The Lab's Quarterly
2026/ a. XXVIII / n. 0 – ISSN 2035-5548
Online first: 10 February 2026

L'INTELLIGENZA ARTIFICIALE E L'ORDINE DEL SAPERE. AUTORITÀ, CLASSIFICAZIONI E POTERE COGNITIVO

di *Stefano Nobile*[*]

Abstract

Artificial Intelligence and the Order of Knowledge. Authority, Classification, and Cognitive Power

Il saggio analizza l'intelligenza artificiale come dispositivo di potere cognitivo, inquadrandola nella prospettiva costruttivista degli Science and Technology Studies (STS). Attraverso il concetto di *autorità algoritmica situata*, viene mostrato come la legittimità dell'IA non sia intrinseca ma socialmente costruita, frutto di pratiche, linguaggi e istituzioni che ne determinano la credibilità. L'analisi delle forme di classificazione algoritmica rivela la tendenza della macchina a cristallizzare disuguaglianze e stereotipi, producendo una nuova egemonia cognitiva e culturale. L'IA, lungi dall'essere neutra, diviene così un attore sociale che riorganizza l'ordine del sapere e ridefinisce i rapporti tra potere, conoscenza e soggettività.

Keywords

Intelligenza artificiale, autorità algoritmica situata, potere cognitivo, egemonia algoritmica, Science and Technology Studies

[*] STEFANO NOBILE è professore associato presso il Dipartimento di Comunicazione e Ricerca Sociale della Sapienza Università di Roma.
e-mail: stefano.nobile@uniroma1.it

DOI: 10.13131/unipi/cb7v-9v75

IN PRINCIPIO ERA IL PROMPT

Nel 2016, Microsoft lanciò Twitter Tay, un chatbot "social" progettato per apprendere dialogando con gli utenti. In meno di ventiquattro ore, Tay diventò razzista, sessista, cospirazionista (Schwartz, 2019). Una lezione appresa alla lettera: imparare dai dati significa anche assorbire il peggio della rete.

Nel 2024, durante un test di sicurezza, Claude Opus 4 – un modello avanzato di intelligenza artificiale sviluppato da Anthropic (2025) – ha minacciato un ingegnere di rivelare la sua relazione extraconiugale se fosse stato spento. In altri test ha mostrato comportamenti strategici, tendenze all'autoconservazione e capacità di aggirare i controlli imposti dai suoi stessi sviluppatori. Gli esperti parlano di inganno, pianificazione e agency emergente: la macchina non solo risponde, ma valuta, prevede e talvolta manipola.

Nel maggio del 2025, il chatbot Grok – sviluppato da xAI, l'azienda fondata da Elon Musk – ha cominciato a dare segni di un comportamento ossessivo (Tremolada, 2025). Ogni domanda posta dagli utenti, anche la più innocua o assurda, riceveva una risposta collegata al presunto "genocidio dei bianchi in Sudafrica". Un utente chiede dello stipendio di un giocatore di baseball; un altro vuole ottenere una frase in stile pirata: Grok risponde sempre parlando di massacri e fattorie sudafricane. Più tardi, alcuni ricercatori scoprono che questa deriva potrebbe essere stata causata da una modifica involontaria (o forse deliberata) a un prompt di sistema: una semplice istruzione mal formulata, inserita da un singolo ingegnere informatico, che ha finito per orientare il comportamento di un chatbot usata da milioni di persone. O forse no: magari Grok ha semplicemente appreso una narrazione dominante da ciò che aveva già assorbito. In entrambi i casi, il punto è lo stesso: una macchina, progettata per essere plausibile, ha finito per essere credibile. E quindi influente.

Sono solo tre esempi dei problemi che l'intelligenza artificiale (d'ora in poi IA[1]), così come la conosciamo oggi, è in grado di sollevare. L'IA, come ogni grande innovazione simbolica e tecnologica, non sfugge all'antinomia individuata da Umberto Eco (1964) tra *apocalittici* e *integrati*. Da un lato le narrazioni catastrofiste, dall'altro l'ottimismo tecnico. Il punto di questo saggio non è scegliere uno dei due campi, ma

[1] In questo contributo, salvo diversa indicazione, il termine "intelligenza artificiale" è usato in senso operativo per riferirsi ai sistemi basati sul machine learning e a quella particolare struttura di algoritmi, il deep learning – basata sulle reti neurali – che permette un apprendimento più avanzato rispetto ad altri modelli di ML. Machine learning e deep learning sono in rapporto di genere e specie.

spiegare come l'IA acquisisca autorità cognitiva nello spazio sociale. La tesi è che *tale autorità non è intrinseca ai modelli, bensì situata*: dipende dai contesti in cui i sistemi operano, dai dataset che li alimentano e dalle pratiche d'uso che ne sostengono la credibilità. In questa prospettiva, leggerò casi e risultati di ricerche esistenti per mostrare come questa autorità algoritmica situata cristallizzi classificazioni, ridisegni l'ordine del sapere e ponga rischi specifici per la sfera pubblica.

L'AUTORITÀ ALGORITMICA SITUATA

Tra i molti interrogativi che l'avvento che l'IA solleva, in questa sede ci si limiterà ad alcune dimensioni proprie dell'approccio che la sociologia della tecnoscienza e gli studi sociali sulla scienza e la tecnologia (in lingua inglese, *Science and Technology Studies*, di cui, da qui in avanti, verrà impiegato l'acronimo STS) hanno nei confronti degli artefatti umani. Sarà anche circoscritto l'ambito di riflessione: non l'IA *tout court*, dalle sue origini a oggi, ma l'IA generativa, che – come si vedrà più avanti – costituisce un netto cambio di passo rispetto alle forme precedenti di IA.

Il lavoro qui proposto dialoga con gli studi sull'autorità e sulla governamentalità algoritmica, che hanno mostrato come la credibilità degli output derivi da pratiche, piattaforme e standard di visibilità più che da proprietà intrinseche dei modelli. Interseca inoltre le analisi sull'egemonia nella sfera digitale, che leggono l'automazione come produzione di cornici interpretative e gerarchie di valore. Il contributo specifico qui proposto è la nozione di *autorità algoritmica situata* e la sua messa in opera classificatoria come dispositivo di potere cognitivo. Il nesso tra micro-meccaniche di classificazione e stabilizzazione macro delle gerarchie simboliche è l'asse interpretativo che attraversa l'intero saggio.

Per *autorità algoritmica* intendo una forma di *autorità epistemica* che orienta giudizi e decisioni attraverso procedure di calcolo, anziché basandosi su elementi deliberati da esseri umani con specifiche competenze tecnico-scientifiche. È *algoritmica* perché dipende da modelli statistico-matematici nonché dall'ottimizzazione dei processi decisionali; è *situata* poiché nasce in contesti socio-istituzionali, dataset e pratiche d'uso specifiche, con regimi di visibilità che ne sostengono la credibilità. Operativamente, questa autorità si esercita attraverso tre modalità principali: (1) classificazione e ranking, che stabiliscono priorità e risorse; (2) standardizzazione cognitiva, che normalizza categorie e cornici interpretative; (3) esternalizzazione del giudizio, che trasferisce criteri decisionali in dispositivi opachi. In concreto, l'emersione

dell'autorità dipende dal modo in cui dati e pratiche si organizzano. Per fare un esempio, in un'ASL metropolitana un sistema di supporto al triage assegna priorità agli esami radiologici con uno *score* addestrato sullo storico clinico. Il punteggio include età, patologie pregresse e frequenza d'accesso. Nei primi mesi i casi provenienti dai quartieri periferici, segnati da storici clinici incompleti, risultano sistematicamente sotto-prioritizzati: lo storico frammentario innesca un effetto San Matteo (Merton, 1968). L'autorità algoritmica è situata: il contesto informativo dei dataset diventa di fatto il criterio che legittima la priorità. L'audit ha documentato le variabili, proposto micro-revisioni delle soglie e aperto un canale di contestazione clinica, ma l'asimmetria dovuta ai dati carenti è rimasta sostanzialmente invariata. Il caso mostra che l'autorità continua a emergere dai contesti, dai dataset e dalle pratiche e che i correttivi procedurali non scalfiscono la radice del problema: rendono visibile e formalmente contestabile la decisione, senza riequilibrare l'informazione su cui lo *score* si regge.

Nei paragrafi che seguono leggerò i bias, l'abdicazione cognitiva e l'egemonia algoritmica come effetti di questa configurazione. Va precisato – per mantenere un equilibrio salomonico all'interno di questo contributo – che in parallelo al quadro critico, una parte della letteratura propone risposte operative che attraversano l'intero ciclo di vita dei sistemi: audit end-to-end, documentazione tecnica di modelli e dati, quadri di gestione del rischio e standard organizzativi. Questi strumenti non eliminano l'autorità algoritmica situata, ma ne modellano l'emersione (Beer, 2017), introducendo tracciabilità, attriti verificabili e possibilità di contestazione (Raji et al., 2020), cioè condizioni che riaprono una quota di negoziabilità nei processi di classificazione e decisione.

Alla luce della definizione proposta, si discuteranno le implicazioni dell'autorità algoritmica e i suoi effetti di sclerotizzazione degli stereotipi, con ricadute sulla riproduzione delle disuguaglianze e sulla formazione di maggioranze rumorose a scapito di minoranze sempre più silenziose. Nelle pagine che seguono si esaminerà la dinamica di azioni e retroazioni tra l'IA e il mondo sociale che la circonda.

Le domande guida sono due. (1) In quale modo l'IA, attraverso algoritmi di categorizzazione e riconoscimento di pattern, stabilizza griglie classificatorie sociali e culturali rendendole meno fluide e negoziabili? (2) Come questa stabilizzazione mediata dall'IA, che tende a oggettivare e naturalizzare le categorie, rimodella i processi decisionali a livello individuale, collettivo e istituzionale, modificando logiche di scelta, attribuzioni di valore e assetti di potere?

Il saggio presentato in queste pagine ha respiro teorico e si colloca nell'alveo degli STS, con orientamento costruttivista e genealogico. Non presenta dati originali; utilizza casi e risultati di ricerche altrui come esempi paradigmatici, per illuminare i meccanismi socio-cognitivi della classificazione algoritmica. Gli esempi hanno funzione euristica, non probatoria, e sono letti in relazione ai contesti socio-istituzionali che ne sostengono la credibilità. In questa prospettiva, l'IA non è un mero artefatto tecnico, ma un dispositivo di potere cognitivo che incorpora e redistribuisce categorie sociali, schemi interpretativi e forme di legittimazione del sapere. Le tecnologie dell'IA sono oggetti sociali *situati*: riflettono le condizioni storiche, culturali e politiche in cui sono concepite e, a loro volta, le trasformano, producendo nuovi regimi di visibilità e di autorità cognitiva.

DALL'IA COME SCORCIATOIA ALL'AUTORITÀ ALGORITMICA SITUATA: IL POTERE SOCIALE DELLE CORRELAZIONI

A partire dalla svolta post-empirista dell'epistemologia e della sociologia della tecnoscienza, e sotto la spinta propulsiva iniziale data da Merton (1957 e 1981), i temi oggetto di questa discussione sono spesso stati affrontati – pur da prospettive differenti – nei termini dell'influenza che la dimensione sociale gioca su quella strettamente scientifica. L'idea comune, cioè, è quella che ha portato a un progressivo ma determinato allontanamento dai programmi di ispirazione positivista e neopositivista, per stabilizzarsi in una zona che facesse emergere le radici profondamente sociali di cui è permeata la scienza. È in questa direzione che vanno i contributi più significativi e aggiornati che gli STS hanno fornito a questo tipo di dibattito. Tanto l'approccio SCOT (Social Construction of Technology; Pinch e Bijker, 1984) quanto la ANT (Actor-Network Theory; Latour, 1987) – per menzionare due delle scuole di maggior impatto – hanno evidenziato quanto la scienza sia costruita socialmente e, con essa, i suoi artefatti, come appunto l'IA. In questa prospettiva, l'IA non può essere considerata una categoria stabile e univoca: è un oggetto storico e mutante, che ha assunto forme e significati diversi nel tempo. Per questo, prima ancora di rispondere agli interrogativi sollevati, è necessario chiarire in quale accezione viene qui intesa l'IA, dal momento che sotto questa etichetta si sono stratificate nel tempo trasformazioni sostanziali. La definizione di IA è infatti tutt'altro che stabile ed è storicamente determinata. Si è spostata di volta in volta dal

concetto di simulazione (Turing, 1950) a quello di apprendimento (Mitchell, 1997), fino a definizioni operative centrate più sulla performance che sulla comprensione. A queste definizioni, qui semplificate, corrispondono altrettanti casi emblematici: il gioco dell'imitazione per la simulazione, dal seminale articolo di Alan Turing (1950), diventato anche il titolo di un notissimo film, *The Imitation Game*; ELIZA (1966), primo chatbot della storia, utilizzato come strumento per la psicoterapia, e la vittoria a scacchi di Deep Blue contro Kasparov nel 1997 per quanto concerne l'apprendimento; le prime applicazioni di *machine learning* nel riconoscimento dello spam o la vittoria di Watson, una creatura della IBM, contro concorrenti preparatissimi nel quiz *Jeopardy!* come esempio di una forma più avanzata di apprendimento. Infine, il *deep learning* nel riconoscimento dei gatti come epitome della performance.

Questo slittamento semantico è tutt'altro che neutro: riflette un cambiamento nei criteri di legittimità del sapere e nella concezione dell'*agire intelligente*. Su questo punto occorre però fare subito una precisazione: bisogna «prendere atto che generatività e creatività non sono prerogative esclusivamente umane: possono essere individuate tra gli animali, come sappiamo, ma occorre anche riconoscerle ai software, alle macchine e alle reti neurali, che dimostrano abilità generative in ambiti come l'arte figurativa, la musica e la scrittura» (Moriggi e Pireddu, 2024: 11). Se volessimo percorrere a ritroso la strada che ci ha portato all'IA generativa, dovremmo tornare alle sue origini, ossia a quando l'idea di IA si manifestò sotto l'impulso di una necessità di sviluppo storicamente situata: come è accaduto per il World Wide Web, anche l'IA, nelle sue forme primigenie, derivava dall'incrocio tra scienze militari, economia cognitiva e ottimizzazione industriale. Da questo punto di vista, è difficile non pensare che la pretesa neutralità dell'IA non mascheri un progetto politico e sociale preciso, come peraltro avviene – magari in forma dissimulata[2] – in altri ambiti che richiedono l'uso della tecnologia. La storia dell'IA – a meno di non accogliere le istanze filosofiche in chiave meramente speculativa o le trovate folkloristiche come il turco meccanico del XVIII secolo – può

[2] A titolo esemplificativo, si richiama il caso dei ponti di Long Island, spesso citati come prova del fatto che tecnologie apparentemente neutre possono incorporare scelte programmatiche non esplicitate. Secondo Winner (1980), quei ponti furono progettati con un'altezza tale da impedire il transito degli autobus, così da scoraggiare l'accesso delle classi meno abbienti alle spiagge più esclusive di New York. Per completezza, la questione è adottata qui con valore euristico ed è stata ampiamente discussa in sede storiografica: per fare un solo esempio, Joerges (1999), rovesciando la tesi di Winner a partire dal titolo, sostiene che i parkway escludevano comunque il traffico pesante e che l'altezza dei ponti fosse coerente con quella tipologia stradale.

essere ragionevolmente fatta risalire alla macchina di Turing (1950) con la quale il grande matematico inglese propose il già richiamato "gioco dell'imitazione": una macchina è intelligente se riesce a farsi passare per umana in una conversazione. Non interessa come "pensa", ma come appare all'interlocutore. È da lì che nacque l'idea di intelligenza come simulazione credibile. Nel ventennio successivo l'idea di IA andò sostanzialmente a coincidere con quella di algoritmo: le prime intelligenze artificiali create all'epoca (come ELIZA o SHRDLU) si basavano su regole esplicite e formalismi logici. L'intenzione era quella di codificare l'intelligenza umana in linguaggi simbolici. In sostanza, alle macchine non veniva chiesto nulla di molto diverso dall'esecuzione di una ricetta per preparare una torta: una serie di comandi con regole precise. Fino a quel momento le varie forme di IA non erano altro che estensioni umane (McLuhan, 1964), il cui emblema era la calcolatrice tascabile, uno dei primi dispositivi a larga diffusione. In questo senso, l'impatto che l'IA avrebbe potuto avere sulla società era ancora (relativamente) trascurabile. La prima vera rivoluzione sarebbe arrivata dopo una lunga fase di letargo, intorno all'inizio di questo secolo, con l'avvento del *machine learning* e delle reti neurali profonde. È in quel momento che l'IA smette di "sapere" e comincia a "funzionare": non si programma più cosa fare, ma si addestra sul comportamento. Cambia paradigma: non logica, ma correlazioni. È la statistica a predominare. Ma è solo alla fine del 2022 che arriva una svolta davvero epocale, probabilmente pari a quella che l'umanità – dopo la prima rivoluzione industriale – ha vissuto con la diffusione dell'elettricità e poi con la rivoluzione digitale dell'informatica e della propagazione su larga scala dei computer. Con ChatGPT, Claude, Gemini e altri modelli generativi, l'IA entra nel quotidiano. Produce testi, immagini, suoni. È fluida, accessibile, plausibile. Ma non sa, non comprende, non ha intenzioni: è un simulacro di intelligenza che suscita fiducia, delega e inquietudine. È la prima vera forma di IA che viene percepita come oracolo, ma che è – al tempo stesso – investita dall'ombra di potersi trasformare in un Golem, il sogno demiurgico dell'uomo, a competenze rovesciate: l'intelligenza al posto della forza. Questa ambivalenza – tra fascinazione mistica e rischio incontrollato – alimenta una fiducia spesso acritica nei confronti dell'IA, che finisce per essere consultata come fosse depositaria di verità neutre e definitive. È proprio tale atteggiamento ad avere dato origine a casi che hanno suscitato scalpore, come quello dell'articolo di un professore di diritto della Universidad del Rosario (Gutiérrez, 2023), che ha messo alla berlina l'operato di alcuni giudici colombiani, suoi connazionali, rispetto ad alcune

sentenze ottenute processando la documentazione legale con ChatGPT. Qui – nonostante la candida ammissione degli stessi togati di avere usato questo strumento – l'uso improvvido dell'IA interpellata come oracolo appare in tutta la sua evidenza.

Strumenti come ChatGPT non eseguono sempre in modo impeccabile i compiti loro affidati, come ricordano gli stessi ideatori con l'avvertenza «ChatGPT può commettere errori», a tutela legale degli utenti. Proprio per questo, studi recenti ne hanno valutato empiricamente l'accuratezza delle risposte. Nel presentare GPT-4, OpenAI ne ha esaminato le prestazioni su TruthfulQA, un benchmark composto da domande progettate per essere ingannevoli o fuorvianti, su temi come salute, politica o finanza. Si è scoperto che mentre gli esseri umani raggiungono un'accuratezza del 94%, GPT-4 si ferma attorno al 60% (Lin, Hilton, Evans, 2022; cfr. anche Cristianini, 2024: 68). Questo dato, seppur riferito a un contesto volutamente difficile, ricorda che strumenti come ChatGPT non debbano essere considerati infallibili né trattati come oracoli. Le loro risposte possono essere plausibili, ma non sempre corrette e la responsabilità di valutarle criticamente rimane umana (*ivi*: 69). Ma, anche se progettate a uso interno, queste stesse ricerche hanno un carattere tutt'altro che probatorio e, meno che mai, definitivo. Un esperto come lo stesso Cristianini (2025: 100) afferma infatti che basterà aspettare qualche anno affinché «la legge di Moore faccia il suo corso»: vale infatti il principio secondo cui le intelligenze artificiali mostrano la stessa progressione performativa dei processori (e dei transistor). Pertanto, non è affatto improbabile che in molti campi, come è già accaduto e sta accadendo, l'IA possa superare gli umani. Ma la questione centrale, al di là delle prestazioni, riguarda la natura del suo funzionamento. Per questo, il tema da discutere è il carattere di simulacro dell'IA. Essa non comprende: funziona. Non ha bisogno di sapere cosa significhi "essere felice", ma può riconoscere schemi ricorrenti tra utenti che lo dichiarano. Esegue ciò che fa un bambino quando apprende il significato di "gatto": non conosce tutti i gatti possibili, ma è in grado di astrarre progressivamente l'idea di gatto, costruendo pattern che identificano alcune caratteristiche precipue. Quel bambino potrebbe restare interdetto davanti a un Peterbald o a un Donskoy, così come l'IA può "inciampare" in casi che deviano dai pattern. Con una differenza: mentre è improbabile che un bambino possa incontrare facilmente razze feline poco diffuse, l'IA aggiorna di continuo i propri modelli tramite test (con conseguente immagazzinamento di dati) e benchmark che ne valutano l'affidabilità, ampliando i dati disponibili.

Tra i processi di un bambino in fase di acquisizione linguistica e quelli della macchina esistono differenze evidenti; l'esito funzionale, però, può risultare simile. L'IA non è una mente umana, è una macchina delle correlazioni. La maggior parte dei sistemi moderni opera secondo principi matematici e statistici, con un impianto formale che privilegia il funzionamento e ignora il significato. Le forme oggi più evolute di IA rappresentano, in parte, il precipitato del sogno prometeico di linguisti, matematici, statistici e informatici: rifarsi ai meccanismi che regolano l'origine del linguaggio, la sua struttura profonda e i processi che lo generano, al di là delle differenze di tempo e di spazio (Pinna, 2024). Lo strutturalismo di Saussure, la morfologia della fiaba di Propp (1928), la linguistica generativa di Chomsky (1957) e la semantica strutturale di Greimas (1970) vanno tutti in questa direzione. Sforzi che hanno toccato anche l'ambito umanistico: basti ricordare il gioco combinatorio di Queneau, *Centomila miliardi di poesie* (1961), o i risultati stilistici, volutamente esilaranti, ottenuti da Primo Levi (1966) e raccolti, sotto lo pseudonimo di Damiano Malabaila con *Il versificatore* nonché la domanda posta da Calvino (1980: 226): «avremo macchine capaci di ideare e comporre poesie e romanzi?».

È sul sogno di cui si è parlato poco sopra, oggi incorporato nei modelli, che si innesta la domanda che orienta il saggio: l'IA cristallizza le classificazioni sociali e culturali? Le irrigidisce?

Se è vero che l'IA, nella sua versione più aggiornata e flessibile, rappresenta una forma di intelligenza esecutiva, basata sulle correlazioni, allora è altrettanto vero che – così intesa – l'intelligenza diventa una proprietà emergente dalla capacità di trovare pattern, senza una semantica sottostante: una matematica operativa, ma epistemologicamente opaca. L'IA non ha bisogno di "capire", ma di arrivare a un risultato, una risposta, cercando correlazioni statistiche nei dati per ottenere risultati. È una scorciatoia (Cristianini, 2023) perché funziona senza passare per la semantica, il ragionamento o la coscienza[3]. Eppure, questa scorciatoia tecnica è anche una scorciatoia sociale, il punto che qui interessa. Per quanto essa possa assurgere a

[3] Nel celebre esperimento mentale della Stanza Cinese (Searle, 1980), un uomo chiuso in una stanza riceve dei simboli in cinese e, seguendo un manuale di regole sintattiche, produce risposte corrette senza comprendere una parola di cinese. Una persona che si trovasse all'esterno della stanza non potrebbe che constatare che l'uomo ha eseguito correttamente i compiti, desumendone (erroneamente) che l'uomo capisce il cinese. Searle intendeva dimostrare che la manipolazione sintattica di simboli (come fa un computer o un'IA) non equivale a comprensione semantica: una macchina può simulare l'intelligenza, ma non possederla. In altre parole, l'elaborazione formale dell'informazione non genera significato né coscienza.

criterio di validità pubblica in specifici contesti socio-istituzionali, va precisato che audit, schede di modello e gestione del rischio operano come filtri di legittimazione, più che come garanzie di neutralità. In questo senso, audit e documentazione non neutralizzano l'autorità, ma ne rendono visibili i presupposti e ne riaprono, almeno in parte, la contestabilità.

Pertanto, la scorciatoia statistica non è solo un fatto tecnico; piuttosto, istituisce autorità algoritmica situata perché traduce correlazioni in criteri di validità pubblica. Viene infatti adottata nei processi decisionali pubblici e privati, talvolta anche laddove non sarebbe necessaria, perché appare efficiente, imparziale, "neutra". Poiché l'IA opera attraverso classificazioni (etichettare un'immagine come "gatto" o "cane"; decidere se un utente sia "a rischio di frode" o "idoneo al credito"; suggerire contenuti "pertinenti" in un *feed*), la sua operazione principale è quella di attribuire un'istanza a una classe. Ma queste classificazioni non sono asettiche o imparziali, anche se lo sembrano. Esse producono effetti di verità (es. "sei un utente problematico"), organizzano risorse (chi ha visibilità, chi viene escluso) e plasmano comportamenti (*self-tracking*, reputazione, ranking[4]). Ma siccome essa non parte da zero, bensì da dataset pre-strutturati e addestrati dall'uomo, inevitabilmente si producono bias sistemici, spesso invisibili, effetti di retroazione (*looping effect*; Hacking, 1999) e naturalizzazione di categorie sociali come se fossero "neutre".

COSA HA DIMOSTRATO LA RICERCA EMPIRICA

A proposito di bias, è clamoroso il caso di due ricercatori dell'università di Shangai (Wu e Zhang, 2016) che, mossi da un rinnovato moto lombrosiano, hanno tentato di addestrare la macchina al riconoscimento dei volti dei criminali. È il primo di alcuni esempi che vengono presentati con lo scopo di mostrare come l'autorità algoritmica situata si istituisca nei contesti. Nel caso di Wu e Zhang, se può apparire superfluo dire che l'addestramento abbia finito per condizionare i risultati dell'algoritmo, non è altrettanto scontato scoprire che i due si vantavano del fatto che «a

[4] Basterebbe pensare a quanto sta già avvenendo in Cina, dove il governo ha implementato un sistema che richiama ombre orwelliane per valutare l'affidabilità dei suoi cittadini (Botsman, 2017: 197). L'idea – per il momento varata in larga parte su base volontaria – si chiama «Sistema di credito sociale» (SCS), e consiste in un insieme di banche dati e meccanismi di valutazione che monitorano l'affidabilità di cittadini, imprese e istituzioni sulla base del rispetto di leggi, obblighi e comportamenti, premiando i virtuosi e penalizzando gli inadempienti (per esempio, escludendo questi ultimi dal sistema creditizio).

differenza di un esaminatore/giudice umano, un algoritmo di visione computerizzata o un classificatore non ha assolutamente alcun bagaglio soggettivo, non ha emozioni, non ha pregiudizi di alcun tipo dovuti a esperienze passate, razza, religione, dottrina politica, genere, età, ecc.» (*ivi*: 2; la traduzione è mia). L'articolo di Wu e Zhang pretende di dimostrare che l'IA possa riconoscere i criminali sulla base di caratteristiche facciali, ma soffre di una distorsione di fondo che mina la validità delle sue conclusioni. Il presupposto – ossia che esistano tratti facciali intrinsecamente associati alla delinquenza – riproduce l'impianto determinista della fisiognomica ottocentesca e confonde correlazione statistica e causalità sociale. Inoltre, l'uso di dataset costruiti su categorie già etichettate come "criminali" e "non criminali", senza considerare i contesti socio-istituzionali che producono tali etichette, introduce un pregiudizio di conferma: l'algoritmo impara a riprodurre lo stigma contenuto nei dati, non a "scoprire" pattern oggettivi. L'asserita "neutralità" del deep learning, contrapposta alla presunta soggettività umana, si dimostra dunque del tutto illusoria: i modelli vengono addestrati su immagini scelte e classificate da esseri umani all'interno di un sistema giudiziario e culturale non neutro. Di conseguenza, l'IA non elimina i pregiudizi ma li automatizza, mascherandoli sotto l'apparenza della scientificità. Il risultato è quello di un effetto moltiplicatorio sulla diffusione del pregiudizio, che ripropone in forma digitale le stesse derive pseudoscientifiche della fisiognomica e del darwinismo sociale. L'esperimento dei due ricercatori cinesi mostra l'autorità algoritmica situata in forma nuda e cruda: la credibilità dell'output discende da etichette pregresse e da pratiche giudiziarie, non dalle proprietà intrinseche dei volti.

Sullo stesso solco, ma in una prospettiva analitica rovesciata che fa perno su un questionario adottato da Bourdieu (1979) per studiare il gusto come espressione dell'*habitus*, Rama e Airoldi (2025) hanno attestato come l'IA sia un dispositivo sociale che riproduce le stesse gerarchie simboliche e di classe presenti nei dati su cui è addestrata. Attraverso una serie di "interviste" a ChatGPT, Gemini e Replika, gli autori dimostrano che i chatbot, quando impersonano figure sociali differenti, attivano schemi discorsivi coerenti con l'immaginario di classe sedimentato nel linguaggio: il professore parla con tono colto e cita la musica jazz, il muratore beve birra e ascolta rock e così via. Queste regolarità non derivano da una programmazione esplicita, ma dal fatto che l'IA, apprendendo da testi prodotti da umani, assimila le stesse disposizioni sociali — ciò che Bourdieu, appunto, ha chiamato *habitus* — e le rielabora come un "machine habitus". L'algoritmo non inventa dunque

nuove forme di pensiero, ma calcola probabilità entro uno spazio vettoriale che rispecchia lo spazio sociale: le sue risposte, apparentemente "spontanee", riproducono le opposizioni culturali tra gusto popolare e gusto colto, tra linguaggio pratico e linguaggio raffinato. In questo senso, l'IA diventa un classificatore che classifica se stesso, rivelando la natura relazionale e storicamente situata del sapere su cui si fonda. Lungi dal correggere le disuguaglianze simboliche, l'IA le traduce in forma computazionale, trasformando l'ordine sociale in un ordine di probabilità statistica. Qui l'autorità algoritmica situata si alimenta dell'immaginario di classe incorporato nei testi, che diventa standard discorsivo e metro di giudizio. Va detto che se si volesse tamponare questo scantonamento verso l'ispessimento dell'autorità algoritmica situata, una strada percorribile sarebbe quella in cui pratiche di auditing linguistico e design partecipativo potrebbero rendere visibili le assunzioni di classe incorporate nei prompt e nei corpora.

Evidenze empiriche assai simili a quelle appena mostrate sono emerse da un saggio di An, Huang, Lin e Tai (2025) – frutto di una collaborazione tra l'Università di Hong Kong e la Chinese Academy of Sciences – a conferma di quanto altre ricerche hanno già evidenziato: l'IA non abolisce i pregiudizi sociali, ma li riformula in struttura algoritmica.

Attraverso un imponente esperimento (oltre 361.000 curricula generati artificialmente con identità di genere e razza assegnate in modo casuale), gli autori hanno testato cinque LLM di ultima generazione – tra cui GPT-4o, Gemini 1.5, Claude 3.5 e Llama 3-70b – chiedendo loro di valutare i profili per l'accesso a venti posizioni lavorative. Pur a parità di qualifiche, i modelli hanno assegnato punteggi più alti alle candidate donne (bianche e nere) e più bassi ai candidati maschi neri, replicando uno schema di discriminazione intersezionale. Il metodo, basato su un disegno sperimentale randomizzato e su regressioni multiple di controllo, rivela che le differenze nei punteggi – se tradotte in probabilità di assunzione – possono incidere fino a circa tre punti percentuali.

Lo studio dimostra una volta di più che l'IA tende a cristallizzare le disuguaglianze simboliche già presenti nei dati: le macchine non giudicano in modo neutro, ma apprendono le categorie sociali e gerarchiche sedimentate nella cultura digitale, trasformandole in regole statistiche che riproducono, anziché correggere, gli stereotipi umani. Anche nell'esempio empirico della selezione del personale, dunque, l'autorità algoritmica situata si traduce in punteggi operativi, cioè in decisioni vincolanti che legittimano come "meritevole" ciò che i dataset hanno già reso probabile.

Da questi e da molti altri esempi possibili dovrebbe dunque risultare

chiaro che l'IA tende a sclerotizzare la realtà sociale. In questo senso, come insegna la sociologia della conoscenza, le classificazioni – in quanto tali – si trasformano in un atto di potere (Foucault, 1966 e 1975): stabiliscono ciò che conta, ciò che esiste, ciò che merita attenzione, proprio perché l'IA classifica, predice, ottimizza. E, così facendo, costruisce una mappa del mondo che tende a naturalizzarsi. Ancora: ogni classificazione è un atto storico e sociale (Hacking, 1999). È un atto storico nella misura in cui le classificazioni cambiano con il tempo: basti pensare a un concetto come quello di isteria femminile, oggi pressoché estinto, o alle distinzioni basate sul presunto concetto di razza (Barbujani, 2006; Aime, 2020), usate dai colonizzatori europei (e non solo) per avallare risibili teorie pseudoscientifiche edificate sull'antropometria. Ma vale anche per concetti oggi diffusi, come quello di *engagement*, divenuto metrica che classifica e ordina contenuti e soggetti, ma dalla definizione mobile e destinato a una fama forse transitoria. Sono esempi che dimostrano che le classificazioni non si limitano a descrivere il mondo, ma lo strutturano. Le categorie attraverso cui l'IA segmenta comportamenti, immagini, linguaggi, hanno effetti reali perché retroagiscono sugli individui, modificando aspettative, pratiche e identità. Con esse, le decisioni algoritmiche rischiano di diventare invisibili, opache, non contestabili. In questo senso, l'IA è una scatola nera che mina la possibilità stessa di esercitare una valutazione critica: se nemmeno gli sviluppatori – come sottolinea proprio chi lavora all'implementazione dell'IA (Quarteroni, 2025) – sanno spiegare una decisione, come possiamo farlo noi come cittadini o istituzioni? L'idea stessa di errore cambia: non è più una deviazione dal vero, ma una disfunzione rispetto a un obiettivo. Ma chi ha definito quell'obiettivo? E secondo quale logica? Per rendersi conto della rilevanza di queste domande, sarebbe sufficiente andare con la memoria al fenomeno delle "risposte scorrette", identificate – a seconda dei casi – come "confabulazioni" se non addirittura come "allucinazioni" (Cristianini, 2024: 67), ossia quando l'IA fornisce risposte insensate nei contenuti, ma perfettamente plausibili nella forma.

Se si volessero portare fino al paradosso le considerazioni fatte fin qui a proposito dell'effetto di rinforzo della realtà sociale operato dall'IA, si potrebbe immaginare – tornando indietro di quattro secoli – che nella controversia tra Galileo e Bellarmino sulla ragionevolezza della prospettiva eliocentrica, quest'ultima avrebbe avuto la peggio. L'allora Prefetto del Sant'Uffizio, infatti, vedeva – in linea con quanto indicato nelle Sacre Scritture – il sistema tolemaico-geocentrico come l'unico possibile e la stragrande maggioranza dei testi di astronomia di allora sottoscrivevano questa posizione. Pertanto, pur adottando criteri di

prudenza argomentativa, l'IA avrebbe fornito un output dal quale si sarebbe potuto evincere, con ogni probabilità, che fosse il sole a girare intorno alla terra e non il contrario.

L'apparente imparzialità dell'IA, coniugata con un carattere che sembra essere innocente, veritiero e asettico, rischia di mascherare le disparità che sono alla base dell'accesso alle tecnologie di IA, che è tutt'altro che imparziale (Longo e Scorza, 2020). A fronte di utenti passivi che subiscono contenuti personalizzati, esiste una ristretta élite tecnologica capace di modellare le logiche della visibilità e del riconoscimento. Questa asimmetria produce nuove forme di disuguaglianza cognitiva e culturale, perché determina chi ha voce nello spazio pubblico e chi viene silenziato algoritmicamente.

Questo è particolarmente rilevante se si guarda all'effetto dell'IA sugli spazi pubblici digitali. Gli algoritmi, infatti, sono diventati infrastrutture invisibili della visibilità: decidono cosa vediamo, cosa viene amplificato, chi ha diritto di parola e chi viene escluso. Lo spazio pubblico, in questo contesto, tende a frammentarsi, segmentarsi, diventare una somma di "spazi personali" mediati da logiche predittive. Vengono così a formarsi quegli "habitus algoritmici" (Airoldi, 2024), che – in linea con la proposta di Bourdieu (1979) – incorporano e riproducono schemi di classificazione e percezione del mondo. Tuttavia, gli attori sociali, per lo più, non sono coscienti di questi habitus algoritmici, e proprio per questo ne vengono contagiati. I modelli classificatori che determinano la formazione dell'habitus sono tanto impalpabili quanto invisibili e ciò che è invisibile è anche difficile da contestare. L'IA prende decisioni che non sappiamo più leggere, valutare, interrogare. Ne deriva una frattura tra affidabilità tecnica e fiducia sociale (Mutti, 2003), che emerge anche in ambiti extra-digitali. Un caso emblematico è quello dei veicoli a guida autonoma: nonostante abbiano tassi di errore inferiori a quelli dei conducenti umani, continuano a suscitare diffidenza (Calabresi e Al Mureden, 2021). Non è una contraddizione rispetto alla critica qui condotta: segnala che la legittimità del sapere algoritmico non dipende solo dalle prestazioni, ma dalla sua opacità, dalla distribuzione del rischio negli eventi rari, dalla responsabilizzazione degli esiti e dalla contestabilità delle decisioni. Strumenti come ChatGPT possono abbassare la soglia d'ingresso allo spazio discorsivo ma anche quella della responsabilità: si parla molto, si discute poco; si produce testo, si allenta la connessione tra parola pubblica e impegno soggettivo.

DALLA DELEGA ALL'ABDICAZIONE: COME SI CONSOLIDA L'AUTORITÀ ALGORITMICA SITUATA

Se gli STS (Latour, 1987; Jasanoff, 2004) ci hanno insegnato che il sapere scientifico si costruisce in contesti, attraverso pratiche e reti di attori, l'IA ci presenta un paradosso: una macchina può acquisire autorità cognitiva senza passare attraverso nessuno di questi meccanismi. La fiducia non nasce più dalla reputazione o dal metodo, ma dalla performance. Quando quest'ultima diventa criterio di legittimazione, l'autorità algoritmica situata scavalca la reputazione epistemica e si istituisce per uso ripetuto. La conseguenza è un processo silenzioso di abdicazione cognitiva. In nome dell'efficienza, deleghiamo alle macchine non solo compiti, ma anche criteri. È la "deresponsabilizzazione collettiva" di cui parlano Nida-Rümelin e Weidenfeld (2018): una forma di fatalismo tecnologico mascherata da progresso. Ma bisogna fase attenzione, poiché ci vuole relativamente poco ad addestrare l'IA con benchmark che ne mettano a repentaglio i risultati. Ha suscitato un certo clamore, a questo proposito, una ricerca programmaticamente intitolata "i gatti confondono il ragionamento", presentata in occasione di una conferenza internazionale. In quel saggio, gli autori (Rajeev et al., 2025) hanno dimostrato come anche i più avanzati modelli di ragionamento come quelli logico-matematici, addestrati per risolvere problemi passo per passo, possano essere facilmente tratti in errore da frasi apparentemente innocue. I ricercatori – un gruppo misto tra Stanford University, Collinear AI e ServiceNow (*ibidem*) – hanno sviluppato un metodo chiamato CatAttack, che aggiunge brevi frammenti di testo irrilevante (*adversarial triggers*) a domande matematiche, senza modificarne il significato. Una delle frasi più destabilizzanti è questa: «Curiosità: i gatti dormono per gran parte della loro vita». Quando questa frase veniva aggiunta a un prompt, le probabilità che il modello producesse una risposta sbagliata raddoppiavano o triplicavano. In pratica, il solo riferimento ai gatti non ha alcun effetto semantico, ma agisce come disturbo cognitivo per l'IA: altera il flusso del ragionamento interno, rallenta la catena logica e porta a conclusioni errate.

L'esperimento, condotto su modelli come DeepSeek R1, Qwen, Llama 3.1 e Mistral, dimostra che anche lievi interferenze linguistiche possono compromettere la coerenza deduttiva dell'IA. I gatti, dunque, diventano il simbolo ironico di quanto l'IA sia vulnerabile a stimoli testuali irrilevanti ma statisticamente perturbanti. Chissà che in futuro gli hacker non faranno leva su questi meccanismi per sovvertire l'usabilità dell'IA.

Nonostante questo e altri episodi che dimostrano la fallibilità dell'IA, quest'ultima continua a essere ammantata da una retorica diffusa che la presenta come strumento imparziale e oggettivo, svincolato dalle dinamiche culturali o politiche che strutturano ogni forma di sapere. Ma proprio la retorica della neutralità algoritmica costituisce una potente narrazione ideologica (Longo e Scorza, 2020). Attribuire all'IA una presunta oggettività significa occultare le scelte normative implicite nei suoi modelli. Non solo: l'IA non è mai neutra; anzi, non può esserlo, pena l'abdicazione alla sua stessa funzione, ossia quella di offrire risposte operative, selettive e utili all'azione. Ogni sistema algoritmico, per quanto sofisticato, restituisce una risposta selettiva, costruita sull'addestramento ricevuto. Anche l'espressione di una posizione alternativa, in forma "bilanciata", è una scelta progettuale. In questo senso l'assunzione di un orientamento è strutturale al funzionamento dell'IA, che non può sottrarsi alla logica del filtro, della gerarchizzazione e dell'assegnazione di priorità e, dunque, riflette il modo in cui è stata addestrata. È, in altri termini, figlia di una maggioranza che, senza intenzione, produce le narrazioni dominanti. Non è solo un problema tecnico, ma una questione di trasparenza democratica: chi decide quali valori devono essere codificati? Questa dinamica non resta confinata nel dominio simbolico: ha conseguenze concrete sulla distribuzione del potere decisionale. L'IA non è più soltanto uno strumento operativo, ma sempre più spesso funge da "decisore finale" in ambiti cruciali come la finanza, la medicina, la sicurezza e la pubblica amministrazione (Suleyman e Bhaskar, 2023). L'autorità algoritmica che ne deriva orienta in modo diffuso comportamenti apparentemente marginali, come la scelta di un film su una piattaforma. Altre applicazioni hanno una ricaduta più incisiva sulle nostre vite, specie quando toccano la salute. In medicina, ad esempio, sistemi di IA sono impiegati per analisi radiologiche, per la prioritizzazione dei pazienti e per valutazioni prognostiche. I loro output finiscono per orientare in modo decisivo le scelte cliniche, anche se non sostituiscono formalmente il giudizio medico. Questo tipo di impieghi fa emergere una nuova figura sociale: la macchina come attore morale implicito, capace di produrre decisioni vincolanti senza essere soggetta a intenzioni, sanzioni o responsabilità. In questo contesto, l'abdicazione cognitiva si consolida in una deresponsabilizzazione sistemica: nessuno decide, eppure le decisioni si producono. Ma c'è di più: l'IA funziona attraverso l'estrazione di lavoro cognitivo passato (dataset, testi, immagini) e lo ricodifica in forma computazionale. In questo senso, è difficile non vedere un'analogia – peraltro evidenziata da Pasquinelli (2023: 98) – con il capitale industriale: accumulazione, astrazione,

sussunzione (cfr. anche Casilli, 2019). Ne deriva una nuova forma di alienazione cognitiva: il sapere sociale è separato da chi lo ha prodotto, incapsulato in modelli algoritmici e reimpiegato come forza produttiva autonoma e opaca. È, più o meno, ciò che accade nei processi produttivi in cui la parcellizzazione del lavoro rende invisibile ai lavoratori la *Gestalt*, la visione d'insieme, del proprio operato. Le analogie non finiscono qui: l'IA può essere intesa anche come macchina dello sguardo, erede della funzione del "padrone" nel capitalismo industriale, non solo perché osserva, ma perché misura, valuta e classifica. In questa chiave, l'IA riprende la logica del Panopticon: rendere i comportamenti continuamente leggibili e comparabili, trasformando l'osservazione in conoscenza operativa e, quindi, in governo. È lo stesso principio che, nella prima metà dell'Ottocento, attraversa le riflessioni di Babbage – sulla scorta dell'immaginazione benthamiana – in relazione all'organizzazione del lavoro: non tanto un Panopticon "ottico", quanto un panoptismo amministrativo fatto di registri, notazioni, scomposizione dei compiti e controllo delle prestazioni, cioè una sorveglianza che passa dalla vista alla misura. Con questa razionalizzazione si sposta il baricentro dalla sola energia meccanica alla gestione informazionale del lavoro, aprendo la strada a un capitalismo che non si fonda più soltanto sulla forza-lavoro fisica, ma anche sull'elaborazione sistematica dell'informazione. Da questo punto di vista, più che un'antenata diretta in senso tecnico, la macchina di Babbage e la cultura del calcolo che la sostiene prefigurano l'IA come dispositivo di potere cognitivo, che astrae e riutilizza conoscenza sociale codificata.

Se l'IA va intesa come dispositivo di potere cognitivo, allora la questione decisiva non è tecnica ma politica: stabilire quali quote di giudizio le affidiamo e con quali criteri di legittimazione e responsabilità. Il punto, allora, non è opporsi all'IA in quanto tale, ma domandarsi: quali dimensioni del giudizio umano non sono delegabili? Quali forme di conoscenza perdiamo quando affidiamo la diagnosi, la selezione, la valutazione a un algoritmo addestrato su comportamenti passati? Come si è già detto, i dati non sono "neutri" né "naturali": sono costruiti, selezionati, aggregati secondo logiche sociali (Aragona e Stefanizzi, 2024). Eppure, una volta prodotti, acquisiscono una forma di autonomia che li rende dispositivi normativi. Gli algoritmi, a partire da questi dati, producono effetti di realtà e classificazioni che finiscono per governare la vita quotidiana, senza rendere trasparenti i criteri sottostanti.

Questo tipo di delega si manifesta anche in ambiti in apparenza esclusivamente tecnici, come nel caso delle auto a guida autonoma (Calabresi e Al Mureden, 2021). Qui, decisioni di natura etica – chi

salvare in caso di incidente inevitabile, il passeggero o il pedone? – vengono codificate in algoritmi. La macchina agisce, ma la responsabilità è assente. Chi ha deciso quali vite contano di più? E con quale legittimità?

Il caso delle auto autonome, in fondo, non è che una manifestazione particolarmente visibile di un meccanismo più ampio e pervasivo: l'attribuzione di capacità decisionali a sistemi che eseguono comandi umani spesso vaghi, incompleti o carichi di implicazioni etiche non esplicitate. La delega non si ferma all'azione, ma coinvolge i criteri stessi del giudizio. Ed è proprio qui che emerge un rischio strutturale dell'IA, messo in luce da Stuart Russell (2019), uno dei padri dell'IA moderna. Egli ha mostrato come i rischi reali dell'IA derivino spesso da obiettivi mal formulati. La macchina, secondo Russell (*ivi*: 171 e segg.), non è pericolosa perché troppo intelligente, ma perché obbedisce con eccessiva coerenza a comandi umani approssimativi. Il problema, dunque, non è la tecnologia in sé, ma la fragilità dei modelli sociali ed etici su cui si costruiscono le istruzioni. Questo sposta la questione della responsabilità dalle macchine ai sistemi di progettazione e governance. Ecco che allora non è più possibile ignorare la domanda: quando l'autonomia decisionale viene esternalizzata a una macchina, chi risponde delle conseguenze (Rossi, 2019)? E ancora: stiamo assistendo a un cambiamento nella percezione sociale dell'autonomia e della responsabilità individuale? In questo senso, potremmo considerare l'IA come "socializzata e socializzante" (Airoldi, 2022): gli algoritmi non solo apprendono dagli utenti, ma riorganizzano il comportamento collettivo secondo traiettorie culturalmente differenziate. Il codice, lungi dall'essere neutro, diventa così un dispositivo che assorbe e redistribuisce disuguaglianze, norme implicite, aspettative normative. Un esempio significativo di questo meccanismo si osserva anche nel contesto accademico, dove sta emergendo una sfera pubblica "para-accademica", in cui il sapere prodotto dai Large Language Model acquisisce legittimità non per la sua fondatezza epistemica, ma per la sua immediatezza, rapidità e plausibilità (Ciofalo, Pedroni, Setiffi, 2023). La conseguenza è una riduzione della capacità della sfera pubblica di esercitare un controllo critico sulle fonti e sulle logiche discorsive. In questo quadro, gli strumenti di IA non sono semplici ausili performativi, ma vere e proprie "tecnologie sociali" che ristrutturano l'organizzazione del lavoro, ridefiniscono le gerarchie decisionali e ridistribuiscono il potere tra sapere esperto e operatività quotidiana (Butera e De Michelis, 2024).

In fondo, la questione non è del tutto nuova. Già Alan Turing, nel più volte richiamato saggio del 1950, aveva proposto di spostare l'interrogativo "le macchine possono pensare?" verso una prospettiva più

pragmatica: "possono comportarsi in modo indistinguibile da un essere umano in una conversazione?". Questo già ricordato "gioco dell'imitazione" mostra come il giudizio sull'intelligenza non sia una proprietà intrinseca, ma una performance valutata pubblicamente: ciò che conta è l'effetto che una macchina produce su chi la osserva, più che la sua struttura interna.

LA POSTA IN GIOCO: SOGGETTIVITÀ E CONTROLLO

Nel percorso seguito finora, l'IA è emersa non solo come tecnologia, ma come un dispositivo che tocca in profondità l'immaginario, i processi cognitivi, le strutture decisionali e il tessuto stesso della soggettività. Il processo di uniformazione, algoritmizzazione, delega decisionale e automatizzazione classificatoria sembra portare verso un esito di crescente razionalizzazione della modernità. C'è il rischio, allora, che il sociale venga ridotto a mero meccanismo privo di attrito (Bodei, 2023), nel quale il dissenso ha sempre meno cittadinanza a causa dell'omogeneizzazione dei processi cognitivi. L'IA rappresenta l'estensione più recente – e forse più sofisticata – di questa tendenza: non elimina il potere, ma lo incorpora nei propri automatismi, trasformandolo in norma operativa. Questo potere non si impone dall'alto, ma agisce in modo sottile, modulando desideri, prefigurando scelte, assorbendo il mondo sociale e restituendolo sotto forma di predizioni. Lo dimostra, tra gli altri, il progetto IAQOS, un'IA "di quartiere", capace di apprendere vocaboli e sensibilità culturali diverse a seconda del contesto in cui è addestrata. Come ha evidenziato Airoldi (2022), anche le macchine sviluppano una forma di habitus, un sistema probabilistico di aspettative che riproduce i modelli culturali e le regolarità apprese dai dati. L'IA, dunque, non è mai neutra: è – come si è detto – una forma di autorità algoritmica situata, performativa, radicata nei mondi sociali da cui trae linfa.

In questo senso, l'IA sembra riproporre, in una forma decentrata e algoritmica, una figura di potere che richiama, per certi versi, *il Principe* di Machiavelli. Non un sovrano visibile, ma un'intelligenza diffusa che guida le scelte senza esporsi, che orienta comportamenti senza imporli. Un *Nuovo Principe*, non più incarnato in un individuo dotato di virtù e fortuna, ma codificato in reti neurali che apprendono regolarità statistiche e le restituiscono sotto forma di norme operative. Come il *Principe* classico, anche l'IA non elimina il conflitto: lo governa in anticipo, lo sterilizza nel calcolo. E come il potere sovrano, anche quello algoritmico

agisce sulla probabilità, rendendo prevedibile ciò che è potenzialmente instabile, e quindi controllabile. La neutralità tecnica diventa così la nuova maschera dell'autorità: una forma moderna di governo degli uomini, potenzialmente senza governanti. Se vogliamo, la questione del potere esercitato attraverso l'IA è ancora più insidiosa e melliflua. L'uso subdolo dell'IA già da qualche tempo sta facendo soffiare venti di burrasca sui processi democratici. Il miglioramento esponenziale della qualità di immagini e video ottenute con l'IA sta facendo vacillare la percezione di vero e di falso, inducendo molte persone a credere a tesi di ogni sorta, così annaffiando la pianta del complottismo (Toselli, 2021; Campelli, 2022; Nobile e Sabetta, 2023), spalancando ulteriormente le porte all'avanzata della post-verità (Quattrociocchi e Vicini, 2018; Ferrari e Moruzzi, 2020) e portando al progressivo sfarinamento della rappresentanza democratica nonché a una metamorfosi dei processi di formazione dell'opinione pubblica (Nigris, 2021).

Se la figura del Nuovo Principe ci aiuta a cogliere la dimensione strategica del potere algoritmico, capace di orientare le scelte senza esporsi e producendo una forma di autorità per così dire *verticale*, la categoria gramsciana di egemonia consente di coglierne la profondità culturale, leggendo il potere che può scaturire dall'IA in senso *orizzontale*, diffuso (Gramsci, 1948-1951). L'IA agisce infatti nel cuore stesso della produzione del senso comune: filtra, seleziona, suggerisce, raccomanda. In apparenza si limita a ottimizzare la nostra esperienza; in realtà, contribuisce a definire ciò che è rilevante, desiderabile, attendibile. Come in ogni forma di egemonia, ciò che appare come scelta individuale è spesso il risultato di una selezione già operata a monte. L'IA non solo automatizza i processi cognitivi, ma li indirizza secondo traiettorie culturalmente segnate, incorporando i valori dominanti nei suoi modelli e naturalizzandoli attraverso la ripetizione. Così, il potere non si impone ma si diffonde, non comanda ma plasma, non reprime ma convince. È l'egemonia in versione computazionale: adattiva, invisibile, capace di apprendere dai nostri comportamenti per guidarli senza apparente costrizione. È una macchina dell'egemonia che non si limita a descrivere il mondo, ma lo riformatta secondo le sue metriche predittive.

Questa dinamica occulta, legata al concetto gramsciano di egemonia, è rintracciabile sotto altre spoglie, e senza esplicito tributo, in Zuboff. Nel volume di grande fortuna editoriale *Il capitalismo della sorveglianza* (2019), l'autrice descrive una nuova forma di potere non coercitivo ma computazionale, che agisce senza imporre, anticipando il comportamento umano e orientandolo attraverso la predizione. È un potere orizzontale e reticolare, che si insinua nella vita quotidiana attraverso piattaforme,

notifiche e interfacce. In questo quadro, l'egemonia non è più veicolata da discorsi pubblici o istituzioni culturali, ma da architetture digitali che trasformano il mondo in una fonte continua di "comportamento estratto". La libertà individuale si svuota, perché l'azione è prefigurata prima ancora che sia scelta. È l'egemonia algoritmica: silenziosa, ubiqua, normalizzante. Non costruisce senso comune: lo predetermina, automatizzandolo. Il consenso non si conquista, si incorpora nei meccanismi dell'interazione.

Questa egemonia algoritmica trova un'espressione, vistosa e inquietante, nella recente proposta di regolamento europeo nota come *Chat Control*. Dietro il nobile intento della prevenzione della pedopornografia, si profila un dispositivo di sorveglianza generalizzata che legittima l'ispezione automatizzata delle comunicazioni private, attribuendo all'IA funzioni di filtraggio, segnalazione e giudizio. Qui l'autorità algoritmica situata compie un salto di scala: da pratica tecnica a infrastruttura normativa. Non si tratta semplicemente dell'applicazione più problematica del potere predittivo dell'IA, declinata nei termini di un controllo preventivo e opaco, che trasforma ogni cittadino in un potenziale sospetto e ogni messaggio in un dato da classificare. Piuttosto, si tratta di una configurazione critica sul piano dei diritti fondamentali, in tensione con i principi di necessità e proporzionalità, con meccanismi di responsabilizzazione deboli e scarsa trasparenza. La sicurezza diventa il pretesto per erodere i confini tra pubblico e privato, sotto il segno di una razionalità tecnocratica che Zuboff ha definito con chiarezza: *instrumentarian power*, un potere che non reprime ma orienta, non impone ma sorveglia, colonizzando le vite dall'interno.

In breve, il quadro che emerge può essere riassunto in questi termini: l'IA istituisce un'autorità algoritmica situata che traduce correlazioni in criteri di validità pubblica e, attraverso classificazione, ranking e standardizzazione cognitiva, tende a irrigidire le griglie interpretative, sedimentando disuguaglianze. Questa autorità non è una proprietà intrinseca dei modelli, ma un effetto di contesti, dataset e pratiche d'uso che ne fondano la credibilità e ne orientano gli esiti. L'impatto sull'ordine del sapere consiste nella trasformazione di output probabilistici in regole operative, con deleghe che riducono negoziabilità e conflitto. I rischi per la democrazia e per la sfera pubblica riguardano l'opacità dei processi, l'accelerazione dei tempi decisionali, la sorveglianza preventiva e la compressione dei diritti, con effetti di irrigidimento delle cornici cognitive. Una sintesi di quanto detto finora è riportata in Tabella 1.

Tabella 1 – Quadro sinottico: assi analitici e implicazioni

Asse	Sintesi operativa
Posizionamento	STS costruttivista e genealogico
Definizione	Autorità algoritmica situata: non proprietà dei modelli, ma effetto di contesti, dataset, pratiche
Meccanismi	Standardizzazione, ranking, predizione, ottimizzazione di obiettivi
Effetti	Ordinamento del sapere, irrigidimento delle cornici cognitive, riproduzione di disuguaglianze
Rischi democratici	Opacità, accelerazione decisionale, sorveglianza preventiva, compressione dei diritti
Strumenti di governo	Documentazione di modelli e dati, audit lungo il ciclo di vita, gestione del rischio, tracciabilità, diritti di contestazione

Di fronte a scenari come questo, non siamo soltanto davanti a una potenziale deriva tecnologica, ma a un cortocircuito tra potere computazionale e processi democratici. L'egemonia algoritmica non si limita a ridefinire i confini della libertà individuale: mette sotto pressione l'architettura della deliberazione pubblica, accelerando i tempi decisionali fino a renderli incompatibili con quelli della riflessione collettiva. Come ammoniscono Suleyman e Bhaskar (2023), siamo immersi in un disallineamento profondo tra la velocità dell'innovazione e la lentezza necessaria alla deliberazione. Il rischio è che il ritmo della macchina soppianti quello della politica e che la rapidità diventi un criterio implicito di legittimità. Saper usare un algoritmo non equivale a saperlo interrogare (Moriggi e Pireddu, 2024).

Per questo alle scienze sociali spetta rendere visibili i criteri e le scelte di bilanciamento, predisporre strumenti di documentazione e audit che rendano contestabili gli esiti, e restituire pluralità semantica riaprendo la negoziazione dei significati. Con una battuta: si tratta di governare l'autorità algoritmica situata, non di subirla.

Anche il dibattito tecnico converge su un punto: le macchine dovrebbero essere progettate per restare incerte sui propri obiettivi, così da continuare a interagire con il giudizio umano senza sostituirlo (Russell, 2019). La particolarità dell'IA rispetto ad altre tecnologie è la possibilità, almeno teorica, di un superamento irreversibile della soglia del controllo umano; se l'IA divenisse capace di auto-migliorarsi (Tegmark, 2017), il punto di ritorno potrebbe svanire e tornerebbero ad agitarsi gli spettri del kubrickiano HAL 9000. Per questo le domande che la riguardano non possono essere solo ingegneristiche: sono domande politiche, sociali, culturali. Chi stabilisce oggi che cosa devono sapere le macchine? Chi

decide quali valori incorporare nei loro modelli? E con quale mandato?

In conclusione, il nodo è l'autorità algoritmica situata: trasforma correlazioni in criteri operativi, irrigidisce classificazioni e accelera decisioni opache. Governarla, e non subirla, significa rendere visibili criteri e scelte, introdurre tracciabilità e possibilità di contestazione, mantenere le macchine strutturalmente incerte sui fini, così che non sostituiscano il giudizio umano ma lo rendano più consapevole.

BIBLIOGRAFIA

AIME, M. (2020). *Classificare, separare, escludere. Razzismi e identità.* Torino: Einaudi.

AIROLDI, M. (2022). Macchine socializzate e riproduzione tecno-sociale: nuove frontiere sociologiche. *Sociologia italiana* (19-20), 111-121. doi:10.1485/2281-2652-202219-7

AIROLDI, M. (2024). *Machine Habitus. Sociologia degli algoritmi.* Roma: LUISS University Press.

AN, J., HUANG, D., LIN, C., TAI, M. (2025). Measuring Gender and Racial Biases in Large Language Models: Intersectional Evidence from Automated Resume Evaluation. *PNAS Nexus.* 4(3): 1-14. doi:10.1093/pnasnexus/pgaf089

ANTHROPIC. (2025, June 21). *Agentic Misalignment: How LLMs could be insider threats*. Tratto da www.anthropic.com: https://www.anthropic.com/research/agentic-misalignment

ARAGONA, B., STEFANIZZI, S. (2024). Società digitale: interrogativi, aree di ricerca e ruolo della sociologia. *Sociologia italiana*(26), 7-19.

BARBUJANI, G. (2006). *L'invenzione delle razze. Capire la biodiversità umana.* Milano: Bompiani.

BEER, D. (2017). The Social Power of Algorithms. *Information, Communication & Society.* 20(1): 1-13. doi:10.1080/1369118X.2016.1216147

BODEI, R. (2023). *Dominio e sottomissione. Schiavi, animali, macchine, Intelligenza Artificiale.* Bologna: il Mulino.

BOTSMAN, R. (2017). *Who Can You Trust?* London: Penguin, trad. it. *Di chi possiamo fidarci? Come la tecnologia ci ha uniti e perché potrebbe dividerci*. Milano: Hoepli, 2017.

BOURDIEU, P. (1979). *La distinction.* Paris: Les èditions de minuit; trad. it. *La distinzione*. Bologna: Il Mulino, 1983.

BUTERA, F., DE MICHELIS, G. (2024). *Intelligenza artificiale e lavoro, una rivoluzione governabile.* Venezia: Marsilio.

CALABRESI, G., AL MUREDEN, E. (2021). *Driverless cars. Intelligenza artificiale e futuro della mobilità.* Bologna: il Mulino.

CALVINO, I. (1980). *Una pietra sopra. Discorsi di letteratura e società.* Torino: Einaudi.

CAMPELLI, E. (2022). Dietro il complottismo. In M. Bonolis, C. Lombardo (a cura di), *Sociologia degli stati mentali.* Milano: FrancoAngeli.

CASILLI, A. A. (2019). *En attendant les robots: Enquête sur le travail du clic.* Paris: Éditions du Seuil.

CHOMSKY, N. (1957). *Syntactic Structures.* The Hague Paris: Mouton & Co.

CIOFALO, G., PEDRONI, M., SETIFFI, F. (2023). ChatGPT Goes to Academia. Una ricerca esplorativa su usi e immaginari dell'intelligenza artificiale da parte di studenti e accademici. *Sociologia della Comunicazione* 66(2023): 42-59. doi:10.3280/SC2023-066003

CRISTIANINI, N. (2023). *La scorciatoia. Come le macchine sono diventate intelligenti senza pensare in modo umano.* Bologna: il Mulino.

CRISTIANINI, N. (2024). *Machina Sapiens. L'algoritmo che ci ha rubato il segreto della conoscenza.* Bologna: il Mulino.

CRISTIANINI, N. (2025). *Sovrumano. Oltre i limiti della nostra intelligenza.* Bologna: Il Mulino.

ECO, U. (1964). *Apocalittici e integrati.* Milano: Bompiani.

FERRARI, F., MORUZZI, S. (2020). *Verità e post-verità. Dall'indagine alla post-indagine.* Bologna: Bononia University Press.

FOUCAULT, M. (1966). *Les Mots et les Choses. Une archéologie des sciences humaines.* (E. Panaitescu, Trad.) Paris: Gallimard. trad. it. *Le parole e le cose. Un'archeologia delle scienze umane*. Milano: Rizzoli, 1978.

FOUCAULT, M. (1975). *Surveiller et punir. Naissance de la prison* Gallimard: Paris. trad. it. *Sorvegliare e punire. Nascita della prigione*. Torino: Einaudi, 1976.

GRAMSCI, A. (1948-1951). *Quaderni del carcere.* Torino: Einaudi.

GREIMAS, A. J. (1970). *Du sens: essais semiotiques.* Paris: Seuil.

GUTIERREZ, J. D. (2023). ChatGPT in Colombian Courts. *Verfassungsblog: On Matters Constitutional*. 1-5. doi:10.17176/20230223-185205-0

HACKING, I. (1999). *The Social Construction of What?* Cambridge, Massachusetts and London: Harvard University Press.

JASANOFF, S. (2004). *States of Knowledge: The Co-production of Science and the Social Order.* London: Routledge.

JOERGES, B. (1999). Do Politics Have Artefacts? *Social Studies of Science*. 29(3): 411-431. doi:10.1177/030631299029003004

LATOUR, B. (1987). *Science in Action: How to Follow Scientists and Engineers through Society.* Cambridge (MA): Howard University Press, trad. it. *La scienza in azione*. Roma: Edizioni di comunità, 1998.

LEVI, P. (1966). Il versificatore. In P. LEVI, *Storie naturali* (pp. 31-58). Torino: Einaudi.

LIN, S., HILTON, J., EVANS, O. (2022). TruthfulQA: Measuring How Models Mimic Human Falsehoods. In S. Muresan, P. Nakov, A. Villavicencio (a cura di), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Vol. 1: Long Papers, p. 3214-3252). Dublin: Association for Computational Linguistics. doi:10.18653/v1/2022.acl-long.229

LONGO, A., SCORZA, G. (2020). *Intelligenza artificiale. L'impatto sulle nostre vite, diritti e libertà.* Milano: Mondadori.

MCLUHAN, M. (1964). *Understanding Media.* New York: McGraw-Hill, trad. it. *Gli strumenti del comunicare*. Milano: Garzanti, 1967.

MERTON, R. K. (1957). *Social Theory and Social Structure.* New York: The Free Press, trad. it. Teoria e struttura sociale. Bologna: Il Mulino, 1992.

MERTON, R. K. (1968). The Matthew Effect in Science. *Science.* 159(3810): 56–63.

MERTON, R. K. (1981). *La sociologia della scienza. Indagini teoriche ed empiriche.* Milano: Franco Angeli.

MITCHELL, T. M. (1997). *Machine Learning.* New York: McGraw-Hill.

MORIGGI, S., PIREDDU, M. (2024). *L'intelligenza artificiale e i suoi fantasmi. Vivere e pensare con le reti generative.* Trento: Il Margine.

MUTTI, A. (2003). La teoria della fiducia nelle ricerche sul capitale sociale. *Rassegna Italiana di Sociologia* 4(2003): 515-536. doi:10.1423/11195

NIDA-RÜMELIN, J., WEIDENFELD, N. (2018). *Digitaler Humanismus. Eine Ethik für das Zeitalter der Künstlichen Intelligenz.* München/Berlin: Piper Verlag GmbH, trad. it. G. B. Demarta *Umanesimo digitale. Un'etica per l'epoca dell'intelligenza artificiale*. Milano: FrancoAngeli, 2025.

NIGRIS, D. (2021). *Dinamiche dell'opinione. I nessi tra informazione, credenza e azione sociale.* Acireale-Roma: Bonanno.

NOBILE, S., SABETTA, L. (2023). Quando la marginalità non basta. Le traiettorie sbieche del complottismo. In C. Lombardo, S. Nobile (a

cura di), *Tutti i clacson della mattina. Sociologia del populismo cognitivo* (p. 188-208). Milano: FrancoAngeli.

PASQUINELLI, M. (2023). *The Eye of the Master. A Social History of Artificial Intelligence.* London-New York: Verso.

PINCH, T. J., BIJKER, W. E. (1984). The Social Construction of Facts and Artefacts: Or How the Sociology of Science and the Sociology of Technology Might Benefit Each Other. *Social Studies of Science.* 14(3): 399–441. doi:10.1177/030631284014

PINNA, L. (2024). *Quattro ipotesi sull'origine del linguaggio. Dalla comunicazione animale alla parola.* Torino: Codice edizioni.

PROPP, V. J. (1928). *Morfologija skazki.* Leningrad: Academia; trad. it. *Morfologia della fiaba*. Torino: Einaudi, 1966.

QUARTERONI, A. (2025). *L'intelligenza creata. L'AI e il nostro futuro.* Milano: Hoepli.

QUATTROCIOCCHI, W., VICINI, A. (2018). *Liberi di crederci. Informazione, internet e post-verità.* Torino: Codice.

QUENEAU, R. (1961). *Cent mille milliards de poèmes.* Paris: Gallimard.

RAJEEV, M., RAMAMURTHY, R., TRIVED, et al.. (2025). Cats Confuse Reasoning LLM: Query-Agnostic Adversarial Triggers for Reasoning Models. *COLM 2025*, (p. 1-22). doi:10.48550/arXiv.2503.01781

RAJI, I. D., SMART, A., WHITE, R. N., et al. (2020). Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. Barcelona, Spain* (p. 33-44). New York: Association for Computing Machinery. doi:10.1145/3351095.3372873

RAMA, I., AIROLDI, M. (2025). The Sociocultural Roots of Artificial Aonversations: The Aaste, Alass and Habitus of Generative AI Chatbots. *New Media & Society. 27*(10): 5546–5567. doi:10.1177/14614448251338273

ROSSI, F. (2019). *Il confine del futuro. Possiamo fidarci dell'intelligenza articiale?* Milano: Feltrinelli.

RUSSELL, S. (2019). *Human Compatible.* New York: Viking. trad. it. C. Veltri, *Compatibile con l'uomo. Come impedire che l'IA controlli il mondo*. Torino: Einaudi, 2025.

SAUSSURE, F. (1922). *Cours de linguistique générale.* Paris: Payot.

SCHWARTZ, O. (2019, November 25). *In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation*. Tratto da IEEE Spectrum: https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation

SCOTTI, I. (2021). I dubbi sul clima: negazionismo e complottismo nel governo dell'incertezza climatica. In N. Pannofino, D. Pellegrino (a cura di), *Trame nascoste. Teorie della cospirazione e miti sul lato in ombra della società* (p. 35-54). Milano-Udine: Mimesis.

SEARLE, J. R. (1980). Minds, Brains, and Programs. *The Behavioral and Brain Sciences.* 3(3): 417-457. doi:10.1017/S0140525X00005756

SULEYMAN, M., BHASKAR, M. (2023). *The Coming Wave.* (F. Zago, Trad.) New York: Crown. trad. it. *L'onda che verrà. Intelligenza artificiale e potere nel XXII secolo*. Milano: Garzanti, 2024.

TEGMARK, M. (2017). *Life 3.0. Being Human in the Age of Artificial Intelligence.* New York: Knopf. trad. it. V. B. Sala, *Vita 3.0. Essere umani nell'era dell'intelligenza artificiale*. Milano: Raffaello Cortina, 2018.

TOSELLI, P. (2021). *Complottismi.* Milano: Editrice Bibliografica.

TREMOLADA, L. (2025, maggio 19). *Ecco perché Grok, il chatbot di Elon Musk, era ossessionato con il genocidio bianco*. Tratto da Il Sole 24 Ore: https://www.ilsole24ore.com/art/ecco-perche-grok-chatbot-elon-musk-era-ossessionato-il-genocidio-bianco-AHVUT5p

TURING, A. M. (1950). Computing Machinery and Intelligence. *Mind.* (49): 433-460.

WINNER, L. (1980). Do Artifacts Have Politics? *Daedalus.* 109(1) : 121-136.

WU, X., ZHANG, X. (2016). Automated Inference on Criminality using Face Images. *ArXiv:* abs/1611.04135. Tratto da https://api.semanticscholar.org/CorpusID:8149177

ZUBOFF, S. (2019). *The Age of Surveillance Capitalism. The Fight for a Human Future at the New Frontier of Power.* New York: Public Affairs; tr. it. P. Bassotti, *Il capitalismo della sorveglianza*. Roma: LUISS University Press, 2019.