

Inferenza. Breve storia di un concetto tra logica, linguistica e intelligenza artificiale

Luca Cignetti (Scuola Universitaria Professionale della Svizzera Italiana)

submitted 30 March 2026, accepted 1 October 2026, published 26 October 2026

<https://doi.org/10.62408/ai-ling.v6i1.81>

luca.cignetti(at)supsi.ch

Abstract

Il contributo propone una breve ricostruzione storico-concettuale del termine “inferenza”, confrontandone le principali accezioni in logica, in statistica, in linguistica e nell’intelligenza artificiale. Muovendo dalla centralità dell’inferenza nei sistemi di IA – dove indica il processo di generazione di output a partire da dati e modelli – il contributo evidenzia il rischio di un impiego dello stesso termine in riferimento a processi differenti. A partire dalla logica aristotelica, dove il ragionamento sillogistico è fondato sulla derivazione della conclusione dalle premesse secondo una relazione necessaria, si procede verso una concezione probabilistica, fondata sulla stima dell’incertezza. In ambito computazionale, ci si sofferma sul passaggio da modelli deduttivi a modelli probabilistici e predittivi, in cui l’inferenza coincide con la stima statistica o la generazione sequenziale. In linguistica e nelle scienze cognitive, l’inferenza è invece intesa come processo interpretativo che consente di costruire significati impliciti integrando informazioni contestuali e conoscenze enciclopediche. L’uso di uno stesso termine per processi eterogenei richiede pertanto una distinzione concettuale che eviti di attribuire ai sistemi di IA, sulla base di una analogia terminologica, proprietà inferenziali proprie del ragionamento umano.

The article offers a brief historical reconstruction of the term “inference”, comparing its main meanings in logic, statistics, linguistics, and artificial intelligence. Starting from the central role of inference in AI systems – where it refers to the process of generating outputs from data and models – the paper highlights the risk of using the same term to denote fundamentally different processes. Beginning with Aristotelian logic, in which syllogistic reasoning is based on deriving a conclusion from premises through a necessary relation, the discussion then traces its development toward a probabilistic conception based on the estimation of uncertainty. In the computational domain, particular attention is paid to the shift from deductive models to probabilistic and predictive ones, in which inference coincides with statistical estimation or sequential generation. In linguistics and the cognitive sciences, by contrast, inference is conceived as an interpretive process that enables the construction of implicit meanings by integrating contextual information and encyclopedic knowledge. The use of the same term for heterogeneous processes therefore calls for a conceptual distinction that avoids attributing to AI systems, on the basis of a terminological analogy, inferential properties characteristic of human reasoning.

Keywords

Inferenza; deduzione; implicatura; presupposizione; schemi di Winograd.

Inference; deduction; implicature; presupposition; Winograd schemas.



Cignetti, Luca. 2026. Inferenza. Breve storia di un concetto tra logica, linguistica e intelligenza artificiale.

AI-Linguistica, Vol. 6 No. 1

DOI: 10.62408/ai-ling.v6i1.81

AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses

ISSN: 2943-0070

© 2026 Luca Cignetti, licensed under CC BY-NC-SA 4.0.

1 Introduzione

Nel campo dell'Intelligenza Artificiale (IA), il termine “inferenza” designa l'insieme delle operazioni mediante le quali un sistema genera output a partire da dati di input, modelli appresi o rappresentazioni formalizzate della conoscenza. In questa accezione, l'inferenza costituisce il nucleo operativo dei sistemi di IA, poiché media tra la fase di acquisizione o apprendimento dei dati e la produzione di risultati potenzialmente rilevanti, come previsioni, raccomandazioni, decisioni o contenuti. La centralità della dimensione inferenziale nei sistemi di IA è riconosciuta esplicitamente nell'*Artificial Intelligence Act*, la prima regolamentazione organica dell'Unione Europea in materia, che individua in questa caratteristica una proprietà costitutiva dei sistemi intelligenti:

Una caratteristica fondamentale dei sistemi di IA è la loro capacità inferenziale. Tale capacità inferenziale si riferisce al processo di ottenimento degli output, quali previsioni, contenuti, raccomandazioni o decisioni, che possono influenzare gli ambienti fisici e virtuali, nonché alla capacità dei sistemi di IA di ricavare modelli o algoritmi, o entrambi, da input o dati. Le tecniche che consentono l'inferenza nella costruzione di un sistema di IA comprendono approcci di apprendimento automatico che imparano dai dati come conseguire determinati obiettivi e approcci basati sulla logica e sulla conoscenza che traggono inferenze dalla conoscenza codificata o dalla rappresentazione simbolica del compito da risolvere. La capacità inferenziale di un sistema di IA trascende l'elaborazione di base dei dati consentendo l'apprendimento, il ragionamento o la modellizzazione. (Unione Europea, 2024, 12).

In linguistica e nelle scienze cognitive “inferenza” designa invece un processo interpretativo mediante il quale un soggetto ricava contenuti impliciti a partire da informazioni esplicite, mobilitando conoscenze enciclopediche, schemi cognitivi e assunzioni contestuali. “Inferire”, in questa prospettiva, richiede di colmare lacune informative, stabilire relazioni di coerenza tra porzioni testuali, attribuire intenzioni comunicative e costruire rappresentazioni semantiche del discorso (de Beaugrande & Dressler 1984; Kintsch, 1998; Cignetti, 2021; Cignetti et al., 2022).

L'impiego della voce “inferenza” per designare tanto processi computazionali automatizzati quanto operazioni proprie del ragionamento umano rischia dunque di attenuare, se non di oscurare, le differenze che separano fenomeni di natura diversa. Con questo contributo ci si propone un confronto critico tra alcune delle principali accezioni di “inferenza” nella tradizione logico-filosofica, linguistica e computazionale, partendo dalla ricostruzione delle tappe fondamentali nella storia del concetto e del termine. Si intende così mostrare come l'adozione di “inferenza” come nozione trasversale nel campo dell'IA, pur risultando funzionale sul piano descrittivo, sollevi questioni teoriche ancora aperte e renda auspicabile una maggiore precisione terminologica.

2 Le origini e le prime attestazioni

La prima sistematizzazione del concetto di inferenza si trova nella logica classica, nella teoria aristotelica del sillogismo. In questo quadro, “inferire” significa stabilire un nesso di necessità che collega le premesse alla conclusione. In tale rapporto formale risiede la validità dell’inferenza: se le premesse sono vere e la struttura argomentativa è corretta, la conclusione ne deriva per necessità. Così nel classico sillogismo “Tutti gli uomini sono mortali. Socrate è un uomo. Dunque Socrate è mortale” la deduzione inferenziale consente di derivare la conclusione a partire da una premessa universale e da una premessa particolare.

Dal punto di vista etimologico, “inferenza” deriva dal latino *inferre* (“portare dentro”, “trarre come conseguenza”), verbo che esprime, con una metafora spaziale, l’idea di un movimento argomentativo attraverso cui da un primo enunciato ne viene ricavato un secondo. Nel latino classico, tuttavia, questa operazione non è ancora stabilizzata come forma sostantivata autonoma, per quanto il verbo *inferre* sia già attestato nel contesto della riflessione logica e dell’argomentazione retorica per indicare l’atto di “trarre una conseguenza”¹. Il verbo assume un impiego più sistematico e tecnicamente marcato nel latino tardo, in particolare nella trattazione delle diverse forme del sillogismo e dei procedimenti dimostrativi. In questi ambiti, *inferre* designa l’operazione attraverso cui una proposizione viene legittimamente derivata da altre secondo regole di natura logico-argomentativa².

Nel latino scolastico medievale si trovano le prime attestazioni del deverbale *inferentia*. Abelardo usa il termine per descrivere la propria teoria delle

¹ Cfr. Cicerone (2015: I, 73): “*Nam aut ita complectemur, ut in unum conducamus propositionem et assumptionem, hoc modo: ‘Quodsi leges omnes ad utilitatem rei publicae referri convenit, hic autem saluti rei publicae profuit, profecto non potest eodem facto et saluti communi consuluisse et legibus non optemperasse’; aut ita, ut ex contrario sententia conficiatur, hoc modo: ‘Summa igitur amentia est in eorum fide spem habere, quorum perfidia totiens deceptus sis’; aut ita, ut id solum, quod conficitur, inferatur, ad hunc modum: ‘Urbem igitur diruamus’*” [‘Infatti, o concluderemo in modo da fondere in una sola proposizione la premessa maggiore e la premessa minore, in questo modo: «Se conviene che tutte le leggi siano riferite all’utilità dello Stato e costui ha giovato alla salvezza dello Stato, certamente non può aver provveduto, con la stessa azione, alla salvezza comune e non aver obbedito alle leggi»; oppure concluderemo a partire da una proposizione contraria, in questo modo: «È dunque somma follia riporre la speranza nella fedeltà di coloro dalla cui perfidia si è stati tante volte ingannati»; oppure inferiremo soltanto ciò che si conclude, in questo modo: «Distruggiamo dunque la città»’].

² Cfr. Capella (1836: IV, 411): “*Si reflexim inferas: ‘omne igitur bonum iustum’, non sequitur, sed particulariter potest inferri: ‘quoddam igitur bonum iustum’*” [‘Se inferisci invertendo il ragionamento «dunque ogni bene è giusto», la conclusione non segue; si può invece inferire in forma particolare «dunque un certo bene è giusto»’]; Boethius (2000: III, 10, 22): “*Super haec igitur ueluti geometrae solent demonstratis propositis aliquid inferre quod illi porismata uocant, ita ego quoque tibi ueluti corollarium dabo*” [‘Pertanto, come i geometri sono soliti inferire qualcosa, che chiamano porismi, dopo aver dimostrato le proposizioni, così anch’io ti offrirò, per così dire, un corollario’].

consequentiae, attribuendo all'inferenza uno statuto di necessità anche nei casi in cui la struttura dell'argomentazione risulti imperfetta³.

Il passaggio dal latino alle lingue volgari europee non segue traiettorie pienamente lineari e riflette percorsi lessicali e concettuali diversi. In italiano, le prime attestazioni del verbo “inferire” risalgono alla fine del XIV secolo⁴. Il verbo si afferma progressivamente tra XVI e XVIII secolo nel lessico filosofico e scientifico, con attestazioni, tra gli altri, in Machiavelli, Galileo e Francesco Algarotti⁵. Più tardi è l'impiego del sostantivo “inferenza”: la prima attestazione nota risale al 1846, nei *Prolegomeni del primato morale e civile degli italiani* di Vincenzo Gioberti⁶.

In francese, il verbo *inférer* è attestato già nel XIV secolo e viene progressivamente impiegato nel senso di «tirer une conséquence»; il TLFi ne documenta questo valore a partire dal 1372⁷. Quanto al sostantivo *inférence*, ne colloca invece la prima attestazione nel 1609⁸.

Quanto al sostantivo *inférence*, il *Dictionnaire historique de la langue française* ne colloca la prima attestazione nel 1606. L'inglese accoglie invece *inference* alla fine del XVI secolo: l'Oxford English Dictionary ne registra una delle prime attestazioni nel trattato *Of the Lawes of Ecclesiasticall Politie* (1593) di Richard Hooker.

3 L'inferenza statistica

La progressiva stabilizzazione, nelle lingue europee moderne, di forme lessicali atte a designare il concetto di inferenza come passaggio dalle premesse a una conclusione costituisce il presupposto terminologico per il successivo impiego di tali forme nell'ambito del ragionamento scientifico e probabilistico.

A partire dall'età moderna, l'inferenza inizia a designare forme di ragionamento che non riguardano più esclusivamente la necessità logica della conclusione rispetto alle premesse, ma anche il grado di attendibilità con cui una conclusione può essere sostenuta sulla base di dati empirici. In questo passaggio, si introduce progressivamente una dimensione probabilistica del ragionamento che

³ Cfr. Abelardo, *Dialec.*, III: “*Similiter enim et in istis qui imperfecti dicuntur, inferentia est necessaria sicut in illis*” [“Allo stesso modo, anche in quelli [argomenti] che sono detti imperfetti, l'inferenza è necessaria come in quelli perfetti”].

⁴ Cfr. Lapo da Castiglionchio, *Epist.*, 24: “Per le predette cose molte cose s'inferiscono”.

⁵ Cfr. N. Machiavelli, *Leg. Comm.* 3, 42: “Volle quasi inferire che, a quesia parte dell'assoldar gente italiana, avevano fatta la loro parte, e che toccava agli amici loro fare il resto”; G. Galilei, *Opere Astronom.*, vol. 4, p. 182: “Dall'apparente moto retto inferivano la cometa muoversi per cerchio Massimo”; F. Algarotti, *Op. Sc.*, vol. I, p. 201: “Omero, Ennio e Dante precedettero Erodoto, Cicerone e il Boccaccio; donde si può inferire che l'arte dello scrivere in versi sia più facile dell'arte della prosa”.

⁶ Cfr. V. Gioberti, *Proleg.*, vol. 2, p. 230 “In ogni qualunque argomento occorrono certe generalità, inferenze e osservazioni accessorie, comuni a molti o a tutti gli autori”.

⁷ Cfr. TLFi, s.v. *inférer*: “1372 “tirer une conséquence” (Denis Foulechat, trad. du *Policraticus* de Jean de Salisbury, III, 8, 48 ds R. *Ling. rom.* t. 33, p. 322)”.

⁸ Cfr. TLFi, s.v. *inférence*: “1609 ‘conséquence’ (G. Vittori, *Tesoro de las tres lenguas*, Genève)”.

amplia il dominio dell'inferenza oltre i confini della sola deduzione formale (Hacking, 1975; Desrosières, 1998; Gigerenzer et al., 1989). Pur continuando a indicare un movimento cognitivo che procede dall'osservato al non osservato, il vincolo tra premesse e conclusione non è ora più garantito esclusivamente dalla validità logica, bensì dalla frequenza, dalla regolarità e dalla stabilità dei dati empirici.

Il momento chiave di questa trasformazione si colloca all'inizio del XVIII secolo, quando Jakob Bernoulli, nella sua *Ars Conjectandi* (1713), presenta una nuova concezione dell'inferenza come strumento di stima, fondato sull'osservazione empirica e sulla stabilità statistica dei fenomeni, offrendo un contributo decisivo alla transizione dalla concezione deduttiva a una concezione probabilistica dell'inferenza (Hald, 2003). In questo contesto, Bernoulli impiega anche forme del verbo *inferre*, ma ciò non autorizza ancora a identificare il suo uso con il concetto moderno di inferenza statistica: la terminologia centrale dell'opera resta quella della *coniectura* e dell'*ars conjectandi*. Il passo in cui Bernoulli scrive "*sic inferri potest*" documenta tuttavia la continuità lessicale tra il verbo latino *inferre* e il successivo sviluppo del termine *inferenza* in ambito probabilistico⁹.

Nel corso del XVIII secolo, tale prospettiva evolve con il lavoro di Thomas Bayes, che introduce un principio che consente di determinare la probabilità di un'ipotesi alla luce di un evento osservato, stabilendo un nesso matematico tra probabilità *a priori* e probabilità *a posteriori* (Bayes 1763). Il risultato, che nella formulazione moderna stabilisce una relazione tra probabilità *a priori* e probabilità *a posteriori*, pone le basi per una concezione dell'inferenza come passaggio dall'evidenza osservata alla valutazione probabilistica di ipotesi non direttamente osservabili. Lo stesso Bayes descrive questo processo nei termini dell'atto di *infer*: il soggetto, osservando il numero di eventi favorevoli e sfavorevoli in una serie di estrazioni, è "*obliged to infer*" la proporzione sottostante e a chiedersi quali conclusioni possa "*reasonably make*"¹⁰. L'atto di inferire assume così la forma di un processo di aggiornamento delle credenze, in cui la probabilità inizialmente attribuita a un'ipotesi viene modificata sulla base delle evidenze empiriche disponibili.

⁹ Cfr. Bernoulli (1713: 70): "*Quod ipsum etiam vel absque calculo utcunque sic inferri potest: A habens omnes nummos praeter unum habet b casus ad vincendum, & B habens omnes nummos praeter unum, habet c casus ad vincendum*" ['La stessa cosa può essere inferita in qualche modo anche senza calcolo: A, possedendo tutte le monete tranne una, ha b casi favorevoli alla vittoria, e B, possedendo tutte le monete tranne una, ha c casi favorevoli alla vittoria'].

¹⁰ Cfr. Bayes (1763: 411): "Let us then imagine a person present at the drawing of a lottery, who knows nothing of its scheme or of the proportion of Blanks to Prizes in it. Let it further be supposed, that he is obliged to infer this from the number of blanks he hears drawn compared with the number of prizes; and that it is enquired what conclusions in these circumstances he may reasonably make" ['Immaginiamo dunque una persona presente all'estrazione di una lotteria, che non sappia nulla della sua struttura o della proporzione tra biglietti perdenti e premi. Si supponga inoltre che sia obbligata a inferire tale proporzione dal numero di biglietti perdenti che sente estrarre rispetto al numero dei premi, e ci si domandi quali conclusioni possa ragionevolmente trarre in queste circostanze'].

Nel corso del XIX secolo e, soprattutto, nella prima metà del XX secolo, questi sviluppi confluiscono nella progressiva formalizzazione della statistica matematica e dell'inferenza statistica, consolidando l'idea che le conclusioni relative a fenomeni non direttamente osservabili possano essere formalizzate matematicamente a partire dai dati osservati (Fisher, 1922; Neyman & Pearson, 1933).

4 L'inferenza nei modelli computazionali

A partire dalla seconda metà del XX secolo, nella terminologia scientifica e, successivamente, in quella computazionale, *inference* viene impiegato per designare procedure attraverso cui, a partire da dati o informazioni disponibili, si ricavano conclusioni, stime o previsioni relative a ciò che non è direttamente osservato. L'espressione *statistical inference* è attestata già nel titolo del volume curato da Koopmans (1950), *Statistical Inference in Dynamic Economic Models*. Salmon (1957) riprende e sviluppa l'espressione *predictive inference* per indicare il passaggio da un campione osservato a uno non osservato. In ambito computazionale, il termine si afferma successivamente nella terminologia dei sistemi esperti, nei quali l'*inference engine* applica regole a una base di conoscenze per produrre nuove conclusioni (Buchanan & Shortliffe, 1984).

Nei primi “sistemi esperti” degli anni '70 e '80 (Buchanan & Shortliffe, 1984; Jackson, 1986) e nella logica computazionale classica (Kowalski, 1974), l'inferenza mantiene un legame diretto con la tradizione logico-deduttiva. I motori inferenziali applicano infatti regole esplicite a fatti contenuti nella base di conoscenze, producendo nuove conclusioni attraverso sequenze di applicazioni successive. Le regole di produzione, generalmente formulate secondo schemi del tipo “se... allora...”, e le corrispondenti catene inferenziali traducono così in procedure computazionali un principio già proprio della logica classica: la derivazione di una conclusione a partire da premesse date secondo relazioni formalmente definite. L'inferenza computazionale si presenta quindi, in questa prima fase, come una formalizzazione operativa di procedure riconducibili al ragionamento deduttivo.

A partire dalla fine degli anni '80, con l'affermazione dei modelli probabilistici (Pearl, 1988; Jordan, 1999), e in modo ancora più evidente nei primi anni del XXI secolo con lo sviluppo del *machine learning* (Bishop, 2006; Murphy, 2012), il termine “inferenza” inizia a essere impiegato per designare operazioni di natura diversa: In questo contesto, non consiste più nella derivazione logica di conclusioni necessarie, ma nella stima di variabili latenti a partire da dati osservabili, secondo criteri statistici e probabilistici.

Attualmente, nel contesto dei modelli contemporanei di IA, e in particolare gli LLM basati sull'architettura Transformer (Vaswani et al., 2017), il nucleo dell'inferenza risiede nella predizione sequenziale dei token. Ogni token successivo viene stimato sulla base della distribuzione di probabilità condizionata sui token precedenti, e selezionato secondo strategie di decodifica o campionamento. In

questo meccanismo, di tipo autoregressivo – vale a dire un processo in cui ogni elemento della sequenza è generato utilizzando come contesto i soli elementi precedenti già prodotti –, la fase di inferenza si distingue dalla fase di addestramento: durante l’addestramento il modello apprende pattern statistici dai dati, aggiornando i pesi della rete neurale, mentre con l’inferenza in senso proprio applica le conoscenze acquisite a input nuovi, senza modificare i parametri del modello¹¹.

5 L’inferenza in linguistica e nelle scienze cognitive

Nel corso del Novecento, il concetto di inferenza si impone progressivamente anche come nozione per la descrizione dei processi interpretativi che regolano la produzione e la comprensione del linguaggio. Entro questo paradigma emerge la consapevolezza che numerosi processi interpretativi richiedono l’attivazione di conoscenze enciclopediche e di schemi situazionali; in tal senso, l’interpretazione coinvolge processi che vanno oltre il significato frasale (Lyons, 1977), aprendo la strada a una concezione del significato come risultato di un’interazione tra codifica linguistica e inferenza, tra “detto” e “non detto” (Sbisà, 2007).

In pragmatica, l’atto di interpretazione viene descritto come un processo inferenziale orientato alla ricostruzione dell’intenzione comunicativa: comprendere un enunciato significa trarre inferenze appropriate circa ciò che il parlante intende comunicare, assumendo che lo scambio linguistico sia regolato da un “principio di cooperazione” (Grice, 1975). Le “implicature” introdotte da Grice corrispondono a inferenze pragmatiche motivate dal contesto e da aspettative condivise tra i partecipanti allo scambio comunicativo, non sono deducibili dalla sola struttura logico-semantica dell’enunciato. Così in (1), assumendo che il parlante partecipi allo scambio in modo cooperativo, il destinatario inferisce che l’impegno menzionato rende impossibile la presenza alla riunione:

- (1) – Vieni alla riunione domani?
– Domani ho lezione tutto il giorno.

A differenza delle implicature, che emergono dall’interazione tra enunciato e contesto, le presupposizioni attivano un contenuto inferenziale che non è asserito, ma recuperato come “condizione di felicità” (Strawson, 1964, Stalnaker, 2002,

¹¹ Cfr. Bergmann (2026): “In AI inference, the trained model then makes predictions on real-world input data. AI inference works by using what it has ‘learned’ – that is, the model parameter updates that were made in order to improve its performance on the training data – to infer the correct output for the new input data. Unlike in model training, inference entails only a forward pass” [“Nell’inferenza dell’IA, il modello addestrato formula quindi previsioni su dati di input del mondo reale. L’inferenza dell’IA utilizza ciò che il modello ha «appreso» – cioè gli aggiornamenti dei parametri apportati per migliorarne le prestazioni sui dati di addestramento – per inferire l’output corretto per un nuovo input. A differenza dell’addestramento del modello, l’inferenza comporta soltanto un passaggio in avanti”].

Ferrari & Zampese, 2016). Per esempio, l'enunciato (2) attiva la presupposizione "Pietro in precedenza fumava":

- (2) Pietro ha smesso di fumare.

Questa linea viene ulteriormente sviluppata nella "teoria della pertinenza" di Sperber e Wilson (1986), che concepisce la comunicazione come un processo fondamentalmente inferenziale. Il destinatario costruisce e valuta interpretazioni possibili sulla base degli indizi forniti dall'enunciato e dal contesto, orientandosi verso quella che presenta il rapporto più favorevole tra effetti cognitivi e sforzi di elaborazione. L'interpretazione che ne emerge è l'esito di un processo inferenziale guidato dalla "pertinenza", in cui il destinatario seleziona l'ipotesi più adatta a rendere comprensibile il comportamento comunicativo osservato nel contesto. In (3), l'inferenza secondo cui l'esame è presumibilmente andato bene deriva dall'assunzione che lo stato d'animo dell'esaminatore è rilevante per l'esito della prova:

- (3) – Com'è andato l'esame?
– Il professore era di ottimo umore.

Analogamente, in (4) l'inferenza secondo cui l'articolo non è stato completato emerge come l'ipotesi più pertinente, in quanto spiega la risposta del parlante con il minimo sforzo cognitivo:

- (4) – Hai finito l'articolo?
– È già mezzanotte.

Questa prospettiva è stata ulteriormente consolidata dalla linguistica del testo, che ha messo in evidenza come l'attività inferenziale costituisca una condizione necessaria per il mantenimento della coerenza testuale. In (5) la connessione tra Matteo e Paolo non è stabilita da meccanismi linguistici di coreferenza, ma passa attraverso un'operazione inferenziale attivata dalla lettura del primo enunciato (Ferrari & Zampese 2016):

- (5) Michela ha due figli molto diversi. Matteo è molto calmo e riflessivo; Paolo invece è molto vivace, non si riesce a tenerlo fermo un attimo.

Un analogo ruolo dell'inferenza è messo in evidenza dai modelli cognitivi di comprensione, come in Kintsch (1988; 1998). All'interno di questo modello, la distinzione tra "inferenze locali" e "inferenze globali" consente di rendere conto del tipo di elaborazione attivato nel processo di costruzione del significato. In (6), l'ipotesi secondo cui stava piovendo costituisce un'inferenza locale di collegamento, necessaria per stabilire la coerenza tra i due enunciati e per costruire una rappresentazione situazionale plausibile della sequenza narrativa:

- (6) Giulia uscì di casa senza ombrello. Pochi minuti dopo era completamente bagnata.

In (7), l'inferenza secondo cui il personaggio agisce per vendetta o per interesse personale non è richiesta per collegare singoli enunciati, ma contribuisce alla costruzione del modello situazionale: si tratta pertanto di un'inferenza globale, che opera a un livello di integrazione semantica superiore.

- (7) Nel corso del racconto, il protagonista agisce ripetutamente contro lo stesso personaggio, anche a costo di gravi conseguenze personali.

Un'ulteriore distinzione riguarda le “inferenze di connessione” e le “inferenze di integrazione” (Cignetti et al. 2022), che mostrano come l'attività inferenziale sia al tempo stesso vincolata dalla struttura testuale e aperta al contributo delle conoscenze enciclopediche. Le inferenze di connessione consentono di istituire nessi tra diverse parti del testo, contribuendo alla costruzione della coerenza testuale. Nell'esempio (8), il lettore inferisce un rapporto causale tra il ritardo del treno e l'arrivo tardivo di Marta, benché tale rapporto non sia esplicitamente formulato:

- (8) Il treno era in ritardo. Gianni arrivò tardi alla riunione.

Le inferenze di integrazione consentono invece di arricchire la rappresentazione del contenuto testuale recuperando informazioni non esplicitate attraverso il ricorso alle conoscenze enciclopediche e agli schemi cognitivi del lettore. In (9), il lettore può inferire che Luigi avesse bisogno di alleviare un dolore, benché questa informazione non sia direttamente formulata, sulla base delle conoscenze relative alla funzione degli analgesici e alle situazioni in cui vengono utilizzati:

- (9) Luigi entrò in una farmacia e chiese un analgesico.

In tutti i casi osservati, il termine “inferenza” è impiegato per designare operazioni cognitive accomunate dalla funzione di colmare lo scarto tra ciò che è linguisticamente espresso e ciò che viene effettivamente compreso.

6 Gli schemi di Winograd

Un'esperienza significativa in cui emerge la differenza tra inferenza umana e comportamento dei modelli linguistici riguarda i cosiddetti “schemi di Winograd” (Levesque et al., 2012). Si tratta di esempi ideati nel quadro di una revisione critica del test di Turing, con l'obiettivo di costruire un banco di prova per valutare la capacità dei sistemi artificiali di effettuare inferenze non riducibili a semplici associazioni superficiali o a strategie di riconoscimento di pattern. Due esempi sono i seguenti:

- (10) I giudici negarono il permesso ai manifestanti perché temevano la violenza.
- (11) I giudici negarono il permesso ai manifestanti perché sostenevano la violenza.

In (10) l'interpretazione più naturale è che a temere la violenza siano i giudici mentre in (11) il verbo si riferisce ai manifestanti: la risoluzione dell'ambiguità richiede un'inferenza basata su conoscenze circa il comportamento tipico dei diversi attori coinvolti nella situazione.

Nella loro formulazione originaria, gli schemi di Winograd rappresentavano un banco di prova per distinguere tra sistemi capaci di ragionamento semantico e sistemi che si limitano a sfruttare regolarità statistiche nei dati linguistici. L'idea di fondo era che, in assenza di segnali lessicali o sintattici decisivi, la risoluzione dell'ambiguità richiedesse il ricorso a conoscenze sul mondo e la loro applicazione flessibile. Con l'avvento dei modelli linguistici di grandi dimensioni, tuttavia, questi esempi si sono progressivamente rivelati meno efficaci. I modelli linguistici contemporanei, infatti, possono individuare la risposta corretta anche senza attivare un processo inferenziale esplicito analogo a quello umano, grazie alla disponibilità di ampi corpora di addestramento, nei quali compaiono anche varianti di scenari analoghi a quelli degli schemi di Winograd (Sakaguchi et al., 2020). La possibilità di ottenere prestazioni linguistiche elevate a partire dalla sola elaborazione della forma linguistica non costituisce, tuttavia, di per sé una prova di comprensione del significato (Bender & Koller, 2020).

In questo contesto, è stata avanzata l'ipotesi che le prestazioni elevate osservate in tali compiti possano dipendere, almeno in parte, da fenomeni di "*verbatim memorization*", ossia dalla memorizzazione – più o meno diretta – di sequenze linguistiche presenti nei dati di addestramento. In altri termini, il modello potrebbe non risolvere l'ambiguità attraverso un processo inferenziale in senso proprio, ma riconoscere configurazioni testuali già incontrate, o sufficientemente simili a esempi precedenti, e sulla base di tali corrispondenze selezionare la continuazione più probabile (Huang et al., 2024).

7 “Inferenza” come termine-ombrello

Nei paragrafi precedenti si è mostrato come le inferenze prodotte dai lettori umani e quelle generate dai modelli linguistici rinviino a processi diversi per natura e funzionamento. Questa divergenza invita a una riflessione più rigorosa sull'uso del termine *inferenza* nelle diverse tradizioni disciplinari. Uno degli aspetti più rilevanti – e al contempo più problematici – che emerge dall'analisi è infatti il suo impiego come “termine-ombrello”, che ricorre in ambiti anche molto distanti (dalla logica alla statistica, dalla linguistica alle scienze cognitive, fino all'intelligenza artificiale) per designare operazioni che, pur condividendo alcune affinità funzionali, differiscono sensibilmente per natura, metodologia e grado di formalizzazione.

Dal punto di vista della realizzazione linguistica, le strategie principali possono essere ricondotte alle seguenti tre tipologie:

- a) l'uso di espressioni polirematiche del tipo “inferenza” + aggettivo;
- b) l'impiego di termini che presuppongono o implicano un processo inferenziale senza nominarlo esplicitamente;
- c) l'uso del sostantivo “inferenza” privo di ulteriori specificazioni.

Al primo gruppo appartengono espressioni come “inferenza statistica”, “inferenza necessaria”, “inferenza locale”, “inferenza globale”, “inferenza di connessione” e simili. Questa strategia presenta il vantaggio di delimitare con maggiore precisione il tipo di processo in gioco, offrendo al contempo etichette riconoscibili e flessibili, adattabili a contesti diversi. Ne risulta una riduzione dell'ambiguità concettuale e una più netta distinzione tra inferenze deduttive, probabilistiche o computazionali. La moltiplicazione delle specificazioni può tuttavia rendere meno riconoscibili i rapporti tra categorie appartenenti a tradizioni disciplinari diverse, soprattutto quando le etichette non sono accompagnate da criteri espliciti che ne definiscano i confini e i rapporti reciproci.

Il secondo gruppo comprende le denominazioni che mettono in rilievo singole manifestazioni entro uno stesso processo, vale a dire il passaggio – non direttamente osservabile – tra premesse e conclusioni, tra dati e ipotesi, oppure tra enunciati e contenuti impliciti. Così, nella tradizione logica classica, l'inferenza necessaria coincide con la “deduzione”, mentre in pragmatica si intreccia con nozioni specialistiche quali “implicatura” e “presupposizione”. Anche questa strategia consente una maggiore precisione rispetto all'uso generico del termine; tuttavia, la sua interpretazione presuppone competenze disciplinari specifiche, rendendone meno immediata la fruizione in contesti interdisciplinari.

Il terzo gruppo, infine, è costituito dall'uso di “inferenza” senza ulteriori qualificazioni. Tale scelta privilegia la semplicità e l'immediatezza, delegando al contesto il compito di disambiguare la natura del processo considerato; ma proprio per questo espone al rischio di appiattare differenze rilevanti tra processi cognitivi, statistici e computazionali, ricondotti così a un'unica etichetta solo apparentemente unificante.

Questa terza modalità è oggi prevalente nel lessico dell'intelligenza artificiale. Ciò rende particolarmente urgente una riflessione terminologica più attenta: l'uso non specificato di “inferenza” può infatti suggerire, in modo fuorviante, una continuità tra inferenza naturale e inferenza computazionale, nonostante le profonde differenze nei meccanismi sottostanti, nei livelli di formalizzazione e nei fini operativi. Sarebbe pertanto opportuno un impiego più consapevole e contestualizzato del termine, capace di preservarne la funzionalità descrittiva in ambito computazionale senza attenuare le distinzioni – teoricamente ed empiricamente fondate – tra processi umani e procedure algoritmiche.

Attributions

Luca Cignetti: Conceptualization, methodology, formal analysis, writing.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this contribution.

Acknowledgments

I acknowledge the use of OpenAI's GPT-4-Turbo in assisting with the editing and refinement of this paper, enhancing both its clarity and presentation in English. However, I ensured that all contributions adhered strictly to the standards and ethical guidelines of academic writing.

Bibliografia

- Abelardo, Pietro. 1970 [1117-1121]. *Dialectica*. A c. di L. M. de Rijk. Assen: Van Gorcum.
- Algarotti, Francesco. 1823. *Opere scelte*. Milano: Società Tipografica de' Classici Italiani.
- Bayes, Thomas. 1763. *An Essay towards Solving a Problem in the Doctrine of Chances*. Philosophical Transactions of the Royal Society of London, vol. LIII, 370-418.
- de Beaugrande, Robert-Alain, & Dressler, Wolfgang U. 1984. *Introduzione alla linguistica testuale*. Bologna: Il Mulino.
- Bender, Emily M. & Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185-5198. <https://doi.org/10.18653/v1/2020.acl-main.463>
- Bergmann, Daniel. 2026. What is AI Inference? IBM. <https://www.ibm.com/think/topics/ai-inference> (ultima consultazione: 20 settembre 2026).
- Bernoulli, Jakob. 1713. *Ars Conjectandi*. Basel: Thurneysen.
- Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. New York: Springer.
- Boethius. 2000 [523-524]. *De consolatione philosophiae*. Ed. Claudio Moreschini. München/Leipzig: K. G. Saur.
- Buchanan, Bruce G. & Shortliffe, Edward H. 1984. *Rule Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Reading, MA: Addison Wesley.

- Capella, Martianus. 1836 [410-429]. *De Nuptiis Philologiae et Mercurii et de septem artibus liberalibus Libri novem*. A c. di Ulricus Fridericus Kopp. Francofurti ad Moenum: F. Varrentrapp.
- Cicerone, Marco Tullio. 2015 [85 a.C.]. *De Inventione*. Torino: UTET.
- Cignetti, Luca. 2021. *Leggere e comprendere a scuola*. In L. Cignetti & E.D. Manetti (a c. di), *Forme e ritmi della lettura nel Cantone Ticino*. Bellinzona: DECS, 16-25.
- Cignetti, Luca, Silvia Demartini, Simone Fornara & Matteo Viale. 2022. *Didattica dell'italiano come lingua prima*. Bologna: Il Mulino.
- Desrosières, Alain. 1998. *The Politics of Large Numbers: A History of Statistical Reasoning*. Cambridge, MA: Harvard University Press.
- Ferrari, Angela & Zampese, Luciano. 2016. *Grammatica: parole, frasi, testi dell'italiano*. Roma: Carocci.
- Fisher, Ronald Aylmer. 1922. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222. 309-368. <https://doi.org/10.1098/rsta.1922.0009>
- Galilei, Galileo. 1843-1853 [1610-1632]. *Opere astronomiche*. A c. di E. Albèri. Firenze: Tipografia. 5 voll.
- Gigerenzer, Gerd, Zeno Swijtink, Terrence Porter, Lorraine Daston, John Beatty & Lothar Krüger. 1989. *The Empire of Chance: How Probability Changed Science and Everyday Life*. Cambridge: Cambridge University Press.
- Gioberti, Vincenzo. 1846. *Prolegomeni del primato morale e civile degli Italiani*. Lugano: Tipografia della Svizzera italiana.
- Grice, H. Paul. 1975. Logic and conversation. In Peter Cole & Jerry L. Morgan (a c. di), *Syntax and Semantics*, vol. 3, 41-58. New York: Academic Press.
- Hacking, Ian. 1975. *The Emergence of Probability: A Philosophical Study of Early Ideas about Probability, Induction and Statistical Inference*. Cambridge: Cambridge University Press.
- Hald, Anders. 2003. *A History of Probability and Statistics and Their Applications before 1750*. New York: Wiley. <https://doi.org/10.1002/0471725161>
- Huang, Jing, Diyi Yang & Christopher Potts. 2024. Demystifying Verbatim Memorization in Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*. Miami, FL: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.emnlp-main.598>
- Hooker, Richard. 1593. *Of the Lawes of Ecclesiasticall Politie*. London: R. Bishop.
- Jackson, Peter. 1986. *Introduction to Expert Systems*. Reading, MA: Addison Wesley.
- Jordan, Michael I. 1999. *Learning in Graphical Models*. Cambridge, MA: MIT Press.

- Kintsch, Walter. 1988. The role of knowledge in discourse comprehension: A construction integration model. *Psychological Review* 95(2), 163-182. <https://doi.org/10.1037/0033-295x.95.2.163>
- Kintsch, Walter. 1998. *Comprehension: A Paradigm for Cognition*. Cambridge: Cambridge University Press.
- Koopmans, Tjalling Charles (a c. di). 1950. *Statistical Inference in Dynamic Economic Models*. New York: John Wiley & Sons.
- Kowalski, Robert. 1974. Predicate logic as a programming language. *Proceedings of IFIP Congress 74*, 569-574.
- Lapo da Castiglionchio. XIV sec. *Epistola o sia ragionamento*. A c. di L. Mehus. Bologna: Corciolani e Colli, 1749.
- Levesque, Hector, Ernest Davis & Leora Morgenstern. 2012. *The Winograd Schema Challenge*. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning (KR 2012)*, 552-561. Toulouse: AAAI Press.
- Lyons, John. 1977. *Semantics*. 2 voll. Cambridge: Cambridge University Press.
- Machiavelli, Niccolò. 1875-1877 [1498-1512]. *Le Legazioni e Commissarie di Niccolò Machiavelli, riscontrate sugli originali ed accresciute di nuovi documenti*. A c. di Luigi Passerini & Gaetano Milanese. Firenze-Roma: Tipogr. Cenniniana, 1875-1877.
- Murphy, Kevin P. 2012. *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press.
- Neyman, Jerzy & Egon S. Pearson. 1933. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231, 289-337. <https://doi.org/10.1098/rsta.1933.0009>
- OED = Oxford English Dictionary. 2026. "Inference" Oxford University Press. https://www.oed.com/dictionary/inference_n (ultima consultazione: 20 settembre 2026).
- Pearl, Judea. 1988. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Morgan Kaufmann. <https://doi.org/10.1016/C2009-0-27609-4>
- Rey, Alain (dir.). 1992. *Dictionnaire historique de la langue française*. Paris: Le Robert.
- Sakaguchi, Keisuke, Ronan Le Bras, Chandra Bhagavatula & Yejin Choi. 2020. *WinoGrande: An adversarial Winograd Schema Challenge at scale*. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(5), 8732-8740. <https://doi.org/10.1609/aaai.v34i05.6399>
- Salmon, Wesley C. 1957. The Predictive Inference. *Philosophy of Science* 24 (2): 180-190. <https://doi.org/10.1086/287532>

- Sbisà, Marina. 2007. *Detto non detto. Le forme della comunicazione implicita*. Roma-Bari: Laterza.
- Sperber, Dan & Wilson, Deirdre. 1986. *Relevance: Communication and Cognition*. Oxford: Blackwell.
- Stalnaker, Robert. 2002. *Common ground*. In *Linguistics and Philosophy* 25(5-6): 701–721. <https://doi.org/10.1023/A:1020867916902>
- Strawson, Peter F..1964. *Identifying Reference and Truth-Values*. In *Theoria* 302: 96-118. Rep. in Idem, *Logico-Linguistic Papers*, London: Methuen, 1971, 75-95. <https://doi.org/10.1111/j.1755-2567.1964.tb00404.x>
- TLFi = *Trésor de la langue française informatisé*. ATILF-CNRS / Université de Lorraine. <https://www.cnrtl.fr/> (ultima consultazione: 20 settembre 2026).
- Unione Europea. 2024. Regolamento (UE) 2024/1689 del Parlamento Europeo e del Consiglio del 13 giugno 2024 che stabilisce regole armonizzate sull'intelligenza artificiale e modifica i regolamenti (CE) e (UE) e le direttive correlate. Gazzetta Ufficiale dell'Unione Europea L 1689, 12.7.2024. Bruxelles: Unione Europea.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser & Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30. 5998-6008. <https://doi.org/10.48550/arXiv.1706.03762>

About the journal

AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses is a Diamond Open Access journal supported by PUBLIA – SLUB Open Publishing.



DOI <https://doi.org/10.62408/ai-ling>

ISSN 2943-0070