Dall'output algoritmico al giudizio valutativo: governare l'evidenza nell'era dell'Intelligenza Artificiale

Francesco Mazzeo Rinaldi*

This article critically examines the role of Artificial Intelligence (AI) in policy and program evaluation, focusing on the processes through which algorithmic outputs become evaluative evidence. Drawing on a critical narrative review of international literature published between 2016 and 2025, the paper identifies three main ways AI is conceptualized within evaluation: as an extension of analytical capacity, as an interpretive and ethical risk, and as an object of evaluation and governance. The analysis shows that these perspectives often leave implicit the inferential chain linking outputs, interpretation, and evaluative judgment. To address this gap, the article proposes an analytical framework structured around five stages - algorithmic output, technical validation, theory-informed interpretation, triangulation, and evaluative judgment - aimed at making visible potential forms of "epistemic compression" in AI-mediated evaluation processes.

Keywords: artificial Intelligence; evaluative evidence; algorithmic governance; critical narrative review.

Introduzione

Negli ultimi anni l'Intelligenza Artificiale (IA) ha acquisito una crescente centralità nel dibattito sulle scienze sociali applicate e, più in generale, sulle pratiche fondate sull'uso dell'evidenza (Athey & Berg, 2023; Sharma, 2025). Nel campo della valutazione di politiche e programmi, questa attenzione si è inizialmente concentrata sulle opportunità offerte dalla disponibilità di grandi volumi di dati e da strumenti computazionali sempre più sofisticati:

* Francesco MAZZEO RINALDI, Università di Catania (fmazzeo@unict.it)
Invio Proposta: 20-05-2026. Accettazione: 15-07-2026.

capacità di analisi su larga scala, automazione di compiti routinari, possibilità di osservare fenomeni sociali attraverso nuove fonti informative (Petersson & Breul, 2017). I Big Data e, successivamente, l'IA venivano interpretati come risorse potenzialmente capaci di rafforzare processi valutativi evidence-based e practice-based, a condizione di saper trasformare i dati grezzi in informazioni rilevanti per le diverse fasi del ciclo delle politiche (Mazzeo Rinaldi & Occhipinti, 2023). Tuttavia, rimaneva sullo sfondo una questione cruciale: quali sono i passaggi attraverso cui l'informazione prodotta mediante strumenti computazionali diventa evidenza valutativa? Il rapido sviluppo dell'IA, e in particolare dei modelli di apprendimento automatico e dei sistemi generativi, ha reso questa domanda non più eludibile. Oggi, una quota crescente di analisi, classificazioni, predizioni e persino testi interpretativi può essere prodotta da sistemi algoritmici con un livello di autonomia e complessità senza precedenti. In questo scenario, il problema centrale non riguarda tanto se utilizzare l'IA nella valutazione, quanto come governare il processo che conduce dagli output algoritmici al giudizio valutativo.

La letteratura internazionale più recente riflette chiaramente questa tensione (Nielsen et al, 2025; Leeuw, & Bamberger, 2025). Accanto a contributi che enfatizzano le potenzialità dell'IA in termini di efficienza, scala e tempestività (Shapiro & Lam, 2025), si moltiplicano le riflessioni critiche sui rischi epistemici ed etici associati all'uso di sistemi opachi, potenzialmente distorti e difficilmente auditabili (Azzam, 2023; Greenstein & Cho, 2025). Recentemente, anche la rivista Evaluation ha dedicato un numero speciale alle sfide della digitalizzazione nella pratica valutativa, affrontano questioni metodologiche, di ownership ed etiche nell'uso dell'IA e Big Data (Potluka *et al.*, 2025). Ciò che appare meno sviluppato è un quadro concettuale integrato che chiarisca il ruolo specifico della valutazione in questo contesto: non come semplice utilizzatrice di nuovi strumenti, ma come campo di sapere responsabile della trasformazione dell'informazione in evidenza e dell'evidenza in giudizio.

È proprio su questo passaggio che si colloca il presente contributo. L'obiettivo non è valutare l'efficacia dell'IA né proporre linee guida tecniche per il suo impiego, ma interrogare il modo in cui gli output algoritmici vengono trasformati in evidenza valutativa. La tesi di fondo è che la sfida posta dall'IA non sia primariamente tecnica, bensì epistemica: riguarda il governo dei processi di inferenza, interpretazione e giudizio che definiscono la pratica valutativa. Attraverso una sintesi critica della letteratura valutativa recente, l'articolo mostra che le diverse modalità di concettualizzare l'IA condividono un punto cieco: i passaggi attraverso cui l'output computazionale diventa evidenza restano in larga misura impliciti. Per affrontare questo nodo,

viene proposto un framework analitico che distingue cinque stadi – dall'output algoritmico al giudizio valutativo – rendendo visibili le tensioni che attraversano ciascun passaggio.

Il contributo dell'articolo non consiste nell'offrire una nuova "buona pratica" per l'uso dell'IA, ma nel fornire una lente interpretativa per analizzare come, dove e a quali condizioni la catena inferenziale viene compressa, delegata o opacizzata. In un contesto in cui la produzione dell'evidenza è sempre più mediata da sistemi algoritmici, il futuro della valutazione dipende meno dalla capacità di adottare nuovi strumenti che dalla capacità di preservare e governare i processi che fondano la legittimità del giudizio.

1. Metodo: una *critical narrative review*

L'articolo adotta una critical narrative review della letteratura su IA e valutazione (2016–2025). Questo approccio non mira a una ricognizione esaustiva né a una sintesi quantitativa degli effetti, ma alla chiarificazione concettuale e alla problematizzazione delle assunzioni epistemiche che attraversano un campo emergente. Nel dibattito metodologico sulle revisioni della letteratura, le critical narrative reviews sono state distinte dalle review sistematiche e meta-analitiche per la loro funzione interpretativa e teorica. Grant e Booth (2009), nella loro tipologia delle revisioni, collocano le narrative e critical reviews tra gli approcci maggiormente orientati alla costruzione di senso (sense-making), alla messa in discussione delle assunzioni dominanti e allo sviluppo concettuale, piuttosto che alla sola aggregazione delle evidenze empiriche. In modo analogo, Saunders e Rojon (2011) sottolineano come una critical literature review non si limiti a descrivere lo stato dell'arte, ma si caratterizzi per la valutazione argomentata delle posizioni esistenti, l'identificazione di lacune concettuali e la riformulazione dei problemi di ricerca. In questo senso, l'elemento "critico" non riguarda la quantità delle fonti analizzate, bensì la profondità dell'analisi e la capacità di esplicitare le assunzioni epistemologiche che attraversano un determinato campo di studi.

Questa impostazione risulta particolarmente appropriata nel caso dell'IA applicata alla valutazione, un ambito in cui la letteratura è ancora fortemente eterogenea, interdisciplinare e caratterizzata da rapide evoluzioni tecnologiche. In tali contesti, come osservano Greenhalgh e colleghi (2018), l'idea di una gerarchia rigida che pone le review sistematiche al vertice e le narrative review a un livello inferiore risulta fuorviante. Gli autori argomentano che la scelta del tipo di review dovrebbe dipendere dalla natura della domanda di ricerca e dagli obiettivi conoscitivi perseguiti, riconoscendo il valore delle

narrative e critical reviews quando si tratta di concettualizzare campi emergenti, chiarire concetti ambigui o problematizzare quadri teorici consolidati. La rapida evoluzione delle applicazioni di IA in ambito valutativo e la natura ancora sperimentale di molte pratiche rendono inoltre necessario considerare le evidenze disponibili come provvisorie e contestuali, rafforzando l'opportunità di un approccio critico e interpretativo alla sintesi della letteratura (Greenhalgh *et al.*, 2018). Coerentemente con questa impostazione, la sintesi proposta in questo articolo non intende offrire una mappatura esaustiva di tutte le applicazioni dell'IA nella valutazione, né una valutazione comparativa delle loro prestazioni. L'obiettivo è piuttosto quello di:

- identificare i principali modi in cui l'IA viene concettualizzata nella letteratura valutativa;
- mettere in luce convergenze, tensioni e punti ciechi ricorrenti;
- interrogare criticamente il rapporto tra output algoritmici, produzione dell'evidenza e giudizio valutativo.

La selezione dei contributi analizzati è pertanto di tipo intenzionale (purposive) e si concentra su lavori pubblicati prevalentemente tra il 2016 e il 2025 che hanno svolto un ruolo significativo nello strutturare il dibattito internazionale. La strategia di selezione dettagliata, articolata in tre fasi, è descritta nella Sezione 1.1. Il corpus selezionato è stato analizzato attraverso alcune lenti critiche ricorrenti, coerenti con la tradizione metodologica della valutazione:

- lo statuto epistemico degli output algoritmici;
- le implicazioni per validità, affidabilità e trasparenza dell'evidenza prodotta;
- le conseguenze etiche e distributive dell'uso dell'IA;
- le ricadute sulla governance della valutazione e sull'identità professionale dei valutatori.

Su queste basi, la sintesi critica si concentra sull'analisi dei principali modi in cui l'IA viene concettualizzata nella letteratura valutativa. La sezione che segue esplicita la strategia di selezione delle fonti e il corpus analizzato.

1.1 Strategia di selezione e corpus analizzato

La selezione dei contributi è avvenuta attraverso una strategia di campionamento intenzionale progressivo (Paterson *et al.*, 2001), articolata in tre fasi:

- Identificazione di anchor texts. Sono stati assunti come testi di riferimento tre lavori che hanno strutturato il dibattito internazionale post-

ChatGPT: (a) lo special issue di New Directions for Evaluation curato da Mason e Montrosse-Moorhead (2023), comprendente 10 saggi; (b) il volume Routledge curato da Nielsen, Mazzeo Rinaldi e Petersson (2025), composto da 14 capitoli; (c) e lo special issue della rivista Evaluation curato da Potluka, Harten, Kocks e Dvorak che raccoglie 8 contributi. I tre volumi sono stati analizzati integralmente e utilizzati come mappa iniziale del campo.

- Espansione attraverso reti citazionali (Google Scholar, Web of Science). A partire dagli anchor texts, sono stati individuati contributi ricorrentemente citati tramite analisi delle reti di citazione (Google Scholar, Web of Science), limitando la ricerca al periodo 2016–2025 e a riviste internazionali di valutazione.
- Selezione finale attraverso criteri di rilevanza epistemica. I testi sono stati selezionati in base a tre criteri: (a) rilevanza per il rapporto tra IA ed evidenza valutativa; (b) diversità di prospettive (strumentali, critiche, orientate alla governance); (c) originalità concettuale (framework, riflessioni teoriche), escludendo contributi orientati a meri casi applicativi.

Il corpus finale comprende diciassette contributi principali (Tabella 1), analizzati attraverso un processo iterativo in quattro stadi: (a) codifica aperta; (b) aggregazione in cluster tematici; (c) analisi comparativa per identificare convergenze, divergenze e lacune; (d) sintesi concettuale che ha generato il framework dei cinque stadi proposto nel paragrafo 4.

1.2. Limitazioni

Tre limitazioni meritano di essere esplicitate. In primo luogo, il focus sulla letteratura in lingua inglese introduce un bias linguistico e geografico che tende a privilegiare le prospettive provenienti dal Global North, con il rischio di sotto rappresentare approcci e contributi elaborati in altri contesti. In secondo luogo, il focus sui tre *anchor texts* comporta un bias legato alla rete accademica di riferimento: questa impostazione favorisce una specifica comunità epistemica, prevalentemente AEA-centrica, e può escludere voci critiche alternative che circolano in altri ambiti disciplinari. Infine, va considerato il posizionamento dell'autore, che ha contribuito direttamente al dibattito analizzato. Ciò comporta il rischio di una selezione non neutrale delle fonti. Questo coinvolgimento viene tuttavia reso esplicito e assunto come parte integrante del metodo: se da un lato consente una comprensione profonda delle dinamiche interne al

campo, dall'altro richiede una costante vigilanza critica per evitare possibili bias confermativi. Queste limitazioni non invalidano l'approccio ma ne definiscono il perimetro di validità. Una critical narrative review non ambisce a rappresentatività statistica ma a profondità analitica e sviluppo concettuale (Greenhalgh *et al.*, 2018).

Tab. 1 - Corpus dei principali contributi analizzati

ANCHOR TEXTS

Autori	**Anno**	**Focus tematico**	**Concettualizzazione IA**
Mason & Montrosse-Moorhead (eds.)	2023	Special Issue *NDE*, Vol. 177-178: Evaluation and AI	Anchor text
Nielsen, Mazzeo Rinaldi & Peterson (eds.)	2025	Volume Routledge: IA and Evaluation	Anchor text
Potluka, Harten, Kocks & Dvorak (eds.)	2025	Special Issue *Evaluation*: Digitalization in evaluations andevaluations of digitalization	Anchor text

PRECURSORI

Autori	**Anno**	**Focus tematico**	**Concettualizzazione IA**
Bamberger	2016	Big Data e M&E	Precursore
Petersson & Breul	2017	Big Data e valutazione	Precursore
Mazzeo Rinaldi	2018	Big Data e valutazione (Italia)	Precursore
Picciotto	2020	Big Data: sfide per la valutazione	Precursore
York & Bamberger	2020	Big Data e valutazione	Precursore

CONTRIBUTI VARI (Special Issues/Routledge/...)

Autori	**Anno**	**Focus tematico**	**Concettualizzazione IA**
Nielsen	2023	Disruption valutativa – S.I.	Framework
Montrosse-Moorhead	2023	Criteri valutativi per IA – S.I.	Framework
Azzam	2023	Validità – S.I.	Rischio epistemico
York & Bamberger	2025	Applications of Big Data (Routledge)	Estensione strumentale
Greenstein & Cho	2025	Etica ed equità (Routledge)	Rischio, epistemico
Mazzeo Rinaldi & Nielsen	2025	Identità professionale (Routledge)	Trasformativo
Nielsen	2025	Industria valutativa e tecnologie emergenti (Routledge)	Governance + Estensione
Mazzeo Rinaldi & Occhipinti	2023	Big Data, IA e valutazione in Italia	Estensione strumentale + Rischio
Rompczyk	2025	Standard professionali	Rischio e opportunità
Shapiro & Lam	2025	Applicazioni pratiche	Estensione strumentale
Pudney *et al.*	2025	Valutazione sistemi IA	IA come oggetto
Wencker *et al.*	2025	Text as data e NLP	Estensione strumentale

Nota: I contributi di Bamberger (2016), Petersson & Breul (2017), Mazzeo Rinaldi (2018), Picciotto (2020) e York & Bamberger (2020) sono classificati come 'precursori' in quanto precedenti al rilascio di ChatGPT e rilevanti per contestualizzare il dibattito su Big Data, IA e valutazione.

2. Modi di concettualizzare l'IA nella letteratura valutativa

La letteratura recente sull'IA applicata alla valutazione non costituisce un campo unitario, ma si articola attorno a modalità ricorrenti di concettualizzazione che riflettono assunzioni differenti sul rapporto tra analisi, evidenza e giudizio. Tali modalità non sono semplici "posizioni" teoriche, bensì configurazioni epistemiche che orientano il modo in cui l'IA viene integrata nei processi valutativi e, soprattutto, il modo in cui gli output algoritmici vengono trattati come evidenza.

La sintesi critica del corpus consente di individuare tre configurazioni prevalenti: (a) l'IA come estensione strumentale; (b) l'IA come rischio epistemico ed etico; (c) l'IA come oggetto di valutazione e governance. Queste prospettive non sono mutuamente esclusive: spesso coesistono nello stesso contributo. Tuttavia, ciascuna tende a enfatizzare specifici stadi del processo che conduce dall'output al giudizio, lasciandone altri in ombra.

2.1 L'IA come estensione strumentale della capacità analitica

Un primo insieme di contributi interpreta l'IA prevalentemente come una estensione strumentale della capacità analitica della valutazione. In questa prospettiva, l'attenzione si concentra sulle opportunità offerte da machine learning, natural language processing e modelli generativi per trattare grandi volumi di dati, analizzare testi non strutturati, automatizzare compiti routinari e produrre sintesi rapide. Questa linea affonda le proprie radici nei lavori sui Big Data che precedono l'attuale ondata di IA generativa. Bamberger (2016) e Petersson e Breul (2017) già sottolineavano come la disponibilità di fonti digitali massive potesse ampliare scala, granularità e tempestività dei processi di monitoraggio e valutazione. York e Bamberger (2020) hanno sistematizzato tale prospettiva mostrando come analytics e tecnologie digitali possano rafforzare la misurazione dei risultati in contesti complessi. Nel contesto italiano, Mazzeo Rinaldi (2018) e Mazzeo Rinaldi e Occhipinti (2023) hanno evidenziato il ritardo della valutazione nell'ingaggiare questo dibattito, ma anche il potenziale trasformativo di tali strumenti.

Nel corpus più recente, questa configurazione assume una forma esplicitamente orientata all'IA. Wencker *et al.*, (2025) mostrano come NLP e large language models consentano di trattare corpus testuali di dimensioni prima inaccessibili, generando nuove forme di "text as data" per l'analisi valutativa. Shapiro e Lam (2025) illustrano applicazioni pratiche dell'IA in diverse fasi del ciclo valutativo, dalla raccolta dei dati alla sintesi dei risultati. Nielsen

(2023) interpreta le tecnologie emergenti come fattori di disruption dell'industria valutativa, capaci di automatizzare una quota crescente di compiti analitici e redazionali. In questi contributi, l'IA è prevalentemente concettualizzata come infrastruttura abilitante: uno strumento che estende ciò che il valutatore può fare in termini di velocità, scala e copertura empirica. Il lessico ricorrente è quello dell'efficienza, dell'ampliamento del raggio osservativo e della riduzione dei costi cognitivi e operativi.

Dal punto di vista analitico, questa configurazione tende a concentrare l'attenzione sugli stadi 1 e 2 del framework: produzione dell'output e sua validazione tecnica. Gli autori discutono in modo approfondito le potenzialità analitiche degli strumenti e, in misura crescente, i problemi di accuratezza e bias. Rimangono invece in larga parte impliciti i passaggi successivi – interpretazione theory-informed, triangolazione, giudizio.

In York e Bamberger (2020, 2025), ad esempio, l'IA e gli analytics sono presentati come mezzi per rafforzare la capacità della valutazione di osservare fenomeni complessi e dinamici. Tuttavia, il modo in cui tali output vengono trasformati in evidenza valutativa è dato per scontato: l'enfasi è sulla produzione di informazione più ricca, non sul governo del passaggio dall'informazione al giudizio. Analogamente, in Wencker *et al.* (2025) l'attenzione è rivolta alle tecniche di estrazione di pattern semantici, mentre resta sullo sfondo il problema di come tali pattern debbano essere interpretati alla luce di teorie di programma o criteri valutativi.

Nell'articolo introduttivo dello special issue di Evaluation dedicato alla digitalizzazione delle valutazioni Potluka *et al.* (2025), identificano tra le sfide principali la necessità di sviluppare conoscenze metodologiche adeguate per evitare approcci "black box". Questa preoccupazione si colloca proprio nello spazio coperto dal filone strumentale: l'enfasi sulla potenza analitica richiede una corrispondente attenzione ai processi che trasformano tale potenza in comprensione valutativa. Tuttavia, come negli altri contributi

di questo filone, i passaggi attraverso cui tale trasformazione avviene restano in larga parte da esplicitare.

In questa prospettiva, l'evidenza tende implicitamente a coincidere con l'informazione prodotta dall'algoritmo. La sofisticazione tecnica diventa una garanzia tacita di qualità dell'evidenza. Il rischio principale non è l'errore puntuale, ma la naturalizzazione dell'output come "fatto". Il passaggio inferenziale che trasforma un pattern computazionale in un'affermazione valutativa rimane opaco.

Questa configurazione illumina con forza il potenziale dell'IA per ampliare le capacità analitiche della valutazione, ma oscura il lavoro interpretativo che rende tali capacità rilevanti per il giudizio. In termini di framework, essa

rafforza i primi due stadi e lascia in ombra i successivi. È in questa compressione che si annida il punto cieco comune alla letteratura: l'assunzione che una migliore analisi produca automaticamente una migliore evidenza.

2.2 L'IA come rischio epistemico ed etico

Un secondo filone della letteratura si sviluppa in chiave critica e interpreta l'IA principalmente come una fonte di rischio per la produzione dell'evidenza valutativa. In questa prospettiva, l'attenzione si concentra sui limiti strutturali dei sistemi algoritmici: bias nei dati di addestramento, opacità dei modelli, automation bias, difficoltà di spiegazione, asimmetrie di potere e potenziali effetti di tecnocratizzazione.

Questo orientamento affonda le proprie radici nei lavori che, già nel contesto dei Big Data, mettevano in guardia contro una lettura ingenuamente positivista dell'innovazione digitale. Picciotto (2020) ha sottolineato come la disponibilità di grandi volumi di dati non risolva problemi fondamentali di veracity, volatility e validity, evidenziando il rischio che l'apparente potenza analitica sostituisca una riflessione sui costrutti valutativi. In modo analogo, Bamberger (2016) osservava che la validazione tecnica degli output non equivale a una validazione valutativa dell'evidenza.

Nel dibattito più recente sull'IA, queste preoccupazioni si intensificano. Azzam (2023) discute in modo sistematico le implicazioni dell'IA per la validità, mostrando come modelli opachi possano produrre risultati formalmente coerenti ma epistemicamente fragili. Greenstein e Cho (2025) collocano l'uso dell'IA entro un quadro di etica ed equità, evidenziando come bias algoritmici e automation bias possano rafforzare disuguaglianze esistenti e ridurre la trasparenza dei processi valutativi. Rompczyk (2025) interroga la compatibilità tra l'uso dell'IA e gli standard professionali della valutazione, mettendo in luce il rischio di un allentamento dei principi di indipendenza, trasparenza e responsabilità. Potluka *et al.* (2025) evidenziano l'urgenza di criteri normativi che governino l'uso dell'IA, riconoscendo implicitamente che i rischi epistemici ed etici non possono essere risolti solo sul piano tecnico. Anche qui, tuttavia, il focus resta prevalentemente sui primi stadi del framework: identificare i rischi più che governare l'intero percorso inferenziale.

In questa configurazione, l'IA si presenta come un elemento ambiguo per la pratica valutativa: introduce opacità dove la valutazione rivendica trasparenza, automatismo dove è richiesto giudizio ed efficienza dove servono tempi di riflessione. Il baricentro dell'argomentazione si colloca sugli stadi 1 e 2 del

framework, ma in senso rovesciato rispetto al filone strumentale: non per esaltarne le potenzialità, bensì per metterne in luce le fragilità.

Questi contributi svolgono una funzione cruciale di problematizzazione. Rendono visibili i rischi epistemici ed etici che accompagnano l'utilizzo dell'IA e contrastano la naturalizzazione dell'output algoritmico come evidenza neutrale. Tuttavia, essi tendono a collocare l'IA come elemento esterno rispetto alla pratica valutativa, rispetto al quale adottare una postura prevalentemente difensiva.

Dal punto di vista del framework, questa letteratura illumina con forza le tensioni degli stadi iniziali, opacità dell'output, confusione tra performance e validità, ma lascia relativamente indeterminato il modo in cui tali rischi possano essere governati all'interno di disegni valutativi concreti. L'attenzione è rivolta più ai pericoli che ai processi attraverso cui l'evidenza viene costruita.

Ne risulta una prospettiva che, pur offrendo diagnosi convincenti, fatica a tradursi in una grammatica operativa per la valutazione. Il passaggio dall'output al giudizio resta un problema evocato ma non articolato. L'IA è qualcosa da contenere, limitare o sorvegliare, più che un oggetto da integrare entro un percorso interpretativo. In termini di prospettiva analitica, questa configurazione rende visibili le fragilità dei primi due stadi, ma lascia in ombra quelli successivi: interpretazione, triangolazione e giudizio. La critica resta concentrata sull'algoritmo, mentre il lavoro interpretativo della valutazione rimane sullo sfondo. È proprio questo scarto che apre lo spazio per una concettualizzazione più integrata del governo dell'evidenza.

2.3 L'IA come oggetto di valutazione e governance

Un terzo insieme di contributi propone un cambio di prospettiva più radicale, concettualizzando l'IA come oggetto diretto di valutazione. In questo filone, la comunità valutativa rivendica un ruolo specifico: non si limita a utilizzare l'IA come strumento, ma mette a disposizione criteri, framework e approcci metodologici per analizzarne funzionamento, impatti e implicazioni sociali. Questa prospettiva è particolarmente visibile nei contributi che applicano approcci standards e criteria based all'analisi dei sistemi di IA, enfatizzando dimensioni quali spiegabilità, equità, affidabilità e coerenza con obiettivi di policy (Montrosse-Moorhead, 2023).

In modo analogo, Pudney *et al.* (2025) sviluppano un quadro per la valutazione dei sistemi di IA in ambito pubblico, insistendo sulla necessità di integrare criteri tecnici con criteri normativi e sociali. Nielsen (2025) colloca

esplicitamente la valutazione come attore chiave nella governance dell'IA, capace di connettere performance tecnica, impatti sociali e decisioni di policy.

In questa prospettiva, la valutazione recupera con forza il proprio mandato normativo. Gli stadi 4 e 5 del framework – triangolazione e giudizio – emergono come centrali: l'IA viene sottoposta a confronto con altre fonti di evidenza, con valori pubblici e con aspettative degli stakeholder, e diventa oggetto di decisioni esplicite. Rispetto ai due filoni precedenti, questa configurazione sposta il baricentro dall'uso dell'IA alla sua regolazione. L'attenzione non è più rivolta solo a ciò che l'IA consente di fare, né esclusivamente ai rischi che introduce, ma alla possibilità di governarla attraverso pratiche valutative.

Tuttavia, anche in questo filone il passaggio dagli output algoritmici all'evidenza tende a rimanere implicito. I sistemi di IA vengono trattati come oggetti complessi da giudicare, ma raramente si tematizza il modo in cui, nelle valutazioni ordinarie, gli output prodotti da tali sistemi attraversino l'intera catena inferenziale. In altre parole, la prospettiva della governance rafforza la dimensione del giudizio, ma non rende necessariamente visibile il percorso che conduce dall'analisi computazionale a tale giudizio. Il rischio è che la valutazione si collochi a valle dei sistemi, pronunciandosi sui loro impatti, senza interrogare sistematicamente il modo in cui l'evidenza algoritmica viene costruita e utilizzata nei processi valutativi quotidiani.

In termini di framework, questa configurazione illumina gli stadi finali – triangolazione e giudizio – ma lascia relativamente in ombra quelli intermedi: interpretazione theory-informed e costruzione progressiva dell'evidenza. La governance si esercita sull'oggetto "IA", mentre la catena inferenziale che trasforma output in evidenza resta in larga parte tacita. È proprio questa asimmetria che completa il quadro: ciascun filone rende visibili alcune dimensioni e ne oscura altre. Nessuno, preso singolarmente, tematizza l'intero percorso che conduce dall'output algoritmico al giudizio valutativo.

2.4 Frammentazione concettuale e punto cieco comune

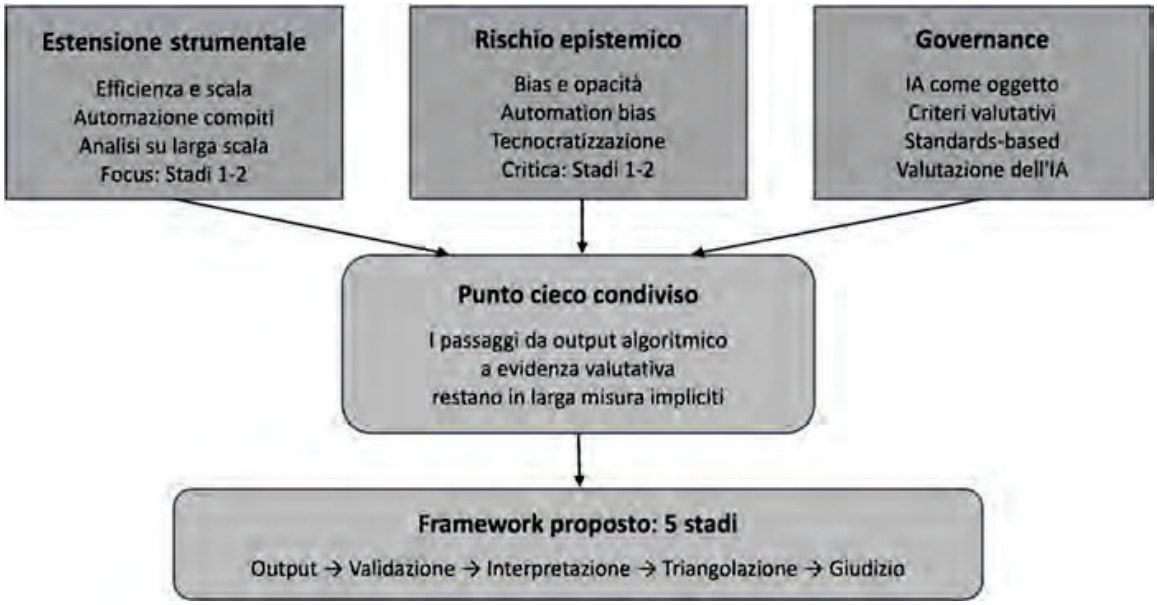
Considerate nel loro insieme, queste tre modalità di concettualizzare l'IA mostrano come la letteratura valutativa sia caratterizzata da una significativa frammentazione concettuale. I contributi differiscono per obiettivi, linguaggi disciplinari e livelli di astrazione, ma condividono un limite ricorrente: la scarsa attenzione ai passaggi attraverso cui gli output algoritmici vengono trasformati in evidenza valutativa. Ogni prospettiva illumina alcuni stadi del framework e ne oscura altri:

- l'estensione strumentale privilegia output e validazione tecnica;

- la prospettiva critica problematizza opacità e bias ma resta esterna ai processi;
- l'IA come oggetto di valutazione enfatizza giudizio e governance senza esplicitare la catena inferenziale.

Il risultato è che il percorso che conduce dall'analisi computazionale al giudizio rimane spesso in larga misura implicito, come rappresentato nella schema sottostante (fig. 1). In questo spazio ancora poco tematizzato si colloca il contributo di questo articolo, che si propone di rendere più espliciti e discutibili i passaggi attraverso cui l'informazione algoritmica viene tradotta in evidenza valutativa.

Fig. 1 - Tre modalità di concettualizzare l'IA nella valutazione



La frammentazione appena descritta suggerisce che il problema non risieda tanto nella presenza dell'IA in sé, quanto nell'assenza di una concettualizzazione condivisa dei passaggi che trasformano l'analisi in evidenza. Per rendere questo nodo maggiormente osservabile e discutibile, appare quindi necessario spostare l'attenzione dagli strumenti ai processi inferenziali che li attraversano.

3. Dagli output algoritmici all'evidenza: il problema

L'evidenza valutativa non coincide con la mera disponibilità di dati o risultati analitici, ma è il prodotto di un processo che integra dati empirici, quadri interpretativi, criteri espliciti e valori normativi. Già Bamberger (2016) e

Picciotto (2020) hanno sottolineato come la qualità dell'evidenza non sia riducibile alla qualità tecnica delle misure, ma dipenda dalla capacità di collegare osservazioni, teoria e giudizio. L'introduzione dell'IA non elimina tali passaggi, ma tende piuttosto a renderli meno visibili, aumentando il rischio di una sovrapposizione impropria tra output algoritmico ed evidenza (Mazzeo Rinaldi & Nielsen, 2025). Classificazioni, cluster, predizioni e testi generati possono essere utilizzati direttamente per informare conclusioni valutative, producendo una compressione epistemica in cui i passaggi interpretativi, inferenziali e valutativi tendono a essere implicitamente fusi: l'output algoritmico viene così tacitamente assimilato all'evidenza. In questo slittamento non risiede semplicemente un rischio tecnico, ma una trasformazione del modo in cui l'inferenza valutativa viene costruita e legittimata.

Le preoccupazioni espresse da Picciotto (2020) rispetto ai Big Data trovano qui una nuova intensità: la potenza analitica può talvolta ridurre lo spazio dedicato alla riflessione sui costrutti e sui criteri di valore. La distinzione tra "Evaluation needs Big Data" e "Big Data needs evaluation" richiama la necessità di un governo metodologico bidirezionale dei passaggi che conducono dall'output all'evidenza. Azzam (2023) e Greenstein e Cho (2025) mostrano come l'opacità algoritmica e l'automation bias rafforzino questa dinamica, rendendo meno trasparenti i passaggi interpretativi che tradizionalmente fondano la credibilità dell'evidenza.

Il problema centrale non è dunque se l'IA produca risultati accurati, ma come tali risultati vengano trasformati in affermazioni valutative. Quando l'output computazionale viene trattato come evidenza autosufficiente, la valutazione rischia di perdere la propria specificità interpretativa, riducendosi a un'estensione dell'analisi dei dati. Per discutere questo passaggio, può essere utile distinguere analiticamente cinque stadi attraverso cui un risultato computazionale può essere trasformato in evidenza valutativa:

- Produzione dell'output algoritmico
- Validazione tecnica
- Interpretazione theory-informed
- Triangolazione
- Giudizio valutativo

Questi stadi non descrivono una sequenza operativa rigida, ma individuano dimensioni epistemiche che possono essere rese più visibili e maggiormente tematizzate. In questo senso, essi ricompongono in forma esplicita aspetti che nella tradizione valutativa risultano spesso distribuiti tra disegno, analisi e uso dell'evidenza (Bamberger, 2016; York & Bamberger, 2020; Picciotto, 2020).

La loro mancata esplicitazione non costituisce soltanto un limite procedurale, ma può incidere sulla trasparenza e sulla legittimazione dell'evidenza prodotta. Rendere esplicita la catena che conduce dall'output al giudizio significa, in questa prospettiva, riaffermare che anche nell'era dell'IA, la valutazione resta una pratica di interpretazione e attribuzione di valore, non una mera funzione di calcolo.

4. Il framework: cinque stadi tra funzione e tensione epistemica

Il framework proposto non descrive una successione standard da applicare meccanicamente a ogni valutazione. Piuttosto, prova a rendere osservabili cinque dimensioni che, nella pratica, possono essere fuse, compresse o solo parzialmente esplicitate. La sua funzione non è prescrittiva, ma analitica: offrire una possibile prospettiva per interrogare il passaggio dall'output algoritmico al giudizio valutativo. Ciascuno stadio individua una funzione che può contribuire alla trasformazione dell'informazione in evidenza, ma richiama anche alcune tensioni che, se non adeguatamente considerate, possono incidere sulla trasparenza e legittimità del processo valutativo.

Tab. 2 - Cinque stadi della trasformazione dell'output algoritmico in evidenza valutativa

Stadio	***Funzione analitica***	***Possibile compressione epistemica***
Output algoritmico	Produce pattern, stime, classificazioni o testi	Naturalizzazione dell'output come dato
Validazione tecnica	Verifica accuratezza, robustezza e bias	Confusione tra performance tecnica e validità valutativa
Interpretazione theory-informed	Collega l'output a teoria di programma, contesto e meccanismi	Riduzione del significato a correlazione o pattern
Triangolazione	Confronta l'output con fonti, metodi e prospettive diverse	Autosufficienza dell'algoritmo
Giudizio valutativo	Formula conclusioni di valore sulla base di criteri espliciti	Delega implicita del giudizio

Un esempio illustrativo può aiutare a chiarire come i diversi stadi del framework possano intrecciarsi nella pratica valutativa e dove si producano possibili forme di compressione epistemica.

Box 1 – Output algoritmico ed evidenza valutativa

Si consideri una valutazione di un programma di formazione professionale che utilizza strumenti di natural language processing per analizzare migliaia di commenti aperti raccolti attraverso questionari online. Il sistema algoritmico produce cluster tematici e indicatori di sentiment, identificando automaticamente pattern ricorrenti nelle esperienze dei partecipanti.
In assenza di ulteriori passaggi interpretativi, tali output possono essere assunti direttamente come evidenza della qualità del programma. Tuttavia, il fatto che un commento venga classificato come "positivo" o associato a un determinato tema non chiarisce necessariamente il significato valutativo di tali risultati, né il modo in cui essi si colleghino agli obiettivi del programma, ai meccanismi di cambiamento attesi o alle differenze tra gruppi di partecipanti. È proprio in questo spazio tra output computazionale e giudizio valutativo che si collocano i passaggi di interpretazione, triangolazione e attribuzione di valore che il framework proposto mira a rendere maggiormente visibili.

4.1 Output algoritmico: potenza analitica e opacità incorporata

Il primo stadio riguarda la produzione dell'output da parte del sistema di IA. Che si tratti di classificazioni, cluster, stime predittive o testi generati, tali output sono il risultato di una serie di scelte: selezione delle variabili, definizione delle categorie, metriche di ottimizzazione, architettura del modello, dati di addestramento.

Dal punto di vista valutativo, il punto cruciale non è tanto cosa produce l'algoritmo, quanto come tali risultati incorporino assunzioni che restano spesso invisibili. Nei modelli di machine learning tradizionali, queste assunzioni riguardano il modo in cui il fenomeno viene reso computabile (Wencker *et al.*, 2025); nei modelli generativi, esse si manifestano nella capacità di produrre output semanticamente plausibili anche in assenza di un ancoraggio empirico verificabile.

La letteratura sull'IA ha ampiamente mostrato che l'output algoritmico non è mai neutro: esso riflette distribuzioni di dati, scelte di design e trade-off tra accuratezza, semplicità e generalizzabilità. Sul piano valutativo, ciò implica che l'output non può essere trattato come "dato grezzo". Esso è già una forma di interpretazione automatizzata.

La tensione strutturale di questo stadio risiede nell'opacità: quanto più l'output appare sofisticato e fluente, tanto più è difficile ricostruirne le premesse. In assenza di un lavoro di esplicitazione, l'output tende a essere assunto come evidenza in sé, innescando un processo di trasformazione che salta i passaggi intermedi.

4.2 Validazione tecnica: necessaria ma non sufficiente

Il secondo stadio riguarda la validazione tecnica dell'output algoritmico. In questa fase vengono valutate accuratezza, stabilità, robustezza e presenza di bias sistematici. Nella letteratura sulla data science, questo livello di analisi è ampiamente sviluppato e rappresenta un prerequisito fondamentale per qualsiasi utilizzo responsabile degli algoritmi (Schwartz *et al.*, 2022). Tuttavia, dal punto di vista valutativo, esso rappresenta una condizione necessaria, ma non sufficiente.

Un modello può risultare altamente performante rispetto alle metriche adottate e, al tempo stesso, produrre informazioni poco rilevanti o fuorvianti rispetto alle domande che la valutazione è chiamata ad affrontare. La tensione strutturale di questo stadio risiede proprio nella possibile sovrapposizione tra performance tecnica e validità valutativa.

Questa distinzione era già stata evidenziata da Bamberger (2016) nel contesto dei Big Data, quando osservava che la validazione tecnica degli output non equivale a una validazione valutativa dell'evidenza prodotta. L'IA tende a rendere questa distinzione meno visibile: l'eccellenza tecnica può essere scambiata per solidità dell'evidenza. Il framework consente di localizzare questo slittamento: l'evidenza diventa fragile non perché l'algoritmo "funziona male", ma perché la validazione si arresta al livello tecnico.

4.3 Interpretazione contestuale e theory-informed

Il terzo stadio introduce un passaggio tipicamente valutativo: l'interpretazione degli output algoritmici alla luce del contesto e della teoria di programma. In questa fase, i risultati computazionali cessano di essere meri pattern e vengono collocati all'interno di un quadro esplicativo più ampio.

La letteratura valutativa sulla theory-based ha da tempo sottolineato che la comprensione dei meccanismi causali e delle condizioni contestuali è essenziale per attribuire significato ai risultati empirici (Chen, 1990; Pawson & Tilley, 1997). In assenza di questo lavoro interpretativo, l'IA rischia di rafforzare un uso meramente correlazionale dell'evidenza, privilegiando pattern predittivi rispetto alla comprensione dei processi sottostanti poiché i modelli tradizionali di machine learning eccellono nell'identificazione di correlazioni ma non distinguono intrinsecamente se queste rappresentino relazioni causali (Pearl & Mackenzie, 2018). La tensione strutturale di questo stadio è dunque la riduzione correlazionale del significato. Un secondo esempio illustrativo

consente di osservare questa tensione nel caso dell'analisi qualitativa assistita da IA generativa.

Box 2 – IA generativa e interpretazione qualitativa

Si consideri una valutazione qualitativa basata su interviste in profondità a beneficiari di un programma sociale. Un large language model viene utilizzato per codificare automaticamente le trascrizioni, identificare temi ricorrenti e produrre sintesi interpretative.
In questo caso, il sistema non si limita a organizzare informazioni, ma partecipa direttamente alla costruzione del significato analitico. Le categorie generate possono apparire coerenti e plausibili, inducendo il valutatore ad assumerle implicitamente come interpretazioni già valide.
Tuttavia, il passaggio tra similarità linguistica e significato valutativo rimane teoricamente e contestualmente mediato. Le categorie prodotte dall'IA possono riflettere regolarità semantiche senza cogliere meccanismi causali, relazioni di potere o dimensioni normative rilevanti per il giudizio.
In questo senso, il framework proposto aiuta a distinguere il momento della produzione automatizzata di pattern dal lavoro interpretativo attraverso cui tali pattern vengono trasformati in evidenza valutativa.

Quando l'interpretazione theory-informed è assente o debole, l'output viene letto come risposta autosufficiente. In questo modo, la potenza predittiva sostituisce la comprensione dei processi. Il framework rende visibile questo slittamento: ciò che manca non è informazione, ma significato valutativo.

4.4 Triangolazione e confronto tra fonti

Il quarto stadio riguarda la triangolazione con altre fonti, metodi e prospettive. In una prospettiva valutativa, nessuna fonte informativa, algoritmica o meno, può essere considerata autosufficiente, la credibilità emerge dal confronto (Picciotto, 2020). La letteratura valutativa mostra come la triangolazione rappresenti un meccanismo chiave per ridurre il rischio di interpretazioni distorte e per rafforzare la credibilità dell'evidenza (Patton, 1999). L'IA introduce una tensione specifica: la pressione all'autosufficienza dell'algoritmo. La scala, la velocità e la granularità degli output possono indurre una fiducia sproporzionata, rafforzata da fenomeni di automation bias (Goddard *et al.*, 2012; Parasuraman & Manzey, 2010). La triangolazione diventa allora non solo una buona pratica, ma un presidio critico contro l'autosufficienza dell'output algoritmico. Come mostrato anche negli esempi illustrativi precedenti, la disponibilità di output altamente granulari o semanticamente plausibili può rafforzare la tendenza a trattare l'analisi algoritmica come fonte autosufficiente di

evidenza. Il framework consente di analizzare quando e come l'output algoritmico viene sottratto al confronto, diventando una fonte privilegiata o esclusiva. In tali casi, la fragilità dell'evidenza non dipende dall'assenza di dati, ma dalla riduzione del pluralismo informativo.

4.5 Giudizio valutativo: il nodo della responsabilità

Il quinto stadio riguarda la formulazione del giudizio valutativo, in cui l'evidenza viene messa in relazione con criteri espliciti, valori e standard condivisi. Questo passaggio implica una responsabilità normativa e istituzionale che non può essere delegata ai sistemi di IA senza snaturare la funzione stessa della valutazione (Montrosse-Moorhead, 2023; Nielsen, 2025). La tensione strutturale risiede nella delegabilità implicita del giudizio. Quando le conclusioni sono fortemente mediate da sistemi opachi, il confine tra supporto decisionale e sostituzione del giudizio tende a sfumare. Il rischio non è che l'IA "decida", ma che il giudizio venga presentato come tecnicamente necessario. Il framework rende visibile questo punto: il giudizio non è un esito automatico dell'analisi. È un atto che implica responsabilità. Rendere esplicito lo stadio finale significa riaffermare che, anche nell'era dell'IA, la valutazione resta una pratica di attribuzione di valore.

5. Implicazioni metodologiche per la valutazione

Il framework proposto non mira soltanto a chiarire un nodo epistemico astratto, ma suggerisce anche alcune possibili implicazioni per il modo in cui la valutazione viene progettata, condotta e utilizzata in contesti mediati da sistemi algoritmici. Tali implicazioni non riguardano tanto l'adozione di specifici strumenti, quanto l'attenzione ai processi attraverso cui si passa dall'output al giudizio valutativo. In questa prospettiva, le principali implicazioni possono essere ricondotte a quattro ambiti: il disegno valutativo; la costruzione dell'evidenza; l'uso dell'evidenza e le relazioni con gli stakeholders; il ruolo professionale del valutatore.

5.1 Implicazioni per il disegno valutativo

L'integrazione dell'IA nei processi valutativi incide, anzitutto, sul disegno della valutazione. L'uso di sistemi algoritmici tende infatti a spostare

l'attenzione verso fasi analitiche avanzate, rischiando di comprimere o dare per scontate le scelte di scoping, di formulazione delle domande valutative e di esplicitazione delle assunzioni teoriche. In contesti caratterizzati da abbondanza di dati e potenza computazionale, il disegno rischia di diventare reattivo agli strumenti disponibili: ciò che è misurabile o modellizzabile orienta ciò che viene considerato rilevante. In questo senso, il framework può offrire una chiave interpretativa utile a rendere più osservabile tale slittamento (York & Bamberger, 2020). Quando gli stadi relativi all'interpretazione, alla triangolazione e al giudizio risultano deboli o poco esplicitati, il processo valutativo rischia infatti di essere orientato prevalentemente dall'analisi algoritmica, con domande valutative che emergono ex post a partire da ciò che l'algoritmo riesce a identificare. In questa prospettiva, la funzione del disegno valutativo si rafforza: esso diventa un vincolo epistemico che precede e orienta l'uso dell'IA. Progettare una valutazione in contesti algoritmici può quindi implicare la necessità di chiarire:

- quali domande di valore devono essere affrontate;
- quali dimensioni del programma risultano osservabili tramite IA e quali tendono invece a rimanere fuori campo;
- in che modo gli output computazionali vengano collegati a una teoria di programma esplicita.

L'IA, in questo senso, non sembra ridurre l'importanza del disegno valutativo; al contrario, può rendere ancora più rilevante la necessità di esplicitarne presupposti, criteri e finalità. Il framework può contribuire a individuare situazioni in cui il disegno appare fortemente orientato dagli strumenti disponibili e a riportare l'attenzione sulle finalità valutative che guidano il processo.

Sul piano pratico, questo comporta anche una dimensione relazionale: la definizione di domande valutative e teoria di programma non è solo un esercizio interno al valutatore, ma un confronto esplicito con committenti e stakeholder, tanto più necessario quando l'IA rende disponibili nuove fonti o granularità di dati capaci di orientare surrettiziamente il fuoco dell'indagine.

5.2 Costruzione dell'evidenza: validità oltre la performance

Un secondo insieme di implicazioni riguarda la validità dell'evidenza valutativa prodotta attraverso sistemi di IA. Come discusso nella sezione precedente, la validazione tecnica degli output rappresenta solo uno dei livelli necessari per garantire la qualità dell'evidenza. Dal punto di vista valutativo, la qualità dell'evidenza dipende dall'allineamento tra ciò che viene modellizzato

e i costrutti teorici e normativi rilevanti per il giudizio (Patton, 2008; Whittemore *et al*., 2023). In questo senso, l'uso dell'IA tende a rendere più visibili alcune tensioni classiche della valutazione, in particolare quella tra precisione analitica e rilevanza per il giudizio. Output altamente performanti possono infatti risultare epistemicamente fragili quando i passaggi di interpretazione e triangolazione rimangono deboli o poco esplicitati. Nei casi illustrativi discussi, tale fragilità non deriva necessariamente da errori tecnici dell'algoritmo, ma dalla parziale esplicitazione dei passaggi che collegano pattern computazionali e giudizio valutativo. In questa prospettiva, il framework può aiutare a individuare dove si collocano tali fragilità: non necessariamente nel funzionamento dell'algoritmo, quanto piuttosto nella parziale esplicitazione dei processi che collegano l'output al giudizio valutativo.

Sul piano operativo, ciò suggerisce l'importanza che la documentazione valutativa renda maggiormente espliciti:

- le assunzioni incorporate nei modelli;
- i criteri attraverso cui l'output viene interpretato;
- le fonti con cui viene triangolato;
- il modo in cui tali elementi informano il giudizio.

In assenza di questa esplicitazione, l'uso dell'IA potrebbe favorire forme dinamiche di tecnocratizzazione del processo valutativo, riducendo lo spazio per il confronto critico e l'uso riflessivo dell'evidenza. Sul piano pratico, ciò richiede di riconsiderare le modalità di triangolazione, affiancando sistematicamente agli output algoritmici metodi qualitativi e approcci interpretativi – interviste, focus group, osservazione – non come verifica accessoria, ma come componente costitutiva della costruzione dell'evidenza, da concordare con i committenti fin dal disegno.

5.3 Uso dell'evidenza e relazioni con gli stakeholder

Un ulteriore insieme di implicazioni riguarda l'uso dell'evidenza e le relazioni con gli stakeholder. La produzione algoritmica di output ad alta granularità può infatti rafforzare forme di razionalizzazione tecnica, riducendo lo spazio per il confronto interpretativo e per la discussione sui criteri di valore sottesi ai risultati.

In questa prospettiva, il framework suggerisce l'opportunità di considerare l'uso dell'evidenza come parte integrante della catena inferenziale. Rendere maggiormente visibili i diversi stadi può contribuire a trasformare l'output algoritmico in un oggetto di discussione e interpretazione, piuttosto che in

"autorità silenziosa". In particolare, interpretazione e triangolazione possono essere lette come spazi deliberativi in cui diversi attori possono interrogare significato, limiti e implicazioni dei risultati.

Da questo punto di vista, l'IA non conduce inevitabilmente alla chiusura del processo valutativo; può, se governata, aprire nuove forme di dialogo basate su evidenze più ricche ma anche più controverse.

Questo implica, sul piano pratico, di ricondurre al confronto con gli stakeholder anche momenti che la prassi valutativa tende a dare per acquisiti a monte, come la definizione della teoria di programma e dei criteri di giudizio, tanto più quando questi vengono messi in tensione da output algoritmici ad alta granularità.

5.4 Ruolo professionale del valutatore

Il quarto ambito riguarda il ruolo professionale del valutatore e la distribuzione delle responsabilità nei processi valutativi mediati dall'IA. L'introduzione di sistemi algoritmici tende infatti a ridefinire, almeno in parte, la distribuzione del lavoro cognitivo, spostando alcune attività analitiche verso strumenti automatizzati e richiedendo al tempo stesso competenze di supervisione, interpretazione e governance.

In questa prospettiva, il ruolo del valutatore non sembra necessariamente ridursi, quanto piuttosto trasformarsi. La letteratura più recente sottolinea, in particolare, come tale evoluzione implichi anche nuove competenze: alfabetizzazione digitale, capacità di interrogare criticamente gli output algoritmici e collaborazione interdisciplinare, con attenzione a privacy, bias e responsabilità (Montrosse-Moorhead, 2023; Mazzeo Rinaldi & Nielsen, 2025; Nielsen, 2025).

In questo scenario, il contributo specifico del valutatore potrebbe risiedere non tanto nella produzione di analisi sempre più sofisticate, quanto nella capacità di rendere espliciti e governare i passaggi che collegano i diversi stadi del processo valutativo. In questo senso, il valutatore può essere interpretato come una figura di mediazione interpretativa, chiamata a esplicitare assunzioni, limiti, criteri interpretativi e responsabilità.

Il framework, da questo punto di vista, può aiutare a individuare situazioni in cui tale funzione tende a indebolirsi: ad esempio quando l'output algoritmico viene fatto convergere direttamente verso conclusioni valutative, riducendo la visibilità dei passaggi interpretativi intermedi. La questione, allora, non riguarda tanto l'uso o meno dell'IA, quanto le modalità attraverso cui viene governato il passaggio dall'informazione al giudizio.

Le competenze richieste non si esauriscono quindi nell'alfabetizzazione digitale individuale, ma includono la capacità di gestire il rapporto con gli stakeholder attorno a questi passaggi – teoria di programma, criteri, triangolazione – ridefinendo il ruolo del valutatore anche come funzione di mediazione tra analisi algoritmica e comunità degli interlocutori della valutazione.

Le implicazioni metodologiche discusse mostrano come la prospettiva analitica proposta possa contribuire a riorientare il disegno, la costruzione e l'uso dell'evidenza. Resta tuttavia importante interrogare criticamente anche il framework stesso, le sue condizioni di applicabilità e i limiti che lo attraversano. La sezione seguente affronta queste questioni.

6. Discussione critica del framework

Il framework proposto non va inteso come una sequenza operativa da applicare integralmente a ogni valutazione. La sua finalità è prevalentemente analitica: offrire una chiave di lettura per rendere più visibili alcune dimensioni epistemiche che tendono a rimanere implicite quando l'evidenza è mediata da sistemi algoritmici. Uno dei rischi principali è interpretare il framework come una definizione di "buona pratica". I cinque stadi non descrivono necessariamente ciò che dovrebbe accadere in ogni valutazione, ma ciò che può essere distinto quando si analizza il passaggio dall'output al giudizio. Nella pratica, tali stadi risultano spesso intrecciati: l'interpretazione può avvenire durante la validazione tecnica; la triangolazione può essere incorporata in procedure automatizzate; il giudizio può essere parzialmente inscritto nelle metriche utilizzate. Il framework non intende negare questa realtà operativa, quanto piuttosto offrire strumenti per osservarla e discuterla. In questo senso, può aiutare a individuare dove avvengano forme di compressione epistemica e quali effetti possano avere sulla trasparenza e sulla legittimazione dell'evidenza. La piena articolazione dei cinque stadi presuppone inoltre condizioni che non sempre risultano disponibili: tempo, competenze interdisciplinari, spazi deliberativi e committenze disposte a confrontarsi con margini di incertezza. In molte valutazioni tali condizioni sono solo parzialmente presenti. Il framework non propone quindi una soluzione generale, ma uno strumento analitico per rendere più espliciti i trade-off che attraversano il processo valutativo: aiuta a riconoscere che fermarsi alla sola validazione tecnica può riflettere non solo vincoli operativi, ma anche specifiche scelte istituzionali. Allo stesso tempo, il framework non elimina il problema dell'opacità algoritmica né garantisce che il giudizio rimanga pienamente umano. Esso può, contribuire a rendere maggiormente visibili le modalità attraverso cui alcune funzioni interpretative vengono

delegate ai sistemi algoritmici. La sua utilità dipende quindi anche dalla disponibilità istituzionale a utilizzare tale visibilità come occasione di discussione, riflessività e assunzione di responsabilità. Il contributo teorico del framework può essere letto come un tentativo di spostare l'attenzione dal solo controllo tecnico dell'IA alle modalità di costruzione e governo epistemico dell'evidenza, in linea con le più recenti proposte di collocare la valutazione come attore chiave nella governance dei sistemi algoritmici (Pudney *et al.*, 2025; Nielsen, 2025). Una parte della letteratura si interroga soprattutto sull'accuratezza, l'equità o la spiegabilità degli algoritmi. Il framework prova ad aggiungere una domanda complementare: attraverso quali passaggi l'output viene trasformato in evidenza valutativa, e chi governa tale trasformazione. Da questo punto di vista, il framework non si pone in alternativa agli approcci regolativi o tecnici, ma tenta piuttosto di integrarli entro una prospettiva valutativa. In questa prospettiva, la legittimità dell'evidenza non dipende esclusivamente dalle proprietà dell'algoritmo, ma anche dai processi inferenziali e istituzionali attraverso cui l'output viene interpretato e utilizzato.

Conclusioni

L'introduzione dell'IA nei processi valutativi non sembra poter essere interpretata soltanto come un'innovazione tecnica o come una semplice estensione degli strumenti di analisi disponibili. Come emerge dalla sintesi critica della letteratura, il dibattito sull'IA nella valutazione è caratterizzato da una pluralità di approcci e concettualizzazioni che, pur evidenziando potenzialità rilevanti, tendono spesso a concentrarsi sugli strumenti e sugli output prodotti, lasciando meno esplorati i passaggi attraverso cui tali output vengono trasformati in evidenza valutativa. L'analisi proposta ha evidenziato una certa frammentazione epistemica del campo, articolabile, almeno in forma analitica, in tre configurazioni principali: l'IA come estensione strumentale, l'IA come rischio interpretativo ed etico, e l'IA come oggetto di valutazione e governance. Ciascun filone mette in luce aspetti importanti: il primo richiama le potenzialità offerte dalla scala e dalla capacità analitica; il secondo rende visibili questioni legate a bias, opacità e possibili dinamiche di tecnicizzazione; il terzo riporta l'attenzione sulla dimensione normativa e sul ruolo della valutazione nella governance dei sistemi algoritmici. La review suggerisce tuttavia l'esistenza di un elemento ancora relativamente poco tematizzato. I passaggi attraverso cui l'output algoritmico viene trasformato in evidenza valutativa rimangono spesso impliciti. L'IA tende così a essere alternativamente valorizzata, problematizzata o regolata, ma più raramente collocata entro un processo

esplicito che colleghi analisi, interpretazione e giudizio. In assenza di tale esplicitazione, l'output rischia di essere assunto come dato autosufficiente oppure trattato come un oggetto esterno al processo valutativo, mentre il lavoro interpretativo che sostiene la costruzione dell'evidenza rimane in secondo piano. Il framework dei cinque stadi proposto in questo articolo nasce dal tentativo di rendere maggiormente osservabili i passaggi attraverso cui l'informazione algoritmica viene trasformata in evidenza valutativa. Distinguere output, validazione, interpretazione, triangolazione e giudizio significa riportare l'attenzione su dimensioni che l'uso dell'IA può comprimere o rendere meno esplicite. Tale esplicitazione non riguarda soltanto la pratica cognitiva del singolo valutatore, ma investe anche il modo in cui teoria di programma, criteri e triangolazione vengono negoziati con gli stakeholder, ridefinendo in parte il management stesso del processo valutativo.

In questa prospettiva, il contributo dell'articolo non riguarda soltanto l'analisi delle implicazioni tecniche dell'IA, ma la possibilità di interrogare criticamente i processi attraverso cui l'evidenza viene costruita, interpretata e utilizzata nei contesti valutativi. Il framework proposto non intende offrire una nuova prescrizione metodologica, ma una possibile chiave di lettura per analizzare dove e come tali processi vengano compressi, delegati o opacizzati. Ulteriori sviluppi empirici potranno contribuire a verificare come le diverse forme di compressione epistemica si manifestino concretamente in specifici contesti valutativi e organizzativi. Nell'era dell'IA, la questione centrale per la valutazione potrebbe non riguardare soltanto l'adozione di nuovi strumenti, ma la capacità di mantenere visibili, discutibili e governabili i passaggi attraverso cui l'informazione viene trasformata in evidenza e l'evidenza in giudizio.

Dichiarazioni

L'autore dichiara che il Software ChatGPT (GPT-5.6 Sol, OpenAI) è stato utilizzato esclusivamente per limitate attività di revisione linguistica e sintesi. Non sono stati inseriti dati riservati. L'autore si assume la piena responsabilità del manoscritto finale.

Riferimenti bibliografici

Athey, S., Berg, A. (2023). AI and the transformation of social science research: Careful bias management and data fidelity are key. *Science*, 381(6663), 1319–1322. https://doi.org/10.1126/science.adi1778

Azzam, T. (2023). Artificial intelligence and validity. *New Directions for Evaluation*, 178–179, 85–95. https://doi.org/10.1002/ev.20565

Bamberger, M. (2016*). Integrating big data into the monitoring and evaluation of development programmes*. UN Global Pulse.

Goddard, K., Roudsari, A., Wyatt, J.C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Grant, M.J., Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal*, 26(2), 91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

Greenhalgh, T., Thorne, S., Malterud, K. (2018). Time to challenge the spurious hierarchy of systematic over narrative reviews? *European Journal of Clinical Investigation*, 48, e12931. https://doi.org/10.1111/eci.12931

Greenstein, N., Cho, S.W. (2025). Ethics & equity in data science for evaluators. In S.B. Nielsen, F. Mazzeo Rinaldi, G. J. Petersson (Eds.), *Artificial intelligence and evaluation: Emerging technologies and their implications for evaluation* (pp. 56–77). Routledge. https://doi.org/10.4324/9781003512493

Leeuw, F.L., Bamberger, M. (2025). *Artificial intelligence and big data: Lessons from evaluations of rule of law and development*. Edward Elgar.

Mason, S., Montrosse-Moorhead, B. (Eds.). (2023). Artificial intelligence and evaluation (Special issue). New Directions for Evaluation, 178–179, 1–134. https://doi.org/10.1002/ev.20554

Mazzeo Rinaldi, F. (2018). Big Data e valutazione: Una relazione ancora da costruire. *RIV Rassegna Italiana di Valutazione*, 68, 83–100.

Mazzeo Rinaldi, F., Nielsen, S.B. (2025). Artificial intelligence: Challenges for evaluators. In S. B. Nielsen, F. Mazzeo Rinaldi, G.J. Petersson (Eds.), *Artificial intelligence and evaluation: Emerging technologies and their implications for evaluation* (pp. 287–306). Routledge. https://doi.org/10.4324/9781003512493

Mazzeo Rinaldi, F., Occhipinti, O. (2023). Big Data, intelligenza artificiale e valutazione: Cosa accade in Italia. *RIV Rassegna Italiana di Valutazione*, 85–86, 185–207. https://doi.org/10.3280/RIV2023-085010

Montrosse-Moorhead, B. (2023). Evaluation criteria for artificial intelligence. *New Directions for Evaluation*, 178–179, 123–134. https://doi.org/10.1002/ev.20566

Nielsen, S.B. (2023). Disrupting evaluation? Emerging technologies and their implications for the evaluation industry. *New Directions for Evaluation*, 178–179, 47–57.https://doi.org/10.1002/ev.20558

Nielsen, S.B. (2025). The evaluation industry and emerging technologies. In S. B. Nielsen, F. Mazzeo Rinaldi, & G. J. Petersson (Eds.), *Artificial intelligence and evaluation: Emerging technologies and their implications for evaluation* (pp. 266–286). Routledge. https://doi.org/10.4324/9781003512493

Nielsen, S.B., Mazzeo Rinaldi, F., Petersson, G.J. (Eds.). (2025). *Artificial intelligence and evaluation: Emerging technologies and their implications for evaluation*. Routledge. https://doi.org/10.4324/9781003512493

NIST - National Institute of Standards and Technology. (2022). AI test, evaluation, validation and verification (TEVV). https://www.nist.gov/ai-test-evaluation-validation-and-verification-tevv

Parasuraman, R., Manzey, D.H. (2010). Complacency and automation bias in the use of imperfect automation. *Human Factors*, 52(3), 381–410. https://doi.org/10.1177/0018720810376055

Paterson, B.L., Thorne, S.E., Canam, C., Jillings, C. (2001). *Meta-study of qualitative health research: A practical guide to meta-analysis and meta-synthesis*. Sage.

Patton, M.Q. (1999). Enhancing the quality and credibility of qualitative analysis. *Health Services Research*, 34(5 Pt 2), 1189–1208.

Patton, M.Q. (2008). *Utilization-focused evaluation* (4th ed.). Sage Publications.

Pawson, R., Tilley, N. (1997). *Realistic evaluation*. Sage Publications.

Pearl, J., Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.

Petersson, G.J., Breul, J. (Eds.). (2017). *Cyber society, big data, and evaluation*. Routledge.

Picciotto, R. (2020). Evaluation and the Big Data challenge. *American Journal of Evaluation,* 41(2), 166–181. https://doi.org/10.1177/1098214019850334

Potluka, O., Harten, S., Kocks, A., Dvorak, J. (2025). Digitalization in evaluations and evaluations of digitalization: The changing landscape of evaluations. *Evaluation*, 31(3), 289-302. https://doi.org/10.1177/13563890251357650

Pudney, S., Mills, D., Alaei, A.R., Sellers, S., Dvorak, J., Potluka, O. (2025). Evaluation of artificial intelligence-enhanced critical infrastructure systems: A conceptual framework. *Evaluation*, 31(3), 412–443. https://doi.org/10.1177/13563890251350677

Rompczyk, K. (2025). Technological revolution in evaluation: Artificial intelligence and the adherence to evaluation standards. *Evaluation*, 31. https://doi.org/10.1177/13563890251331066

Saunders, M.N.K., Rojon, C. (2011). On the attributes of a critical literature review. *Coaching: An International Journal of Theory, Research and Practice*, 4(2), 158–162. https://doi.org/10.1080/17521882.2011.596485

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., Hall, P. (2022). *Towards a standard for identifying and managing bias in artificial intelligence* (NIST Special Publication 1270). National Institute of Standards and Technology.

Shapiro, S., Lam, V. (2025). Artificial intelligence in program evaluation: Insights and applications. *Canadian Journal of Program Evaluation*, 39(2), 382–391. https://doi.org/10.3138/cjpe-2024-0027

Sharma, V.D. (2025). Artificial intelligence as a research tool in social sciences and emerging challenges. *International Journal for Multidisciplinary Research*, 7(3). https://doi.org/10.36948/ijfmr.2025.v07i03.49410

Wencker, T., Borst-Graetz, J., Niekler, A. (2025). Text as data for evaluation: Natural language processing and large language models to generate novel insights from unstructured text data. *Evaluation*, 31(3), 369–393. https://doi.org/10.1177/13563890251330911

Whittemore, R., Vilar-Compte, M., De La Cerda, S., Delvy, R., Jaser, S., Stowell, C., Breen, C., Gobeil-Lavoie, A.P., Naranjo, D., Crimston, C., Weinger, K., Pérez-Escamilla, R. (2023). Getting rigor right: A framework for methodological choice in adaptive monitoring and evaluation. *Global Health: Science and Practice*, 11(Suppl. 2). https://doi.org/10.9745/GHSP-D-22-00243

York, P., Bamberger, M. (2020). *Measuring results and impact in the age of big data: The nexus of evaluation, analytics, and digital technology*. Rockefeller Foundation.

York, P., Bamberger, M. (2025). The applications of Big Data to strengthen evaluation. In S.B. Nielsen, F. Mazzeo Rinaldi, G.J. Petersson (Eds.), *Artificial intelligence and evaluation: Emerging technologies and their implications for evaluation* (pp. 37–55). Routledge.