su due scale separate, l'impatto sull'apprendimento e il costo, assumendo ogni volta che l'intervento fosse realizzato bene.¹¹ Il risultato attraversa lo schieramento politico: i voucher, bandiera della destra, e le classi più piccole, care alla sinistra, finiscono entrambi fra i cinque interventi peggiori per costo-efficacia. Il numero c'è, si lascia discutere e misura la cosa giusta.

L'obiezione è giusta, e il caso che porta conferma la tesi invece di smentirla, perché fa esattamente ciò che Taleb chiede ai tornei di previsione: tiene separata la probabilità di funzionare dal costo, invece di fonderle in un punteggio solo. Le autrici, inoltre, dichiarano che cosa il numero non è, «una misura utile di come i ricercatori esperti vedono queste idee» e non un sondaggio scientifico in senso stretto, cioè dicono per prime che cosa stanno misurando. È il controllo che manca a tutti gli altri casi di questa catena, e si può fare prima di conoscere l'esito: chiedere a un numero di dichiarare l'oggetto, le condizioni in cui vale e la distanza fra ciò che misura e ciò per cui verrà usato. Una previsione che regge questo controllo può ancora risultare sbagliata, e allora almeno avrà sbagliato sulla cosa giusta.

Nella catena

1. Nassim Nicholas Taleb / arXiv, *Election Predictions as Martingales: An Arbitrage Approach*, 18 mar 2017. Strumento concettuale.
<https://arxiv.org/abs/1703.06351>
Scheda: <https://claudiocammarano.com/curated/2017-03-18-taleb-election-predictions-as-martingales-arxiv/>
2. Nassim Nicholas Taleb / arXiv, *On the Statistical Differences between Binary Forecasts and Real World Payoffs*, 24 lug 2019. Strumento concettuale.
<https://arxiv.org/abs/1907.11162>
Scheda: <https://claudiocammarano.com/curated/2019-07-24-taleb-binary-forecasts-payoffs-arxiv/>
3. Nassim Nicholas Taleb, Ron Richman, Marcos Carreira e James Sharpe / arXiv, *The Probability Conflation: A Reply*, 26 gen 2023. Strumento concettuale.

<https://arxiv.org/abs/2301.10985>

Scheda: <https://claudiocammarano.com/curated/2023-01-26-taleb-probability-conflation-arxiv/>

4. Nassim Nicholas Taleb, Yaneer Bar-Yam e Pasquale Cirillo / arXiv, *On Single Point Forecasts for Fat-Tailed Variables*, 31 lug 2020. Dato che corregge.
<https://arxiv.org/abs/2007.16096>
Scheda: <https://claudiocammarano.com/curated/2020-07-31-taleb-single-point-forecasts-arxiv/>
5. Nassim Nicholas Taleb / arXiv, *A Short Note on P-Value Hacking*, 24 mar 2016. Dato che corregge.
<https://arxiv.org/abs/1603.07532>
Scheda: <https://claudiocammarano.com/curated/2016-03-24-taleb-p-value-hacking-arxiv/>
6. John Pavlus / Quanta Magazine, *He Wants AI Benchmarking To Be a Science, Not a Dark Art*, 16 ott 2026. Strumento concettuale.
<https://www.quantamagazine.org/he-wants-ai-benchmarking-to-be-a-science-not-a-dark-art-20261016/>
Scheda: <https://claudiocammarano.com/curated/2026-10-16-pavlus-hardt-benchmark-quanta/>
7. The Economist (Schumpeter), *The AI boom is threatened by an epidemic of unreliable numbers*, 14 ott 2026. Lettura di mercato.
<https://www.economist.com/business/2026/10/14/the-ai-boom-is-threatened-by-an-epidemic-of-unreliable-numbers>
Scheda: <https://claudiocammarano.com/curated/2026-10-14-economist-numeri-inaffidabili-ai/>
8. Matt Levine / Bloomberg Opinion, *Someone Has to Take the AI Risk*, 20 ott 2026. Lettura di mercato.
<https://www.bloomberg.com/opinion/newsletters/2026-10-20/someone-has-to-take-the-ai-risk>
Scheda: <https://claudiocammarano.com/curated/2026-10-20-levine-someone-has-to-take-ai-risk-bloomberg/>
9. The Economist, *Could AI help scientists with their paperwork?*, 14 ott 2026. Dato che corregge.
<https://www.economist.com/science-and-technology/2026/10/14/could-ai-help-scientists-with-their-paperwork>
Scheda: <https://claudiocammarano.com/curated/2026-10-14-economist-ai-burocrazia-scienza/>
10. Luciano Floridi, Jessica Morley e Claudio Novelli / SSRN, *Not Even Wrong 2: An Audit of Public AGI Prediction, 1950–2026*, 14 set 2026. Strumento concettuale.
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=7456700
Scheda: <https://claudiocammarano.com/curated/2026-09-14-floridi-not-even-wrong-2-agi-ssrn/>
11. Sarah Mervosh, Francesca Paris e Claire Cain Miller / The New York Times, *30 Ideas for Fixing U.S. Schools, Ranked by Experts*, 28 set 2026. Dato che corregge.
<https://www.nytimes.com/2026/09/28/upshot/ideas-fixing-schools-ranked.html>
Scheda: <https://claudiocammarano.com/curated/2026-09-28-mervosh-paris-miller-cosa-funziona-davvero-nella-scuola-nyt/>

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

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

Versione originale e aggiornata: <https://claudiocammarano.com/writings/2026-10-27-la-previsione-che-misura-la-cosa-sbagliata/>. Le parti interattive della catena sono disponibili solo online.

Questa catena fa parte della serie *Lo standard di prova*: <https://claudiocammarano.com/serie/lo-standard-di-prova/>.

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