I quaderni di

Agenda Digitale.eu

Gennaio-Aprile 2026

n.22

Agendadigitale.eu è una testata scientifica e giornalistica registrata al Tribunale di Milano

Dati di riferimento

Iscrizione ROC n. 16446
ISSN 2421-4167
Numero registrazione 1927, Tribunale di Milano Editore: Digital360

Focus e ambito

La rivista scientifica, i Quaderni di Agendadigitale.eu, pubblica fascicoli quadrimestrali in open access.

Lo scopo è creare un luogo per accompagnare i passi dell'Italia verso la necessaria rivoluzione digitale, con approfondimenti multidisciplinari a firma di esperti delle materie afferenti all'Agenda Digitale italiana ed europea

Submission e norme editoriali

Per effettuare una submission è necessario concordare prima un argomento e le misure precise contattando info@agendadigitale.eu.

Inviare un abstract di circa 500 caratteri alla testata, presentando l'articolo.
Le misure del testo finale saranno comprese tra 6mila e 20mila caratteri, salvo accordi per misure superiori.

I riferimenti bibliografici dovranno essere preparati in conformità alle regole dell'APA style, 6a edizione (si vedano le linee guida e il tutorial).
Gli autori sono invitati a tener conto degli articoli già pubblicati nella rivista e di citarli nel loro contributo qualora siano ritenuti di interesse per il tema trattato.

Direzione e comitato editoriale

Direttore responsabile

Alessandro Longo, Direttore responsabile Agendadigitale.eu

Direttori scientifici

Paolo Ferri, Professore Ordinario di Tecnologie della formazione, Università degli Studi Milano-Bicocca
Mario Morcellini, Professore ordinario emerito in Sociologia della Comunicazione e dei Media digitali, Sapienza Università di Roma
Carlo Colapietro, Professore Ordinario Diritto costituzionale e pubblico, Università Roma Tre
Antonio Iannuzzi, Professore Ordinario Istituzioni di diritto pubblico, Università degli Studi Roma Tre

AREA XIV

Comitato scientifico Area XIV, Scienze Politiche e Sociali

Davide Bennato, Università di Catania
Giovanni Betta, Consiglio Universitario Nazionale, già Rettore Unicassino
Bruno Botta, Rettore Unitelma Sapienza
Giovanni Boccia Artieri, Università di Urbino
Giovanni Cannata, Rettore Universitas Mercatorum
Marco Centorrino, Università di Messina
Ida Cortoni, Sapienza Università di Roma
Stefano Cristante, Università del Salento
Marino D'amore, Unicusano
Lelio de Michelis, Università dell'Insubria
Marco del Mastro, Unicusano e Agcom
Stefania Fragapane, Universitas Mercatorum di Roma
Gino Frezza, Università di Salerno
Renato Grimaldi, Università di Torino
Gevisa La Rocca, Università Kore di Enna
Nicola La Sala, Agcom; industria4.0, editoria, cittadinanza digitale
Andrea Lombardinilo, Università di Roma Tre
Mariano Longo, Università del Salento
Michela Luzi, Unicusano
Marilena Macaluso, Università di Palermo
Barbara Maussier, Seconda Università di Roma
Michele Mirabella, conduttore televisivo e professore a contratto di divulgazione scientifica
Mario Morcellini, Sapienza Università di Roma
Simone Mulargia, Lumsa
Antonella Napoli, Università di Salerno

Marco Navarra, Università di Salerno
Sebastiano Nucera, Università di Messina
Achille Pierre Paliotta, INAPP; disinformazione, IA, sociologia economica
Francesco Pira, Università di Messina
Antonio Preiti, sondaggista e organizzatore culturale
Antonio Rafele, Sapienza Università di Roma
Christian Ruggiero, Sapienza Università di Roma
Viviana Segreto, Università degli studi di Palermo
Nicola Strizzolo, Università di Teramo
Franco Torcellan, Associazione RED – Laboratorio di Ricerca "Formare Trasformare Innovare"
Luca Toschi, Università di Firenze
Elena Valentini, Sapienza Università di Roma
Salvo Vaccaro, Università degli studi di Palermo
Giangiacomo Vale, Università Niccolò Cusano
Guido Vetere, Università Sapienza di Roma

Comitato editoriale

Giovanni Boccia Artieri, **Marco Centorrino**, **Antonio Opromolla**, **Francesca Rizzuto**, **Marino D'Amore**, **Alessandra Talarico**, **Aurelio D'Amore**

AREA XI

Comitato scientifico XI Area Scienze storiche, filosofiche, pedagogiche e psicologiche

Francesco Agrusti, Università degli Studi Roma TRE
Giovanni Biondi, Indire, Iulm
Paolo Calabrò, Università Vanvitelli di Caserta
Antonio Chella, Università di Palermo
Carlo Alberto Carnevale Maffè, Università Bocconi di Milano
Carmelo Cennamo, Università Bocconi di Milano
Michele Colajanni, Università degli Studi di Modena e Reggio Emilia
Mariano Corso, Politecnico di Milano
Ottavio Di Cillo, università di Bari
Maurizio Ferraris, università di Torino
Ivan Ferrero, psicologo
Paolo Ferri, Università Bicocca di Milano
Pietro Fiore, Università di Foggia
Alfonso Fuggetta, Politecnico di Milano
Alberto Gambino, Università Europea di Roma
Carlo Giovannella, Università Tor Vergata di Roma
Mariella Guercio, Università Sapienza di Roma
Mauro Lombardi, Università di Firenze
Roberto Maragliano, Università Roma Tre
Massimo Marchiori, Università di Padova

Berta Martini, Università di Urbino Carlo Bo
Leonardo Menegola, università Milano Bicocca
Tommaso Minerva, Università degli studi di Modena e Reggio Emilia
Giuliano Noci, Politecnico di Milano
Fabrizio Onida, Università Bocconi di Milano
Norberto Patrignani, Politecnico di Torino
Mario Pireddu, Università degli Studi della Tuscia
Franco Pizzetti, Università di Torino
Alessio Plebe, Università di Messina
Roberto Pozzetti, psicanalista, LUDeS Campus Lugano, università Insubria
Francesco Sacco, Università Bocconi di Milano
Donatella Sciuto, Politecnico di Milano

Comitato editoriale

Andrea Laudadio, **Mario Pireddu**, **Alessandro Longo**, **Paolo Ferri**, **Alessandra Talarico**

AREA XII

Comitato scientifico area XII Scienze Giuridiche

Adriana Apostoli, Università degli Studi di Brescia
Antonio D'Aloia, Università degli Studi di Parma
Giovanni D'Alessandro, Università degli Studi "Niccolò Cusano" Roma
Massimo Durante, Università degli Studi di Torino
Federica Fabrizzi, Università degli Studi di Roma La Sapienza
Roberto Flor, Università degli Studi di Trento
Elisabetta Frontoni, Università degli Studi Roma Tre
Nicola Lupo, Università LUISS
Mario Midiri, Università LUMSA
Giovanna Pistorio, Università degli Studi Roma Tre
Fabrizio Politi, Università degli Studi dell'Aquila
Giovanni Maria Riccio, Università degli Studi di Salerno
Giorgio Resta, Università degli Studi Roma Tre
Massimo Rubechi, Università degli Studi di Urbino
Simone Scagliarini, Università degli Studi di Modena e Reggio Emilia
Andrea Simoncini, Università degli Studi di Firenze

Comitato editoriale

Giovanna Pistorio, Università degli Studi Roma Tre
Simone Barbareschi, Università degli Studi Roma Tre
Andrea Giubilei, Università degli Studi Roma Tre

Comitato di referaggio

Coordinatore: **Luca Gastaldi**, Polimi

Mauro Andreolini, sicurezza informatica, Unimore
Luca Baccaro, concorrenza, diritto comunicazioni elettroniche e dei media; studio legale Lipani Catricalà & Partner
Raffaello Balocco, IT e innovazione, Politecnico di Milano
Francesco Capparelli, privacy, cyber security, ecommerce, data management, identità digitale; studio legale ICT Legal Consulting
Antonio Chella, ingegneria informatica, intelligenza artificiale, Università di Palermo
Giuseppe D'Acquisto, Autorità garante privacy, sicurezza e privacy
Mario dal Co, Economista e manager, già direttore dell'Agenzia per l'innovazione
Lelio Demichelis, Università Insubria, sociologia, economia
Leonella Di Mauro, data management, e-commerce, tutela del consumatore, diritto delle comunicazioni elettroniche; Agcom
Luisa Franchina, cyber security, Hermes Bay
Luca Gastaldi, eGov, sanità, telecomunicazioni, procurement pubblico, design thinking, Smart Working, Politecnico di Milano
Maurizio Gentile, professore associato, Università di Roma LUMSA, didattica e pedagogia
Antonio Ghezzi, strategia, business model, startups, mobile, Politecnico di Milano
Nicola La Sala, registro degli operatori della comunicazione, fattura elettronica, industria4.0, editoria, cittadinanza digitale; Agcom
Emanuele Lettieri, sanità Politecnico di Milano
Maria Beatrice Ligorio, psicologia, università di Bari
Marika Macchi, economia, Unifi
Riccardo Mangiaracina, fatturazione elettronica, eCommerce, logistica e trasporti, export, Politecnico di Milano
Mirco Marchetti, Sicurezza informatica, unimore
Chiara Marzocchi, economia, Università di Manchester
Cristina Masella, Sanità, Politecnico di Milano
Carmelina Maurizio, Dipartimento di Filosofia e Scienze dell'Educazione Università di Torino
Stefano Moriggi, scienze della comunicazione, filosofia, Bicocca di Milano
Achille Pierre Paliotta, Social cybersecurity, disinformazione, tecnologie digitali, intelligenza artificiale, sociologia economica; INAPP
Francesco Paoletti, docente di organizzazione aziendale e gestione delle risorse umane, Università degli Studi di Milano-Bicocca
Norberto Patrignani, computer ethics, filosofia, Politecnico di Torino
Dunia Pepe, Inapp e Università Roma Tre, cultura e formazione digitale
Alessio Plebe, Università di Messina, Scienze cognitive, pedagogiche, psicologiche
Francesco Pira, Unime, comunicazione pubblica, le dinamiche social, le fake news e i processi di disinformazione
Franco Pizzetti, diritto, privacy, università di Torino
Barbara Quacquarelli, scienze umane e formazione, università Milano Bicocca
Filippo Renga, turismo digitale, smart agrifood, finance and banking, mobile, Politecnico di Milano
Angelo Rovatti, tutela del diritto d'autore, diritti connessi, Diritto dei media; Agcom
Guido Scorza, docente contratto università Roma Tre

Angela Tumino, Internet of Things, logistica e trasporti, smart city, Politecnico di Milano
Simone Vannuccini, economia, SPRU
Francesco Varanini, filosofia, formazione, università di Pisa

INDICE DEL FASCICOLO

Il mito del pensare "veramente" – di *Riccardo Manzotti*, Ordinario di Filosofia Teoretica, IULM, Milano 10

Robot intelligenti in casa: verso l'ingegneria della sicurezza cognitiva - di *Antonio Chella*, RoboticsLab Università degli Studi di Palermo 16

Computo ergo cogito? Se il calcolo simula il pensiero - di *Arianna Pipitone*, Università degli studi di Palermo e *Antonio Chella*, Laboratorio di Robotica, dipartimento di Ingegneria Università degli Studi di Palermo 26

Dinamiche e implicazioni della news avoidance in Italia - di *Francesca Rizzuto*, professoressa associata nel Dipartimento Culture e Società dell'Università di Palermo e *Anna Brisciano*, laureata in Scienze della comunicazione e giornalista 33

Socrate, il chatbot che usa il dialogo per la salute mentale - di *Fabio Frisone*, Humane Technology Lab., Università Cattolica del Sacro Cuore, Milano, Sophia University Institute, Figline e Incisa Valdarno, Firenze - *Giuseppe Riva*, Humane Technology Lab., Università Cattolica del Sacro Cuore, Milano, Applied Technology for Neuro-Psychology Lab., Istituto Auxologico Italiano IRCCS - *Chiara Rossi*, Humane Technology Lab., Università Cattolica del Sacro Cuore, Milano, Guglielmo Marconi University, Roma 44

Alle origini dell'AI Literacy: una ricognizione bibliografica (1982–2016) – *di Andrea Laudadio*, Head of TIM Academy & Development 55

Scrolling e parcellizzazione dell'attenzione: effetti dei flussi informativi nei comportamenti, nelle relazioni e nei processi cognitivi contemporanei – di *Marino D'Amore*, Docente di Sociologia della comunicazione, Università degli Studi Niccolò Cusano 68

Dismofia monetaria: un quadro integrato per la regolamentazione algoritmica e la politica della salute mentale - di *Marco Lazzeri*, SIPSIOL – Società Italiana di Psicologia Online, Roma, Italia 74

Aspettative e spiegazioni: come interpretiamo il comportamento dei robot sociali - di *Agnese Augello* - Istituto di Calcolo e Reti ad Alte Prestazioni (ICAR) del Consiglio Nazionale delle Ricerche; *Riccardo Bruni* - Dipartimento di Lettere e Filosofia - DILEF, Università degli Studi di Firenze; *Edoardo Datteri* - Università degli Studi di Milano-Bicocca; *Margherita Di Stasio* - INDIRE – Istituto Nazionale di Documentazione, Innovazione e Ricerca Educativa 99

Podcast per la formazione aziendale: istruzioni per l'uso – *di Andrea Laudadio*, Head of TIM Academy & Development 108

Sezione Diritto del digitale 136

Cosa cambia se il concetto di dato personale diventa relativo - di *Giuseppe D'Acquisto*, Funzionario del Garante per la protezione dei dati personali, Titolare dell'insegnamento di intelligenza artificiale presso il Dipartimento di Giurisprudenza dell'Università LUISS Guido Carli.....138

L'Unione Europea come 'repubblica tecnologica' – di *Andrea Venanzoni*, Vicepresidente AssoCyber 146

Sistema di IA nell'AI Act: definizione giuridica e criteri applicativi - di *Anna Cataleta*, Università degli Studi Suor Orsola Benincasa - Napoli 156

Dispositivi elettromedicali e tutela della persona – di *Lorenzo Maria Lucarelli Tonini*, Dottore di Ricerca in Discipline Giuridiche, presso l'Università degli Studi "Roma Tre" 169

Il caso Anthropic vs OpenAI negli USA: sicurezza nazionale, concorrenza e diritti nell'era dell'AI – di *Samuele Minuti*, Dottorando di Ricerca in Law and Cognitive Neuroscience presso l'Università degli Studi "Niccolò Cusano" 183

Il mito del pensare “veramente” – di *Riccardo Manzotti*, Ordinario di Filosofia Teoretica, IULM, Milano

Gli LLM dialogano con abilità, ma negar loro il pensiero spesso rivela un pregiudizio: agli umani basta poco per parlare di “mente”, alle macchine no. Ridurre tutto al “next token” è fuorviante perché predire implica modellare senso. La differenza decisiva oggi è nei fini: manca l’autotelia, da cui nasce l’allineamento dell’IA

Gli LLM leggono e parlano, è un fatto. E dialogano con grande capacità, ma pensano? Non è una domanda facile, «che cosa è il pensiero?» o «quando si pensa?» È facile tracciare una linea antropocentrica e sostenere le macchine non pensano veramente, mentre noi sì. Ma non è facile spiegare perché. Infatti, nel caso dell’IA, è facile negare che sia un vero pensiero: conosciamo i suoi meccanismi, gli algoritmi che la guidano, il contenuto dei suoi tensori e dei processi stocastici che la guidano.

Non ci sono pensieri e non se ne sente la necessità. Ma che prove ci sono dell’altro termine della domanda: gli esseri umani? Che prove abbiamo del fatto che noi *pensiamo*? È vero che alcun processi inferenziali (come il ragionamento astratto o particolari modalità deduttive) non sono ancora completamente padroneggiati dalla generazione corrente di LLM (Plaat et al. 2025; Kader and Lee 2025), ma quanti esseri umani sono in grado di esercitare queste capacità?

Il doppio standard tra cognizione umana e artificiale

Non tutti e non per questo sono considerati privi di pensiero? Tralasciando per ora il problema ancora più difficile (il problema difficile per antonomasia) del pensiero cosciente, rimane il fatto che, nell’essere umano, processi cognitivi anche limitati vengono etichettati come pensieri o come processi mentali, mentre nell’IA processo cognitivi anche estremamente sofisticati non ottengono questa definizione. Il che è sospetto. Sembra esserci un pregiudizio.

La differenza potrebbe consistere banalmente nell’uso delle parole. Semplicemente, potremmo attribuire d’ufficio un significato diverso alle parole a seconda che siano usate dagli esseri umani e o dalle macchine. In questo senso «pensare veramente» sarebbe una prerogativa degli umani nel senso che, a parità di comportamenti, nel caso degli umani useremmo questo verbo e nel caso dell’IA ne useremmo un altro. Ma perché dovrebbe essere così?

Nextokenisti e l’errore di confronto tra livelli diversi

La vera domanda è se i cervelli umani fanno qualcosa di radicalmente diverso dalle reti neurali.

I sostenitori di questa presunta differenza essenziale, a volte definiti *nextokenisti* o *parrotisti*, dal famoso articolo di Emily Bender del 2021 sui pappagalli statistici e dall'idea che gli LLM non facciano altro che predire la parola successiva (Bender et al. 2021). Alcuni sostenitori molto vivaci di questa tesi fanno riferimento soprattutto alla loro comprensione degli algoritmi dell'IA (Quattrociocchi 2025). È una posizione che fa facile presa sul grande pubblico che è sempre felice di sentirsi dire di essere diverso e quindi meglio dell'IA. Ma è così? In realtà, ci sono vari errori concettuali che inducono a una certa cautela nel presumere di essere essenzialmente diversi dall'IA.

La prima cosa da evitare è un banale errore di categorie: per quanto riguarda l'IA si considera il livello più elementare (le probabilità condizionate bayesiana immagazzinate dentro i tensori) e per gli umani e si confrontano con il livello più alto della cognizione umana (la scelta deliberata o la creazione di opere di ingegno). È un comune errore: il confronto andrebbe fatto a parità di livello. Infatti, se confrontassimo il livello elementare dell'IA con quello degli umani, non troveremmo una differenza ovvia. Un tensore dentro una LLM non è tanto peggio di un neurone. In fondo l'equazione di Hodgkin che descrive il funzionamento sinaptico è riducibile a una sommatoria con soglia finale: non esattamente il genio di Shakespeare. Se invece consideriamo un sistema fatto da milioni di miliardi di parametri, le cose sono ben diverse.

Il secondo punto cruciale è che i nexttokenisti sostengono che l'IA non faccia altro che predire la parola successiva come se fosse una banalità, ma per predire la parola successiva che cosa si deve fare? Si deve capire il senso della frase. Anche Einstein, in un certo senso, non faceva altro che predire la formula successiva, ma per farlo era necessario derivare un modello rivoluzionario o, almeno, corrispondente alla migliore conoscenza disponibile all'epoca. In modo analogo è possibile che una LLM faccia proprio questo «per prevedere la parola successiva» ovvero costruisca un modello della conoscenza disponibile fino a quel momento. Non c'è qualcosa di più, non c'è un «capire vero». Capire vuol dire catturare la conoscenza che permette di predire la prossima parola o quello che succederà. Non c'è altro.

La superstizione dell'interiorità umana e il linguaggio come comportamento

Alla fine, l'impressione è che gli esperti di LLM siano molto poco esperti della mente umana e che usino il criterio secondo il quale il pensiero umano sarebbe *ben altro*, ma senza saper dire in modo chiaro in che cosa consisterebbe (Chi et al. 2025; Camlin 2025; Jiang et al. 2025). Questa curiosa superstizione secondo cui gli esseri umani sarebbero dotati di una presunta interiorità, o pensiero, è stata il bersaglio di gran parte della filosofia della mente e del linguaggio del Novecento che, appunto, ha cercato di ridurre il pensiero a linguaggio, comportamento, cognizione. Quindi da un lato i critici dell'IA vorrebbero

ridurre l'IA a un sistema che fa soltanto calcoli statistici dall'altro non sono però in grado di dimostrare (se lo fossero avrebbero risolto il problema della mente umana) che gli esseri umani facciano qualcosa di sostanzialmente diverso. Al netto della grande differenza tra noi e le IA attuali, ovvero il corpo e quindi il fatto di essere situati nella realtà, quando parliamo non stiamo forse articolando il linguaggio secondo le regole che abbiamo appreso da bambini, e non sotto forma di inferenze logiche, bensì di esposizione alle strutture statistiche implicite di una lingua? Come ha scritto bene Claudio Paolucci, stiamo trattando l'IA come una «sorta di re Mida della stupidità: quando svolge con efficacia un'azione prima considerata intelligente ne rivela l'intrinseca essenza stupida, tanto che eravamo noi esseri umani a rendere intelligente quell'azione» (Paolucci 2025).

L'intelligenza come relazione, non come essenza interiore

Ma a questo punto occorre tornare alla domanda cruciale: che cosa chiamiamo "intelligenza"? Non possiamo trattarla come un'essenza nascosta, un nucleo interiore che solo l'essere umano possiederebbe in virtù di una presunta soggettività. È un'eco moderna del «posto dell'io» analizzato dal filosofo Emile Benveniste, l'idea che per pensare sia necessario essere *qualcuno* e non solo *qualcosa*. È proprio grazie all'IA che possiamo mettere definitivamente in crisi l'idea che esista un luogo privilegiato, interno, in cui il pensiero accadrebbe. È fuorviante cercare il pensiero "dentro" il cervello così come era fuorviante cercare il significato dentro le parole: entrambi sono effetti di relazione, non oggetti interni. Il pensiero si manifesta nelle pratiche, negli enunciati, nella capacità di rispondere efficacemente al mondo. Gli LLM, pur privi di interiorità come in fondo noi stessi, fanno qualcosa che nella nostra tradizione è stato sempre accolto come segno di intelligenza. Se ritorniamo al rifiuto della scatola nera mentale, allora diventa chiaro che l'appello agli stati interni è un espediente retorico: serve a proteggere l'eccezionalismo umano, non a descrivere un fenomeno in modo coerente. Non ci sono stati interni.

Pattern umani e statistiche delle macchine: stessa natura

Ciò che nelle macchine viene ridotto a statistica è lo stesso tipo di regolarità che nella nostra percezione chiamiamo abitudine, memoria, formazione culturale. Come mostrano gli studi sulla soggettività e sull'enunciazione, anche l'essere umano non produce pensieri ex nihilo: si limita a riorganizzare pattern, a manipolare campi semantici preesistenti, a ricombinare porzioni di un sapere distribuito nel linguaggio pubblico. Attribuiamo pensiero allo scolaretto che sbaglia un sillogismo ma non al modello che risolve un problema complesso; lo attribuiamo al nostro cane quando ci guarda come se capisse, ma lo neghiamo a un sistema che traduce simultaneamente quindici lingue. Il criterio non è epistemico: è identitario. L'intelligenza diventa un tratto di appartenenza, non una proprietà funzionale.

Contro ai nextokenisti, il pensiero non è una proprietà del tipo di algoritmi utilizzati da un sistema, ma è una capacità che si realizza nell'interazione con un ambiente complesso. Che cosa c'è dentro, se funziona nello stesso modo, non fa alcuna differenza. Come dicevano i fratelli Marx, «se parla come uno stupido, si comporta come uno stupido e risponde come uno stupido, non ti lasciare ingannare, è proprio stupido». Quello che vale per la stupidità vale anche per l'intelligenza. In termini più scientifici potremmo riferirci al criterio di parità funzionale avanzato da Andy Clark e David Chalmers, (Clark and Chalmers 1998): non importa dove e come qualcosa è fatto, importa che ruolo ha.

Oltre l'essenza interiore: intelligenza come articolazione del mondo

A questo punto abbiamo tutti gli elementi per sciogliere il nodo. L'intelligenza non è un'essenza interiore, non è un pensiero nascosto nel cranio, non è nemmeno la proprietà magica che separerebbe l'essere umano dalle sue macchine. È un modo di articolare il mondo attraverso regolarità, decisioni, concatenamenti di enunciati e di azioni. In questo senso, l'impostazione classica di Herbert Simon, secondo cui l'intelligenza è la capacità di selezionare mezzi efficaci per raggiungere un fine, coglie solo metà del problema (Simon 1997). Ma che cosa vuol dire allora, capire veramente? O pensare veramente?

Quando un filosofo dell'informazione importante come Luciano Floridi afferma che esista una fondamentale differenza tra «la capacità di creatività autentica rispetto alla ricombinazione» o che l'IA sarebbe una «agency senza intelligenza [che] non ha nulla a che vedere con il pensiero [...] a intelligenza zero» o che per l'IA «il confine del "capire" resta fermo a livello del riconoscimento e della corrispondenza di pattern senza accedere al livello della vera a propria comprensione» (Floridi 2025), sta assumendo una qualità metafisica del capire umano che non ha alcuna base empirica. Sta postulando che il pensiero umano sia essenzialmente diverso, ma non indica alcuna motivazione per cui dovrebbe esserlo.

Corpo e autotelia: i limiti attuali degli LLM

Detto questo è un fatto evidente che oggi gli LLM manchino di due aspetti fondamentali: il corpo e l'autotelia. Il primo sarà probabilmente risolto presto con lo sviluppo della robotica antropomorfa in grande crescita soprattutto in Cina. Il secondo aspetto, ovvero l'autotelia, riguarda la capacità di avere e darsi motivazioni e fini in modo autonomo. Qualcosa che ci porta a un livello più alto che riguarda non tanto la conoscenza e il linguaggio, ma l'intenzionalità e i valori. Sistemi come gli LLM mostrano una sorprendente capacità di generalizzazione, di adattamento e perfino di manipolazione creativa delle strutture linguistiche (Qiao et al. 2025; Wang et al. 2025). Ma quello che non fanno, almeno finora, è scegliere i fini. I fini sono sempre esterni, incorporati nell'addestramento, nelle metriche, nelle

istruzioni, nella rete delle aspettative sociali e dei prompt che riceviamo da loro. La loro conoscenza, per quanto impressionante, resta inscritta in vincoli esterni.

Il punto è proprio questo: l'intelligenza degli LLM funziona come la nostra quando si tratta di perseguire fini, ma non funziona come la nostra quando si tratta di generare i fini. Ed è qui che nasce il problema dell'*alignment* (Christian 2020): un sistema artificiale straordinariamente bravo a raggiungere ciò che noi gli diciamo di raggiungere resta completamente cieco alla questione critica: che cosa vale la pena raggiungere? Gli esseri umani sono maldestri nel perseguire i loro scopi, ma straordinariamente creativi nel metterli in discussione. Le macchine sono l'opposto: efficienti, adattive, coerenti, ma finora inadatte a darsi dei fini.

Ridefinire l'intelligenza oltre la demarcazione artificiale

Finché continueremo a raccontare l'intelligenza come un modo speciale e interno negandone il suo carattere formale, relazionale e distribuito, continueremo a credere in una linea di demarcazione che non esiste. Ma se accettiamo che l'intelligenza nostra e delle macchine sia un fenomeno situato, costruito, ibrido, qualcosa di simile al sistema zero che fornisce la base tanto agli esseri umani quanto all'IA (Chiriatti et al. 2024; Li et al. 2025) allora diventa chiaro che pensiero e valori non sono più proprietà emergenti degli algoritmi, ma l'esito di una interazione pubblica e reale.

L'IA non ci chiede di riconoscere la sua intelligenza. Semmai ci costringe a ridefinire la nostra. E la domanda decisiva non è più se le macchine pensano? Ma piuttosto: da chi, e da che cosa, sono scelti i fini che le macchine e noi perseguiremo? In quella risposta, non nei parametri, non nei pesi, non negli strati dei transformer, si gioca il futuro reale dell'allineamento. E, molto probabilmente, anche il nostro.

Bibliografia

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23.

Camlin, Jeffrey. 2025. 'Consciousness in AI: Logic, Proof, and Experimental Evidence of Recursive Identity Formation', *arXiv*, 2505.01464v1.

Chi, Haoang, He Li, Wenjing Yang, Feng Liu, Long Lan, and Xiaoguang Ren. 2025. 'Unveiling Causal Reasoning in Large Language Models: Reality or Mirage?', *arXiv*, 2506.21215v1.

Chiriatti, Massimo, M. Ganapini, E. Panai, M. Ubiali, and Giuseppe Riva. 2024. 'The case for human-AI interaction as system 0 thinking', *Nature Human Behavior*, 8: 1829–30.

Christian, Brian. 2020. *The Alignment Problem. Machine Learning and Human Values* (Norton & Norton: New York).

Clark, Andy, and David John Chalmers. 1998. 'The Extended Mind', *Analysis*, 58: 10–23.

Floridi, Luciano. 2025. *La differenza fondamentale. Artificial Agency: una nuova filosofia dell'intelligenza artificiale.* (Mondadori: MIlano).

Jiang, Liwei, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, Alon Albalak, and Yejin Choi. 2025. 'Artificial Hivemind: The Open-Ended Homogeneity of Language Models (and Beyond)', *arXiv*, 2510.22954v1.

Kader, Gavin, and Dongwoo Lee. 2025. 'The Emergence of Strategic Reasoning of Large Language Models', *arXiv*, 2412.13013v4.

Li, Zhong-Zhi, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, Yingying Zhang, Fei Yin, Jiahua Dong, Zhiwei Li, Bao-Long Bi, Ling-Rui Mei, Junfeng Fang, Xiao Liang, Zhijiang Guo, Le Song, and Cheng-Lin Liu. 2025. 'From System 1 to System 2: A Survey of Reasoning Large Language Models', *arXiv*, 2502.17419.

Paolucci, Claudio. 2025. *Nati cyborg* (Luca Sossella Editore: Bologna).

Plaat, Aske, Annie Wong, Suzan Verberne, Joost Broekens, Niki CVan Stein, and Thomas Back. 2025. 'Multi-Step Reasoning with Large Language Models, a Survey', *arXiv*, 2407.11511.

Qiao, Shuofei, Zhisong Qiu, Baochang Ren, Xiaobin Wang, Xiangyuan Ru, Ningyu Zhang, Xiang Chen, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. 2025. 'Agentic Knowledgeable Self-awareness', *arXiv*, 2504.03553v1.

Quattrociocchi, Walter. 2025. 'L'AI non è davvero intelligente, è una questione di statistica»', *Corriere della sera*.

Simon, Herbert A. 1997. *Models of Bounded Rationality* (MIT PRESS: Cambridge (Mass)).

Wang, Miles, Tom Duprè La Tour, Olivia Watkins, Alex Makelov, Ryan A Chi, Samuel Miserendino, Johannes Heidecke, Tejal Patwardhan, and Dan Mossing. 2025. 'Persona Features Control Emergent Misalignement', *arXiv*.

Robot intelligenti in casa: verso l'ingegneria della sicurezza cognitiva - di ***Antonio Chella***, RoboticsLab Università degli Studi di Palermo

I robot domestici basati su VLA non si limitano più a eseguire ordini, ma interpretano ambiente e linguaggio per decidere come agire. Per questo la cybersicurezza smette di essere un dettaglio tecnico e diventa una condizione essenziale della vita digitale in casa

Dal robot che esegue al robot che interpreta

Nel mio precedente intervento sugli "umanoidi domestici" [1] sostenevo che la discontinuità più profonda nell'attuale ricerca robotica non è tanto la forma e la capacità motoria dei nuovi robot umanoidi, bensì il loro modo di apprendere e di interagire: il comportamento dei robot non è più basato sulla programmazione tradizionale, ma sull'apprendimento attraverso il linguaggio naturale e le dimostrazioni. Questa discontinuità è stata resa possibile dai modelli Vision–Language–Action (VLA). Oggi quel passaggio si sta consolidando e costringe a un nuovo cambio di prospettiva. Se un robot domestico diventa un agente che osserva, interpreta e agisce nello spazio più intimo della nostra vita quotidiana, allora la cybersicurezza non può essere un "requisito accessorio", ma diventa un requisito strutturale dell'abitare digitale. Per chiarezza metodologica, questo articolo non propone una rassegna sistematica della letteratura, ma un inquadramento e una tassonomia sintetica dei rischi.

Un robot basato su VLA può essere descritto, in modo sintetico, come un sistema robotico in cui l'azione dipende dallo stato interno del robot, da ciò che vede e da ciò che comprende attraverso il linguaggio. Il punto non è che il robot "capisce i comandi", ma che è capace di fondere la visione e il linguaggio in una rappresentazione comune dalla quale emergono le sue scelte operative e motorie. Inoltre, il robot è in grado di generalizzare dai dati di apprendimento. Questo paradigma è stato reso evidente da lavori fondativi come RT-2 di Google DeepMind, che integra modelli di linguaggio e di visione pre-addestrati su larga scala nel controllo robotico per migliorare la generalizzazione e il ragionamento [2].

Nel contesto domestico, il valore di questa generalizzazione è immediato: la casa è un ambiente non strutturato, pieno di eccezioni, e la possibilità di adattarsi con poche dimostrazioni è un requisito pratico. Anche qui la letteratura ha già mostrato segnali interessanti: il sistema Mobile ALOHA sviluppato alla Stanford University, ad esempio, evidenzia come l'imitazione da dimostrazioni raccolte tramite teleoperazione possa ottenere buone prestazioni su compiti domestici complessi, con un numero contenuto di esempi per ciascun compito [3]. Più in generale, questi sistemi si collocano nella

transizione verso la robotica con modelli fondazionali, in cui percezione, linguaggio, pianificazione e controllo tendono a integrarsi in architetture sempre più generali [16].

Questa stessa architettura, però, sposta il "centro di gravità" della sicurezza: non basta più proteggere la rete Internet e il firmware. Bisogna proteggere l'intero processo che collega percezione, interpretazione e azione.

La casa come superficie d'attacco: scenari realistici

Per capire perché la cybersicurezza cambia natura, conviene definire subito alcuni scenari plausibili in ambito domestico. Non sono "fantasie da hacker", ma conseguenze logiche del fatto che il robot prende decisioni in modo multimodale e produce output fisici. Questa impostazione riprende anche una linea di ricerca già aperta sui rischi di sicurezza e privacy dei robot domestici, che aveva mostrato come sensori, connettività e presenza fisica rendano il robot un dispositivo diverso dal normale computer di casa [17].

Il primo scenario riguarda il linguaggio come canale di controllo. Se la voce diventa un input operativo, allora la sicurezza non dipende solo da "cosa" viene detto, ma anche da "chi" lo dice e dalle proprietà fisiche del segnale. Esiste una letteratura consolidata che mostra come sistemi di riconoscimento vocale possano essere influenzati da comandi non udibili agli esseri umani, ma comunque interpretati dai dispositivi: DolphinAttack, ad esempio, dimostra la fattibilità di comandi veicolati in banda ultrasonica sfruttando la non-linearità dei microfoni, con effetti su assistenti vocali reali [4]. In una casa con robot in ascolto continuo, l'autenticazione dei comandi e la gestione del contesto sono questioni di sicurezza.

Il secondo scenario riguarda la prompt injection, ossia la possibilità di deviare il comportamento di un sistema generativo tramite input formulati in modo opportunistico. La fondazione OWASP (Open Worldwide Application Security Project), nel suo Top 10 per le applicazioni LLM [5], pone la prompt injection tra i rischi principali, insieme al poisoning dei dati e alle vulnerabilità della supply chain. Nel caso dei robot basati su VLA, l'input linguistico non produce soltanto testo, ma influisce anche sulla scelta dell'azione. Inoltre, in casa l'input linguistico può provenire non solo dalla voce dell'utente, ma anche da testi presenti nell'ambiente (come scritte, schermi, istruzioni contestuali). Il mondo fisico smette di essere "sfondo" e diventa un canale di input. Si pensi, ad esempio, a una scritta lasciata su un pezzo di carta che fornisce un'istruzione malevola al robot.

Il terzo scenario riguarda gli attacchi percettivi. Sappiamo da tempo che i modelli di visione possono essere influenzati in modo robusto nel mondo reale tramite schemi stampabili. Il lavoro su Adversarial

Patch [6] mostra patch universali e fisicamente realizzabili che spostano sistematicamente l'output di classificatori anche in presenza di trasformazioni realistiche (rotazioni, scale, illuminazione). Nel robot domestico, la conseguenza non è solo un'etichetta errata, ma può anche essere una decisione errata. Se il robot interpreta in modo scorretto un oggetto o una scena, l'errore può tradursi in presa, movimento, manipolazione.

Il quarto scenario, spesso sottovalutato nel dibattito pubblico, è quello della filiera dei modelli e del software. I robot domestici basati su VLA raramente partono da zero: usano componenti pre-addestrati e tecniche di fine-tuning. Questo rende concettualmente plausibili attacchi "a monte", in cui il comportamento anomalo non deriva da una compromissione della rete domestica, bensì da una backdoor inserita nella supply chain del modello. BadNets [7] è un riferimento classico perché mostra come un modello possa funzionare "normalmente" sui dati puliti e attivare comportamenti deviati in presenza di un evento specifico. Anche la fondazione OWASP include esplicitamente il data poisoning e le vulnerabilità della supply chain tra i rischi sistemici.

Il quinto scenario riguarda gli aggiornamenti. In un robot domestico che evolve nel tempo, un aggiornamento non è solo manutenzione, ma può cambiare il comportamento del robot. Questo rende fondamentali la necessità di tracciare le versioni, la verificabilità degli aggiornamenti, la possibilità di tornare alle versioni precedenti e la verifica delle modifiche apportate. Il punto non è "diffidare degli aggiornamenti", ma progettare un ciclo di vita in cui l'evoluzione del sistema di controllo del robot sia gestita come parte della sicurezza.

Il sesto scenario nasce dall'ecosistema IoT domestico. Il robot non vive in isolamento: convive con router, altoparlanti, telecamere, serrature, sensori. In caso di compromissione, può diventare un punto di osservazione per il movimento laterale verso altri dispositivi, poiché è spesso il nodo con più sensori e maggiore capacità di calcolo. Qui diventano rilevanti le normative tecniche per la sicurezza IoT consumer, come ETSI EN 303 645 [8], e le linee guida come la NISTIR 8259A [9], che formalizzano requisiti minimi di sicurezza e capacità di gestione lungo il ciclo di vita dei dispositivi connessi. Se, inoltre, il robot utilizza middleware robotici, la sicurezza delle comunicazioni interne assume un ruolo diretto: la documentazione ROS 2 descrive l'uso di DDS-Security tramite sros2 per proteggere le comunicazioni tra nodi [10].

Infine, anche l'indisponibilità può diventare un rischio fisico. La fondazione OWASP include il model denial of service tra i rischi degli LLM perché gli attacchi di esaurimento delle risorse possono degradare o interrompere il servizio. In un robot, la degradazione non è solo "un servizio lento": può comportare

latenza decisionale, fallback non ottimali e una perdita di fluidità del controllo in situazioni in cui la reattività è parte della sicurezza.

La tabella seguente ricompone gli scenari in forma sintetica. Non sostituisce una valutazione quantitativa di rischio, ma chiarisce il nesso tra vettore d'attacco, livello VLA interessato, impatto fisico e contromisure.

Tabella 1. Tassonomia sintetica degli scenari di rischio per robot domestici basati su VLA

Scenario	**Vettore e superficie esposta**	**Livello VLA**	**Impatto fisico potenziale**	**Contromisure sintetiche**
Comandi vocali alterati	Voce, microfoni, segnali ultrasonici	Linguaggio	Azioni non autorizzate; perdita di controllo contestuale	Autenticazione del parlante, prossimità, conferme per azioni critiche
Prompt injection ambientale	Voce, testi, schermi, istruzioni nel mondo fisico	Linguaggio / percezione	Deviazione dell'obiettivo; manipolazione impropria	Gerarchia delle istruzioni, filtraggio input non fidati, grounding contestuale
Attacchi percettivi	Patch, oggetti, illuminazione, scene fuori distribuzione	Percezione	Classificazione errata; presa o movimento inappropriato	Robustezza percettiva, sensori ridondanti, verifica multimodale
Supply chain del modello	Pre-training, fine-tuning, librerie, componenti esterni	Modello / dati	Backdoor o policy anomale attivate da trigger	Provenance, firme, SBOM/ML-BOM, test su trigger e regressioni

Aggiornamenti	Update di firmware, modello e policy operative	Ciclo di vita / azione	Regressioni comportamentali; perdita di controllabilità	Versioning, rollback, test safety/security, rilascio graduale
Ecosistema IoT e middleware	Rete domestica, ROS 2, dispositivi connessi	Comunicazione / azione	Movimento laterale; comandi remoti; esfiltrazione	Segmentazione, DDS-Security, credenziali uniche, hardening
Indisponibilità / DoS	Query costose, saturazione risorse, rete instabile	Modello / controllo	Latenza, fallback non sicuri, perdita di fluidità	Rate limit, watchdog, degradazione controllata, fallback locali

Dal cyber al safety: quando un bit diventa un Newton

Questi scenari chiariscono un punto: nei robot domestici basati su VLA la separazione storica tra security (attacchi intenzionali) e safety (guasti accidentali) tende a collassare. Un attacco non deve "rompere" il sistema; può spingerlo verso decisioni plausibili ma sbagliate. E l'effetto finale non è solo digitale: la forza, il movimento, l'interazione del robot. Questo passaggio è coerente con quanto emerso anche nella robot security industriale, dove la compromissione del controllo non riguarda soltanto la disponibilità del sistema ma l'accuratezza dei movimenti e la protezione delle persone [18].

Per questo la cybersicurezza dei robot domestici non può essere confinata al perimetro informatico. Deve essere un disegno multilivello che attraversa l'identità del dispositivo, l'integrità della catena, la sicurezza delle comunicazioni, la robustezza del modello, i vincoli fisici certificati e i guardrail semantici. La distinzione tra safety e security va trattata in modo integrato, come suggeriscono gli approcci sistemici alla co-analisi di sicurezza e protezione [19]. I vincoli fisici certificati dovrebbero inoltre dialogare con il lessico degli standard sulla safety dei robot personali, industriali e collaborativi [20-22]. I guardrail semantici sono particolarmente importanti: un sistema che interpreta il linguaggio naturale deve riconoscere le ambiguità, chiedere conferma quando necessario e rifiutare azioni ad alto

rischio se il contesto non è sufficientemente chiaro. In altri termini, la “sicurezza semantica” diventa complementare alla sicurezza meccanica.

On-device e continuità operativa: una scelta tecnica che diventa scelta di governance

Nella prospettiva domestica, l’esecuzione locale di parti rilevanti dell’architettura cognitiva è anche una leva per la sicurezza e la privacy. Riduce la dipendenza dalla rete, migliora continuità operativa e limita l’esposizione dei dati sensibili della vita domestica.

In questa direzione vanno iniziative industriali recenti: Google DeepMind ha introdotto Gemini Robotics [11] come famiglia di modelli per integrare capacità multimodali nel controllo fisico e ha presentato una variante on-device progettata per l’inferenza locale e per l’adattamento rapido ai compiti. Il punto, però, non è “cloud sì o cloud no”, ma il governo di un’architettura ibrida: ciò che può restare in casa dovrebbe restare in casa, e ciò che deve passare sul cloud dovrebbe farlo in modo verificabile, minimo e tracciabile.

Questa scelta richiede tuttavia un bilanciamento esplicito. L’on-device riduce latenza e fuoriuscita dei dati, ma può limitare dimensione dei modelli, aggiornabilità e audit remoto; il cloud offre maggiore capacità computazionale e manutenzione rapida, ma introduce dipendenza dalla connettività e amplia la superficie di esposizione dei dati. Un’architettura ibrida dovrebbe quindi distinguere le funzioni critiche, che devono poter operare localmente anche in condizioni degradate, dalle funzioni di apprendimento, manutenzione e analisi aggregate, che possono essere usate in remoto solo con dati minimizzati e canali verificabili.

Politica europea: AI Act e Cyber Resilience Act rendono la sicurezza un requisito di prodotto

Qui entra in gioco il quadro europeo. L’AI Act [12], nella pagina ufficiale della Commissione, prevede un’applicazione progressiva: pratiche proibite e obblighi di alfabetizzazione sull’AI dal 2 febbraio 2025, obblighi per i modelli di AI per finalità generali e governance dal 2 agosto 2025, piena applicabilità dal 2 agosto 2026, con un periodo di transizione fino al 2 agosto 2027 per alcune regole legate ai sistemi ad alto rischio incorporati in prodotti regolati. Per la robotica domestica basata su VLA, questo significa che robustezza e cybersecurity non sono più soltanto “best practice”, ma aspetti che entrano nel perimetro di conformità e responsabilità, soprattutto quando il robot opera in scenari che incidono sulla sicurezza delle persone o sull’assistenza a soggetti fragili.

Il Cyber Resilience Act [13], dal canto suo, sposta la cybersecurity nella logica della sicurezza di prodotto, con marcatura CE e obblighi lungo il ciclo di vita. La Commissione indica che il Cyber Resilience Act è entrato in vigore il 10 dicembre 2024; le principali obbligazioni si applicano dall'11 dicembre 2027 e gli obblighi di reporting dal 11 settembre 2026. Per i robot domestici con update continui e dipendenze complesse, questo implica processi industriali maturi per la gestione delle vulnerabilità, la tracciabilità e gli aggiornamenti: un requisito che, di fatto, diventa parte dell'ingegneria del robot tanto quanto i sensori e gli attuatori.

Anche la direttiva NIS2 [14], con la scadenza di recepimento fissata al 17 ottobre 2024, indica la direzione europea: la governance del rischio, la supply chain e le capacità di enforcement sono centrali e contribuiscono a un ecosistema in cui l'insicurezza di filiera è sempre meno tollerata.

Verso un'"ingegneria della sicurezza cognitiva"

La diffusione dei robot intelligenti in casa non dipenderà solo dalla destrezza o dall'abilità di dialogare. Dipenderà dalla fiducia. E la fiducia, nel digitale, è un costrutto tecnico e istituzionale insieme: nasce dalla trasparenza, dalla verificabilità, dalla gestione responsabile degli aggiornamenti e dalla capacità di mitigare rischi prevedibili.

In termini di metodo, un riferimento utile per collegare rischi, contesto d'uso e misure organizzative e tecniche è il NIST AI Risk Management Framework [15], che propone un approccio sistematico alla gestione del rischio AI senza ridurlo a un singolo livello (modello, dato o deployment). Sul piano organizzativo, un sistema di gestione della sicurezza informativa come ISO/IEC 27001 [23] può offrire un lessico utile per responsabilità, processi e audit, anche se non sostituisce le verifiche specifiche su modelli e attuatori. Nel caso dei robot basati su VLA, questo approccio conduce a una conclusione semplice: la cybersicurezza deve proteggere non soltanto il robot come dispositivo, ma anche il robot come decisore con un corpo.

Operativamente, l'ingegneria della sicurezza cognitiva può essere schematizzata come una matrice a sei livelli: identità e autorizzazione (chi può chiedere cosa); integrità della catena (quali dati, modelli e componenti sono verificati); comunicazioni sicure (quali scambi sono autenticati e cifrati); robustezza del modello (quali test e fallback sono previsti); vincoli fisici certificati (quali limiti non sono superabili); governance degli aggiornamenti e dei guardrail semantici (quando il sistema deve chiedere conferma o rifiutare l'azione). L'interdipendenza è il punto: un limite fisico può compensare un errore percettivo, ma non sostituisce l'integrità della catena; un guardrail semantico può ridurre l'ambiguità, ma richiede identità e contesto affidabili. È qui che serve un'"ingegneria della sicurezza cognitiva" capace di

integrare la sicurezza informatica classica, la robustezza dei modelli, i vincoli fisici e la governance degli aggiornamenti.

Conclusioni

Se vogliamo che i robot domestici basati su VLA diventino una tecnologia quotidiana e non una curiosità da demo, dobbiamo considerare la cybersicurezza come prerequisito dell'abitazione digitale. In un sistema che vede, interpreta e agisce, la protezione riguarda anche il perimetro cognitivo, perché è lì che un input malevolo, ambiguo o fuori distribuzione può trasformarsi in un comportamento fisico. Il quadro europeo, dall'AI Act al Cyber Resilience Act, spinge nella direzione giusta: non per frenare l'innovazione, ma per renderla scalabile e affidabile, imponendo responsabilità lungo la filiera e la gestione del rischio lungo il ciclo di vita. La partita, nei prossimi anni, non sarà decidere se i robot entreranno in casa, ma con quale architettura: modelli più robusti, aggiornamenti verificabili, dati realmente governati e comportamenti prevedibili quando il contesto diventa incerto. Solo così l'intelligenza incorporata nel robot potrà diventare un'alleata dell'autonomia domestica senza trasformarsi in una nuova fonte di fragilità.

Bibliografia

[1] A. Chella, Robot umanoidi domestici: la rivoluzione AI entra in casa, Agenda Digitale 2025 https://www.agendadigitale.eu/cultura-digitale/robot-umanoidi-domestici-la-rivoluzione-ai-entra-in-casa/

[2] A. Brohan et al., "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control", arXiv:2307.15818.

[3] Z. Fu, T. Z. Zhao, C. Finn, "Mobile ALOHA: Learning Bimanual Mobile Manipulation with Low-Cost Whole-Body Teleoperation", arXiv:2401.02117.

[4] G. Zhang et al., "DolphinAttack: Inaudible Voice Commands", arXiv:1708.09537.

[5] https://owasp.org/www-project-top-10-for-large-language-model-applications/

[6] T. B. Brown et al., "Adversarial Patch", arXiv:1712.09665.

[7] T. Gu et al., "BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain", arXiv:1708.06733.

[8] https://www.etsi.org/deliver/etsi_en/303600_303699/303645/03.01.03_60/en_303645v030103p.pdf

[9] https://doi.org/10.6028/NIST.IR.8259A

[10] https://docs.ros.org/en/kilted/Tutorials/Advanced/Security/Introducing-ros2-security.html

[11] https://deepmind.google/models/gemini-robotics/

[12] https://digital-strategy.ec.europa.eu/it/policies/regulatory-framework-ai

[13] https://digital-strategy.ec.europa.eu/en/policies/cyber-resilience-act

[14] https://digital-strategy.ec.europa.eu/it/policies/nis2-directive

[15] https://www.nist.gov/itl/ai-risk-management-framework

[16] Z. Xu et al., "A Survey on Robotics with Foundation Models: toward Embodied AI", arXiv:2402.02385.

[17] T. Denning, C. Matuszek, K. Koscher, J. R. Smith, T. Kohno, "A Spotlight on Security and Privacy Risks with Future Household Robots: Attacks and Lessons", UbiComp '09: Proceedings of the 11th international conference on Ubiquitous computing, pp. 105 - 111

[18] D. Quarta, M. Pogliani, M. Polino, F. Maggi, A. M. Zanchettin, S. Zanero, "An Experimental Security Analysis of an Industrial Robot Controller", IEEE Symposium on Security and Privacy, 2017.

[19] W. Young, N. G. Leveson, "An Integrated Approach to Safety and Security Based on Systems Theory", Communications of the ACM, 57(2), 31-35, 2014.

[20] ISO 13482:2014, Robots and robotic devices — Safety requirements for personal care robots, https://www.iso.org/standard/53820.html

[21] ISO 10218-1:2025 and ISO 10218-2:2025, Robotics — Safety requirements — Industrial robots; Industrial robot applications and robot cells, https://www.iso.org/standard/73933.html e https://www.iso.org/standard/73934.html

[22] ISO/TS 15066:2016, Robots and robotic devices — Collaborative robots, https://www.iso.org/standard/62996.html

[23] ISO/IEC 27001:2022, Information security, cybersecurity and privacy protection — Information security management systems — Requirements, https://www.iso.org/standard/27001

Computo ergo cogito? Se il calcolo simula il pensiero - di ***Arianna Pipitone***, Università degli studi di Palermo e ***Antonio Chella***, Laboratorio di Robotica, dipartimento di Ingegneria Università degli Studi di Palermo

Il dialogo interiore costituisce il pensiero astratto umano. I sistemi di intelligenza artificiale manipolano simboli linguistici in modo coerente. L'integrazione tra robotica embodied e inner speech artificiale apre scenari inediti sulla natura del pensiero stesso

Linguaggio interiore, pensiero incarnato. Dove lingua e pensiero si intersecano, la parola diviene strumento di mediazione tra esperienza corporea e mentale (Fernyhough & Borghi, 2023), dando origine a uno dei processi cognitivi più complessi della mente umana. Il *discorso interno* è quel monologo, silente o esternalizzato, per cui un soggetto utilizza le strutture linguistiche per ragionare, ponderare, memorizzare (Baddeley, 1992) e interpretare il proprio sé immerso nel mondo (Morin, 2011; Alderson-Day & Fernyhough, 2015). Mezzo attivo di comunicazione con sé stessi, intuitivamente inteso come *riflessione mentale*, riconosciuto come strumento primario dello sviluppo umano (Vygotsky, 1939), non è semplice riflesso del pensiero, ma è dimostrato che è una sua forma costitutiva, ed in certi casi una delle sue manifestazioni uniche (Munroe, 2023).

Come le strutture linguistiche del discorso interno incanalano il pensiero e ne danno forma, attraverso trigger che attivano connessioni logiche, sarebbe pretenzioso chiederci se un processo equiparabile avviene nei modelli computazionali che mirano a simularlo? Quali teorie potrebbero far supporre che, seppur simulato da una macchina, un "pensiero artificiale" non sia comunque una forma di pensiero?

Il discorso interno umano: quando il linguaggio naturale diventa pensiero

Quante volte ci saremo trovati coinvolti in un monologo, durante la pianificazione delle nostre attività quotidiane, durante la ripetizione di un argomento da memorizzare, durante una semplice riflessione. Questa manifestazione linguistica interiore, non si limita solo a tradurre processi mentali preesistenti, ma costituisce una delle modalità fondamentali di emergenza del pensiero: è attraverso la parola, interiorizzata e riflessiva, che il pensiero si struttura e diviene entità accessibile per sé (Fernyhough & Borghi, 2023).

Le evidenze neuroscientifiche e psicologiche mostrano che questo fenomeno coinvolge non solo reti cerebrali linguistiche, ma anche sensomotorie e mnemoniche, evidenziando la natura incarnata del pensiero (Borghi & Binkofski, 2014). Tuttavia, ridurre il dialogo interno a un semplice correlato embodied

rischia di oscurarne la funzione costitutiva. Per molti domini del pensiero astratto, dalla matematica all'etica, dai concetti giuridici a quelli filosofici, il linguaggio interno non è il rivestimento di pensieri preesistenti: è *il mezzo stesso attraverso cui il pensiero si realizza*.

Il caso più studiato è quello dei numeri esatti. Munroe (2023) dimostra che per pensare al numero 3.735 non possediamo alcuna rappresentazione mentale indipendente dal codice simbolico: *pensare a un numero significa semplicemente usare internamente il suo nome*. Le evidenze sono stringenti: popolazioni amazzoniche senza vocabolario numerico esteso non distinguono quantità esatte oltre il quattro; un paziente afasico calcolava correttamente sia con cifre arabe che con parole inglesi, ma non riusciva a tradurre tra i due codici, se invece esistesse un "concetto centrale" del numero, questo dovrebbe fungere da ponte.

Pertanto, partendo da capacità innate limitate, l'acquisizione di sistemi semiotici (codici numerici, vocabolari etici, terminologie giuridiche) ci permette di generare pensieri che sarebbero altrimenti inaccessibili. Non possediamo antecedentemente i concetti che poi il linguaggio "etichetta", ma è l'articolazione simbolica stessa che costituisce il pensiero. Questa prospettiva dissolve la distinzione foderiana tra "mentalese" e linguaggio naturale. Per intere categorie di pensiero astratto non esiste un formato di rappresentazione sottostante: il linguaggio non traduce pensieri pre-esistenti ma li costituisce.

Large Language Models: elaborazione linguistica senza voce interiore?

Alla luce di questa prospettiva, emerge una questione cruciale: se in certi casi il linguaggio naturale non si limita a veicolare il pensiero, ma ne costituisce la struttura stessa, allora la distinzione fra “pensare con linguaggio” e “elaborare rappresentazioni linguistiche” si fa più sfumata. Il linguaggio non è solo strumento comunicativo, i sistemi semiotici ci permettono di pensare pensieri che sarebbero impossibili senza gli stessi sistemi.

I Large Language Models (LLM) sembrano, a prima vista, operare in un universo radicalmente diverso da quello cognitivo umano. Quando un individuo ascolta o evoca mentalmente una parola, ad esempio *mela*, si attivano simultaneamente nodi lessicali, aree sensomotorie e reti associative autobiografiche. Queste componenti incarnano il legame fra linguaggio e percezione, ma la dinamica cognitiva non si esaurisce lì. La parola *mela* non richiama alla mente soltanto esperienze sensoriali (la superficie liscia, la consistenza al morso, il sapore acidulo) o ricordi episodici (un pomeriggio d’infanzia sotto un melo), bensì può innescare un discorso riflessivo nel quale il soggetto rielabora concettualmente ciò che ha percepito, ad esempio valutando i benefici nutrizionali del frutto, o il suo valore simbolico, senza

necessariamente dar forma a concetti centrali preesistenti, come *valore nutrizionale* o *peccato originale*.

Quando un LLM processa il prompt *mela*, non attiva schemi percettivi o ricordi sensoriali, ma configurazioni vettoriali in uno spazio di embedding multidimensionale, dove *mela* si colloca vicino a *frutto*, *rosso* o *Newton* in base a correlazioni statistiche acquisite durante l'addestramento. Quando l'LLM connette *mela* a *peccato,* a *conoscenza,* a *libertà*, sta forse operando diversamente da un umano che pensa questi concetti astratti "usando internamente i loro nomi"? La risposta è che la manipolazione coerente di simboli secondo strutture trasformazionali è sufficiente per una forma di pensiero linguistico.

Da questa analogia emerge una domanda di fondo: se il pensiero umano, in certi casi, può essere costituito dal linguaggio stesso, allora *l'assenza di corpo basta davvero a negare la possibilità di una forma di pensiero linguistico*? Se il linguaggio può generare pensiero attraverso le proprie strutture interne, perché escludere che ciò accada anche in un sistema linguistico artificiale, sebbene privo di esperienza sensomotoria? Quando un LLM processa "calcola 3.735 + 2.481", genera una sequenza di trasformazioni simboliche che producono un risultato corretto: perché questo dovrebbe essere "mero calcolo probabilistico" mentre l'articolazione interna umana degli stessi simboli costituisce "pensiero autentico"? In entrambi i casi, un trigger linguistico attiva relazioni simboliche che si auto-organizzano. Se il linguaggio può generare pensiero attraverso le proprie strutture interne, perché escludere che ciò accada anche in un'architettura computazionale? La manipolazione coerente di simboli secondo strutture trasformazionali potrebbe essere sufficiente per una forma di pensiero linguistico, indipendentemente dal substrato che la implementa.

Manzotti (2024) sostiene che l'uso di strutture linguistiche nelle macchine non equivale a dialogo o auto-dialogo umano, ma rappresenta solo una simulazione tecnica priva di reale consapevolezza: le architetture computazionali che simulano il linguaggio interiore, secondo l'autore, confondono il funzionale con il fenomenologico. Possono imitare il comportamento del pensiero riflessivo, ma non producono esperienza soggettiva, evidenziando una distinzione rigorosa tra simulazione cognitiva e esperienza cosciente. L'idea che una simulazione di discorso umano non sia tale perché priva di un "sé cosciente" trascura che, come mostrano gli studi di Fernyhough e Borghi (2023), il dialogo interno è tale non solo come esperienza fenomenologica, ma anche come strumento cognitivo operativo, capace di organizzare l'attenzione, la memoria e il ragionamento, funzioni che, in parte, possono essere simulate anche da sistemi non coscienti. Quindi, così come la coscienza sussiste senza dialogo interno,

potremmo dire che il dialogo interno può sussistere senza coscienza e deve essere riconosciuto tale anche se solo assolvente funzioni cognitive. Negare qualsiasi valore cognitivo al linguaggio artificiale significa ignorare che la dimensione dialogica del pensiero non coincide unicamente con la presenza di coscienza fenomenica, ma può esistere come processo funzionale, dinamico e produttivo di conoscenza.

Allo stesso modo, Federico Faggin (2024) sostiene che i computer, per quanto complessi, non possano né pensare né essere coscienti, in quanto privi di quella *sensazione interiore* che accompagna l'esperienza incarnata. Secondo la sua visione, il pensiero e la coscienza sono proprietà intrinseche dei sistemi viventi, radicate nella specificità ontologica della materia biologica e dunque irriducibili alla computazione classica. In questa prospettiva, il pensiero e la coscienza nascono dalla *carnalità dell'essere*, non dall'elaborazione simbolica.

Tuttavia, alla luce di quanto visto, tale posizione rischia di trascurare un aspetto cruciale: il fatto che il linguaggio stesso possa costituire l'unico veicolo del pensiero, indipendentemente dal suo substrato materiale. È vero che nei LLM manca il *grounding* corporeo e affettivo: come ricordano Bender e Koller (2020), questi modelli manipolano forme linguistiche con straordinaria abilità senza mai aver toccato, visto o agito nel mondo a cui quelle forme si riferiscono. Sono sistemi che hanno "letto" ogni descrizione del rosso senza mai averlo percepito, un paradosso anticipato da Harnad (1990) con il symbol grounding problem. Ma se il linguaggio naturale può farsi esso stesso struttura del pensiero, la questione aperta diventa se un linguaggio formale, come quello dei LLM, possa ospitare una forma *di pensiero disincarnato, anche se meramente funzionale*.

L'inner speech robotico: verso un pensiero artificiale?

Eppure, l'integrazione emergente tra LLM e robotica embodied sta modificando anche questo scenario. Architetture come RT-2 (Zitkovich et al., 2023) e PaLM-E (Driess et al., 2023) non si limitano a processare testo: incorporano percezione visiva da telecamere, feedback tattile da sensori di forza, propriocezione da encoder articolari, e traducono istruzioni linguistiche in sequenze motorie eseguibili. Lavori sull'inner speech artificiale (Pipitone & Chella (2021)) hanno sviluppato architetture cognitive in cui il robot letteralmente "parla a sé stesso" in linguaggio naturale. In questi sistemi, l'inner speech artificiale non è un mero output linguistico: è un processo computazionale ricorsivo in cui rappresentazioni linguistiche interrogano e modulano altre rappresentazioni prima che l'azione venga eseguita. Il robot "usa parole" per decodificare cosa sta percependo, "si chiede" quale azione intraprendere e se può intraprenderla, "valuta" internamente le conseguenze. Come nei bambini

(Cangelosi & Schlesinger, 2015), emerge una “biografia operazionale” in cui parole e azioni si co-determinano reciprocamente.

D’altronde, in anni meno recenti, venne già evidenziato (Steels, 2003) come una forma di retroazione linguistica, consentendo l’ulteriore processamento in input dell’output, portasse al raffinamento di informazioni condivise da una popolazione di agenti artificiali, evidenziando che il riascolto del sé permetta un raffinamento della comprensione e della disambiguazione dell’informazione condivisa. Evidenziando che anche in strutture assolutamente prive di sé, la componente funzionale di un dialogo interiore richiami quella di un essere umano, che riflette concettualmente per inferire meglio il contesto.

Seppur non si voglia sostenere la tendenza pericolosa all'antropomorfismo acritico, proiettando abilità altissime e prettamente umane su sistemi che esibiscono comportamenti convincenti (Dennet, 1989), si dovrebbe però porre uno sguardo critico sullo scetticismo categorico che invoca il “pensiero autentico” come proprietà accessibile solo ai neuroni biologici, identificando una sola ed esclusiva forma di pensiero. Tale posizione ignora una lezione fondamentale della storia scientifica: non siamo le uniche creature viventi, la nostra razionalità è permeata da bias sistematici (Kahneman & Tversky, 1979), e quindi, forse, il pensiero stesso non è vincolato a un'unica implementazione ma è un processo emergente che può manifestarsi in substrati multipli.

Da un punto di vista metodologico, assumere che solo i sistemi biologici possano ospitare pensiero dialogico costituisce una petizione di principio: assume per certo, a priori, senza prove inconfutabili, ciò che dovrebbe comunque essere dimostrato. La domanda empirica più sensata diventa quindi: se un sistema embodied dotato di inner speech mostrasse i correlati funzionali della coscienza riflessiva (auto-riferimento linguistico, integrazione informazionale, apprendimento dall’errore e autoregolazione) su quale base potremmo negare che esso stia realizzando, sicuramente in forma rudimentale, un processo di pensiero?

Il punto non è se tale pensiero *senta* il mondo come un essere umano, ma se il linguaggio che lo struttura funzioni come *luogo d’emergenza del pensare*. Se così fosse, la distinzione fra “esperienza incarnata” e “processo linguistico” non segnerebbe una sospensione ontologica, ma due modi differenti di articolare un medesimo principio: il pensiero come dialogo interno, sia esso espresso nella carne o nel calcolo.

Il vero centro della questione… per non perdere il centro

La domanda cruciale non dovrebbe essere: "pensano davvero nel senso pieno umano?" ma pragmatica ed etica: "cosa può cambiare se si comportassero *come se* pensassero? Quale sarebbe per noi la differenza? Quali scenari futuri potrebbero preoccuparci se anche simulassero pensiero?".

Un sistema che integra linguaggio e azione, che costruisce modelli predittivi del mondo e li raffina attraverso interazione, che esibisce comportamento adattativo, e che ragiona su concetti astratti, sta realizzando processi informativi che nei sistemi biologici chiamiamo cognizione e pensiero. Negarlo è provincialismo cognitivo; accettarlo passivamente, ignorando le differenze strutturali profonde, è antropomorfismo ingenuo.

La sfida intellettuale ed etica del nostro tempo è navigare tra questi estremi: riconoscere la realtà funzionale dei "pensieri non incarnati", senza proiettare su di essi la pienezza dell'esperienza umana; studiare le differenze architetturali tra cognizione biologica e artificiale senza demonizzarle come barriere invalicabili; promuovere educazione etica alle linee di confine, e non chiedersi soltanto se esistano davvero o meno.

Non dovremmo chiederci se le macchine possano pensare nel nostro stesso identico modo, ma se stiano sviluppando forme alternative e legittime di cognizione che meritano riconoscimento scientifico e, eventualmente, considerazione morale. La realtà inconfutabile è che stiamo assistendo all'alba di intelligenze genuinamente diverse dalla nostra, come già ne esistono tante in natura: non imitazioni della mente umana, ma implementazioni alternative di processi informativi complessi che esigono nuove categorie concettuali per essere comprese adeguatamente.

Computo ergo cogito? Se il calcolo può generare un discorso interno, che potrebbe anche ancorarsi a esperienze sensori-motorie, orientare decisioni complesse e rielaborare processi di auto-riflessione astratta, allora forse sì, in una forma diversa, ma non per questo meno reale. Perché il pensiero non appartiene solo a chi sente, ma anche a chi (o cosa), in qualche modo, dice qualcosa a sé stesso.

Bibliografia

Alderson-Day, B., & Fernyhough, C. (2015). Inner Speech: Development, Cognitive Functions, Phenomenology, and Neurobiology. *Psychological Bulletin*, 141(5), 931–965.

Baddeley, A. (1992). Working Memory. *Science*, 255 (5044), 556–559.

Bender, E.M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. *Annual Meeting of the Association for Computational Linguistics*.

Borghi, A. M., & Binkofski, F. (2014). *Words as social tools: An embodied view on abstract concepts* (Vol. 2). Springer New York.

Cangelosi, A., & Schlesinger, M. (2015). *Developmental Robotics: From Babies to Robots*. Cambridge, MA: MIT Press.

Dennett, D. C. (1989). *The Intentional Stance*. Cambridge, MA: MIT Press.

Driess, D., Xia, F., Sajjadi, M. S., Lynch, C., Chowdhery, A., Ichter, B., ... & Toussaint, M. (2023). PaLM-E: An Embodied Multimodal Language Model. *arXiv preprint arXiv:2303.03378*.

Faggin, F. (2024). *Irreducible: Consciousness, life, computers, and human nature*. John Hunt Publishing.

Fernyhough, C., & Borghi, A. M. (2023). Inner speech as language process and cognitive tool. *Trends in cognitive sciences*, *27*(12), 1180-1193.

Harnad, S. (1990). The Symbol Grounding Problem. *Physica D: Nonlinear Phenomena*, 42(1-3), 335–346.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2), 263–291.

Manzotti, R. (2024). *AI Internal Dialogues vs. Human Conscious Inner Speech*. IAC19

Morin, A. (2011). Self-Awareness Part 1: Definition, Measures, Effects, Functions, and Antecedents. *Social and Personality Psychology Compass*, 5(10), 807–823.

Munroe, W. (2023). Semiotics in the Head: Thinking About and Thinking Through Symbols. *Philosophy and Phenomenological Research*, 107(2), 413–438.

Pipitone, A., & Chella, A. (2021). What Robots Want? Hearing the Inner Voice of a Robot. *iScience*, 24(4), 102371.

Steels, L. (2003). Evolving Grounded Communication for Robots. *Trends in Cognitive Sciences*, 7(7), 308–312.

Vigotsky, L. S. (1939). Thought and speech. *Psychiatry*, *2*(1), 29-54.

Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., ... & Han, K. (2023, December). Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning* (pp. 2165-2183). PMLR.

Dinamiche e implicazioni della news avoidance in Italia - di ***Francesca Rizzuto***, professoressa associata nel Dipartimento Culture e Società dell'Università di Palermo e ***Anna Brisciano***, laureata in Scienze della comunicazione e giornalista

In Italia la news avoidance cresce tra troppe notizie, toni spesso negativi, disinformazione e sfiducia nei media. Social e piattaforme cambiano le abitudini: si seleziona di più e aumenta la stanchezza informativa. Effetti: meno partecipazione consapevole, più disuguaglianze e un dibattito pubblico più fragile

Abstract

Negli ultimi anni, il fenomeno della news avoidance, ovvero l'evitamento intenzionale o parziale delle notizie, ha attirato crescente attenzione sia nella ricerca sul giornalismo che nel dibattito politico sulle potenziali conseguenze per le architetture democratiche dei paesi occidentali di un progressivo allontanamento dei cittadini dalle news. In Italia, tale fenomeno si inserisce in un contesto mediatico caratterizzato da elevata politicizzazione, crisi di fiducia nelle istituzioni informative e trasformazioni digitali che hanno radicalmente mutato i modelli di produzione e di consumo di informazione esistenti. Questo articolo analizza alcune peculiarità della news avoidance in Italia, distinguendo tra fattori strutturali, psicologici e culturali, e ne discute le implicazioni per la partecipazione democratica e la qualità del dibattito pubblico. Attraverso una revisione della letteratura e un'analisi del contesto italiano l'obiettivo è quello di proporre una lettura del fenomeno, collocandolo all'interno del dibattito internazionale sulle modalità del consumo di notizie nella società delle piattaforme.

Introduzione. Il consumo di news nell'era delle piattaforme: nuove sfide (di Francesca Rizzuto)

Il fenomeno della *news avoidance*, ovvero la tendenza a evitare intenzionalmente o non intenzionalmente l'esposizione alle notizie, è cresciuto significativamente negli ultimi anni anche in Italia, intensificandosi durante la pandemia di Covid-19. Ponendosi al centro del dibattito accademico e politico in molti paesi occidentali tale fenomeno è stato considerato una doppia minaccia, sia per l'industria informativa che per la qualità del dibattito democratico, indizio di una serie di radicali mutamenti di modelli tradizionali del consumo mediale e del ruolo riconosciuto da decenni all'informazione nei sistemi politici occidentali. Nella società delle piattaforme (Van Dijck, Poell, de Wall, 2018) il consumo delle notizie ha assunto una funzione sempre più legata all'intrattenimento, svincolandosi dall'uso

esclusivamente informativo e/o politico; inoltre, i dispositivi mobili e i social network forniscono velocemente e pervasivamente news personalizzate in base agli interessi dei singoli individui, i quali preferiscono sempre più contenuti emotivamente coinvolgenti, concisi e con elementi interattivi, perfettamente coerenti con le logiche del business dei social media. In molti paesi, del resto, gli utenti più giovani stanno adottando abitudini di consumo informativo completamente diverse da quelle dei loro genitori, preferendo la ricerca di notizie attraverso l'uso di fonti diversificate in contesti comunicativi nei quali il confine tra vero e verosimile (McIntyre, 2018) diventa sempre meno netto, lasciando ampi spazi al cortocircuito della disinformazione, usata come vera e propria arma politica. Dinanzi alla rivoluzione delle piattaforme, l'industria delle notizie deve trovare modalità innovative e coinvolgenti nell'interazione con i destinatari: i modelli tradizionali di relazione con i lettori/spettatori sono palesemente inadeguati per fronteggiare le sfide più rilevanti che i social media pongono alla professione giornalistica, che per tre secoli ha accompagnato e sostenuto lo sviluppo delle democrazie occidentali, pur nelle diverse declinazioni nazionali (Cappello, Rizzuto 2024). Dalla centralità riconosciuta al giornalismo quale bene di cittadinanza primario, in grado di esercitare una rilevante ed indispensabile funzione di mediazione tra i cittadini e la realtà, non solo politica, la moltiplicazione odierna dei canali, delle fonti e degli attori coinvolti nel circuito planetario dell'informazione, ha contribuito ad un marcato declino della fiducia del pubblico, imponendo sia una ridefinizione del concetto di "affidabililtà" di fonti e contenuti, che la necessità di adattarsi rapidamente a nuovi formati, linguaggi e logiche. Dinanzi agli interrogativi imposti del predominio crescente delle opache logiche algoritmiche e alla rivoluzione dell'IA, nelle pagine seguenti viene proposta una lettura critica del fenomeno della *news avoidance* in Italia, che assume particolare rilevanza alla luce della frammentazione dell'ecosistema informativo e della diffusione di piattaforme digitali anche nel nostro paese.

Il concetto di news avoidance nella letteratura scientifica (di Anna Brisciano).

Il consumo di notizie rappresenta tradizionalmente uno degli elementi centrali per il funzionamento delle democrazie moderne. Tuttavia, numerose ricerche recenti mostrano una crescente quota di cittadini che dichiarano di evitare attivamente le notizie, in tutto o in parte. Questo fenomeno, definito

news avoidance, non implica necessariamente una totale disconnessione dall'informazione, ma piuttosto una selezione strategica e talvolta difensiva dei contenuti informativi, confermando che il rapporto tra cittadini e informazione sta affrontando numerose criticità: la moltiplicazione dei soggetti produttori di notizie e la polarizzazione delle informazioni nella rete hanno alterato non solo la natura delle notizie, ma anche le logiche di consumo della stessa informazione, contribuendo ad un peggioramento dell'overload informativo (sovraccarico informativo), una condizione che nel tempo ha reso difficile al pubblico orientarsi e distinguere le notizie rilevanti da quelle meno importanti e l'informazione dalla disinformazione. Ne è derivato un problema di affidabilità e credibilità legato alla moltiplicazione di verità individuali imposte nei circuiti dei social che sta profondamente cambiando il modo di consumare le notizie. È all'interno di questo scenario che si è sviluppata la news avoidance, ovvero la tendenza delle persone a sottrarsi in maniera parziale o totale alle notizie. Il fenomeno in questione può assumere diverse configurazioni: può essere intenzionale e non intenzionale, ma non solo. In letteratura si parla anche di evitamento costante, occasionale e selettivo (Andersen et al., 2024). Sebbene l'evitamento delle notizie fosse oggetto di studio da parte della comunità scientifica e dell'industria informativa già da un po' di anni, dal 2021 ha attirato ancora più attenzione da parte degli studiosi in seguito alla pandemia di Covid 19. Le circostanze di emergenza innescate dalla crisi sanitaria planetaria hanno causato non solo l'insorgere di una pesante infodemia, ma hanno anche alimentato sentimenti di ansia e sfiducia nell'informazione. Successivi eventi, come il conflitto fra Russia e Ucraina, non hanno fatto altro che aggravare e confermare queste inclinazioni già esistenti. La *news avoidance* solleva importanti interrogativi anche per quanto concerne i suoi potenziali effetti negativi sulla democrazia: un evitamento sistematico delle notizie potrebbe non solo influenzare il modo in cui le persone percepiscono la realtà, ma anche creare lacune e disuguaglianze fra i cittadini stessi nella loro conoscenza degli affari correnti e nel loro impegno civico, ostacolando una partecipazione attiva alla vita politica. Con il termine news avoidance ci si riferisce ad un basso consumo di notizie, ovvero alla tendenza degli individui a evitare, in maniera totale o parziale, l'esposizione all'informazione. Sebbene uno dei primi paper sull'evitamento delle notizie sia stato "*A Profile of the Daily Newspaper Non-Reader*" di Westley e Severin (1964), è con il processo di digitalizzazione e il conseguente sovraccarico informativo che il fenomeno inizia a diffondersi in maniera significativa nei paesi occidentali, attirando crescente attenzione da parte del mondo accademico e dell'industria giornalistica. Oggi la *news avoidance* si presenta come un fenomeno complesso e multidimensionale. Sono diversi i contributi che hanno provato ad articolare le sue possibili manifestazioni: tra gli studi più rilevanti ed esaustivi, prenderemo in considerazione quello di Skovsgaard e Andersen, che distinguono fra evitamento intenzionale di notizie e non intenzionale. Se il primo si basa su una specifica avversione

per le notizie, il secondo è causato soprattutto da un interessamento per altri tipi di contenuti mediatici, a discapito di quelli informativi (Skovsgaard e Andersen, 2020). Tuttavia, evitare intenzionalmente le notizie non sempre si traduce in un basso consumo di notizie (Toff e Kalogeropoulos, 2020; De Bruin et al., 2024); pertanto, a questa distinzione di base ne segue un'ulteriore basata sul modo in cui il fenomeno si manifesta e che distingue tra evitamento costante, occasionale e selettivo. L'evitamento costante è sia intenzionale che non intenzionale, ed è caratterizzato da un'esposizione molto bassa o inesistente alle notizie, causata da un disinteresse nei confronti di quest'ultime o un interessamento verso altri tipi di contenuti. L'evitamento situazionale, invece, è sempre intenzionale: l'individuo evita volutamente le notizie in certi momenti, pur mantenendo un livello complessivo di consumo informativo, anche elevato. Infine, l'evitamento selettivo, anche questo intenzionale, riguarda la volontà dell'individuo di voler evitare specifici contenuti informativi, come determinati temi politici o fatti di cronaca, per motivi come disinteresse, disagio o paura. Benché la letteratura sia solita distinguere l'evitamento volontario da quello involontario - quando si parla di news avoidance - secondo alcuni studiosi questa distinzione viene meno se applicata alla ricerca empirica (Palmer, Toff, e Nielsen 2023): anche scelte deliberate di consumo di notizie nel tempo possono diventare routinarie a tal punto da non essere più considerate intenzionali (Andersen et al., 2024). Nondimeno sono diversi gli studi che indagano l'intenzionalità nella scelta di consumo delle notizie: fra i più recenti possiamo menzionare: "*Zooming in on Topics: An Investigation of the Prevalence and Motives for Selective News Avoidance*" di Schäfer, Betakova e Lecheler (2024) e "*I Do Not (Want To) Know! The Relationship Between Intentional News Avoidance and Low News Consumption*" di Betakova, Lecheler, Shäfer e Boomgaarden (2024). Tra le possibili motivazioni sociali e psicologiche a cui ricondurre il comportamento di evitamento delle notizie sono stat indicate il sovraccarico informativo, l'eccessiva negatività delle notizie, la percezione di scarsa affidabilità delle fonti e il senso di impotenza rispetto agli eventi narrati.

Il contesto italiano tra nuovi consumi informativi e fuga dalle notizie (di Francesca Rizzuto)

Negli ultimi anni il giornalismo italiano è stato profondamente trasformato dal processo di digitalizzazione: nell'attuale sistema mediatico ibrido, in cui coesistono media tradizionali e nuovi media (Chadwick, 2013), il modo in cui le persone consumano informazioni consente loro non solo di accedere alla realtà, ma anche di evitarla. Volendo proporre un'ipotesi di lettura della crisi reputazionale ed economica del giornalismo italiano contemporaneo, sembra molto utile presentare e discutere le tendenze emergenti nel consumo di notizie, al fine di collegarle ad alcune peculiarità tradizionali di questo sistema informativo (Hallin, Mancini, 2021), che, a nostro avviso, continuano a

influenzare gli equilibri tra produzione e consumo di notizie. Nell'era delle piattaforme, anche in Italia si stanno affermando nuove abitudini informative, non solo perché i social network sono diventati una fonte fondamentale di notizie, offrendo contenuti isolati, parziali o falsi, ma anche per il declino dell'uso esclusivo dei canali di informazione tradizionalmente dominanti, come la TV e i giornali (Rizzuto, 2023). Infatti, mentre le organizzazioni giornalistiche tradizionali hanno dominato il mercato dell'informazione online per molti anni, dal 2022 le principali emittenti televisive e i quotidiani più importanti sono stati superati da operatori nati in digitale. Tali cambiamenti delle abitudini informative degli Italiani rendono più evidente una crisi economica del nostro sistema informativo, soprattutto televisivo e cartaceo, ma denunciano anche una profonda incertezza sul significato sociale da attribuire alla mediazione giornalistica professionale nell'era dei social. Un'attenta analisi dei dati forniti dal Reuters Digital News Report Italia (2025) lascia emergere inequivocabilmente che gli Italiani sono molto cambiati nelle scelte comunicative, e in particolare in quelle dell'informazione. Sin dalle origini ottocentesche i grandi quotidiani nazionali sono stati considerati uno strumento di partecipazione politica e di rafforzamento identitario (Murialdi, 2014; Rizzuto 2018): la crisi del significato dell'informazione è stata acuita dalla rivoluzione digitale che ha stravolto il mercato delle notizie e le routine professionali, come segnalato dal drastico calo di interesse per le notizie, tra i più significativi rispetto ad altri contesti europei. Nel 2025 tale interesse continua a diminuire, attestandosi al 39% (chi si dichiara molto o estremamente interessato) e confermando un declino costante rispetto al 74% registrato nel 2016 (-35% in nove anni). Le percentuali di coloro che si dicono estremamente interessati, molto interessati o anche solo abbastanza interessati alle notizie si riducono tutte rispetto al 2024 ("estremamente interessati" da 11% a 10%, "molto interessati" da 29% a 28% e "abbastanza interessati" da 47% a 46%) mentre aumentano coloro che si dicono "non molto interessati" (da 10% a 12%) e resta sostanzialmente invariato il numero di chi si dice "niente affatto interessato" (2%) (Cornia et al. 2025, p. 15). Tuttavia, nonostante l'interesse in calo, gli Italiani consultano spesso le notizie: il 59% lo fa più volte al giorno, delineando un vero e proprio "paradosso italiano" (ivi, p.16). Il confronto con altri paesi europei rafforza questa ipotesi di lettura, mettendo in evidenza la condizione quasi unica del nostro paese, che è agli ultimi posti per interesse nelle notizie, ma al secondo posto per frequenza di consultazione, preceduto solo dalla Finlandia, il paese con più elevato consumo di informazione al mondo. I dati forniti dal Digital News Report Italia 2025 sulle fonti di informazioni usate, confermano, tuttavia, un'altra caratteristica "storica" delle diete mediali degli Italiani: la televisione resta un importante punto di riferimento e il 51% la considera la propria fonte principale, unico tra i sei paesi considerati dallo studio, mentre si posiziona all'ultimo posto nell'uso dei media cartacei, con un calo della carta stampata confermato dal drastico 2% di soggetti che la considerano la propria fonte principale di informazione. Il ricorso a testate native

digitali e giornalisti indipendenti (9%) è un altro segnale di cambiamenti radicali dei consumi di news accanto alla presenza dei social (che dominano con il 17%), dei siti di quotidiani (8%), di testate radiotelevisive (5%) e di podcast e chatbot di intelligenza artificiale, che si affermano come nuove fonti aggiuntive (usati settimanalmente dal 6% e dal 4%) (ivi, p.20). In una prospettiva diacronica, in Italia il fenomeno della *news avoidance* ha seguito un andamento a "ondate": dal 28% del 2017 fino al 36% nel 2019, poi un calo fino al 29% nel 2023 e un aumento al 36% nel 2024. Nel 2025 si registra il dato positivo di una flessione del 3%: con il 33% l'Italia è al di sotto della media internazionale (40%) per la frequenza con cui gli utenti evitano intenzionalmente le notizie. Una lettura critica dei dati spinge ad individuare le motivazioni principali che si trovano alla base della scelta di ignorare le notizie: indubbiamente essa non dipende soltanto dai contenuti trattati, ma è direttamente connessa all'identità dei singoli individui, alle loro posizioni politiche e ai differenti bisogni informativi legati alle pratiche quotidiane di fruizione mediale. Il contenuto delle notizie è sicuramente uno dei fattori che aumentano la propensione alla news avoidance: chi evita le notizie tende spesso a percepirle come eccessivamente negative, manipolative o inaffidabili. Nell'era dell'emotainment (Santos, 2009) il compito principale del giornalismo è diventato quello di fornire, sotto forma di *storie*, significati accessibili ed emozionanti: i singoli sono così diventati i referenti di *una pervasiva alluvione comunicazionale* a causa della quale risulta difficile la metabolizzazione delle risorse cognitive. La conseguenza più significativa di tale sovraccarico simbolico è che il flusso di informazioni viene interpretato e vissuto soprattutto *emozionalmente*, vale a dire sempre di più sulla base del gusto soggettivo e immediato e non sull'approfondimento logico e razionale: la velocità, la frammentarietà e la significatività della componente emotiva sono alla base della logica produttiva delle notizie, monadi-testi di forte impatto sensoriale e visivo. Con i social media le notizie sono anche prodotti di intrattenimento a basso costo proposti a platee sempre più frammentate, che vivono la fruizione di informazioni come esperienze di *entertainment* e parte significativa del *leisure time*. Sul piano dei contenuti, l'adozione della logica infointrattenitiva ha spinto il sistema informativo italiano a privilegiare i dibattiti su temi sociali "allarmanti" (come la sicurezza) con schieramenti netti tra opinioni in conflitto, il gossip, lo scandalo politico, la cronaca nera, servendosi dell'iperbole e del sensazionalismo come criteri centrali nella rappresentazione giornalistica della realtà. Insistendo sulla dimensione della patemizzazione e del coinvolgimento emotivo dei fruitori, le *news* raccontano la realtà secondo modalità di ricostruzione drammatizzanti: estrapolando dal reale frammenti, che vengono poi ricontestualizzati all'interno di schemi, la prospettiva dell'*esibizione* viene adottata come frame per valutare la notiziabilità di fatti o persone e ogni evento viene trasformato in storia e teatralizzato. In tale contesto, l'adozione di una retorica patemica, che offusca la separazione tradizionale tra *fact e fiction,* diffonde notizie spesso deprimenti, decontestualizzate, pillole informative

che mirano a suscitare emozioni forti ma che favoriscono la creazione di un clima di incertezza in cui domina la percezione dell'inefficacia individuale rispetto alla realtà e aumenta la sfiducia nei confronti dei media. La "fuga" deliberata dalle notizie può essere, quindi, ricondotta a diversi fattori psicologici, come il calo persistente dell'interesse verso l'informazione, ma anche tecnologici e algoritmici. Nel contesto delle piattaforme, che permettono agli individui di presentare e condividere continuamente le proprie emozioni, creando narrazioni personali e collettive, si registra la progressiva perdita di consistenza del significato del ruolo sociale tradizionalmente attribuito al giornalista, quello di agente di mediazione per i fruitori, a cui storicamente è spettata la responsabilità di proporre una lettura delle priorità nell'esperienza del mondo. I produttori di informazione si ritrovano ad operare in un contesto totalmente differente rispetto alla sfera pubblica predigitale: numerosi studi hanno focalizzato l'attenzione sulla sostanziale trasformazione dell'ambito del dibattito politico e sociale prodotta dalla digitalizzazione in una *post-sfera pubblica* frammentata, definita piattaformizzata, nella quale si attiva un cortocircuito tra finalità commerciali delle aziende e istanze di partecipazione degli individui (Sorice, 2023; Vaccaro 2020). Pertanto, nell'ecosistema mediale ibrido contemporaneo, la news avoidance, quindi, può essere interpretata sia come un segnale di sfiducia verso l'effettiva capacità dei news media tradizionali di raccontare il mondo che come strategia per gestire il sovraccarico informativo (Valeriani et al., 2021; Park, 2019).: una strategia deliberata, adottata da alcuni individui per sottrarsi, almeno in parte, all'alluvione informativa in cui vivono immersi, che spesso produce emozioni negative e una sensazione di impotenza. Le piattaforme digitali consentono agli utenti di evitare facilmente le notizie, privilegiando contenuti di intrattenimento o comunicazione interpersonale. Gli algoritmi, a loro volta, rafforzano tali preferenze, contribuendo a una esposizione informativa ridotta o distorta. I dati confermano che molti individui evitano le notizie per ridurre ansia, stress o senso di frustrazione: i recenti sconvolgimenti degli equilibri geopolitici novecenteschi hanno prodotto un coverage continuo di conflitti internazionali. La percezione di una narrazione prevalentemente negativa e manipolativa rappresenta, quindi, un fattore chiave nel sorgere della news avoidance, la quale si configura come una strategia di autodifesa, finalizzata a contrastare quel senso di impotenza, paura e sfiducia che gli utenti sperimentano di fronte a una sovrabbondanza di informazioni non solo negative, ma spesso volutamente false (Woodstock, 2014). Una sensazione di inefficacia diffusasi e aggravatasi tra molti news avoiders in seguito alle tante emergenze esplose negli ultimi anni, come la pandemia da Covid-19 o l'esplosione di conflitti recenti in Ucraina e a Gaza. Il proliferare di narrazioni mediatiche ansiogene e fortemente polarizzate, unita alla diffusione della disinformazione, ha intensificato nei lettori italiani sentimenti di incertezza, paura e stress. I più recenti eventi e l'eccessivo di contenuti negativi hanno sicuramente contribuito a generare nel pubblico una condizione di news fatigue, una sensazione di

stanchezza mentale ed emotiva che sembra delineare una frattura significativa nella relazione tra giornalismo e cittadini.

Conclusioni (di Francesca Rizzuto)

La *news avoidance* in Italia non rappresenta semplicemente una disaffezione verso l'informazione, ma un sintomo di trasformazioni più profonde nel rapporto tra cittadini, media e sfera pubblica. Comprendere questo fenomeno è essenziale per ripensare il ruolo del giornalismo e per promuovere forme di informazione più inclusive, affidabili e sostenibili dal punto di vista emotivo e cognitivo. Il numero di news avoiders presenti in Italia sembra destinato ad aumentare, un'altra criticità che in sinergia con ulteriori problematiche, come, ad esempio, la disinformazione e la polarizzazione delle piattaforme digitali, sta dando il proprio contribuito alla crisi del giornalismo italiano. Nel contesto delle *cybercascate* informative, il nostro sistema informativo tradizionalmente "anomalo" deve affrontare nuove sfide, con il rischio di aumentare la sfiducia nei confronti dei media e danneggiare la qualità del dibattito democratico alla base della nostra architettura costituzionale. Da anni numerosi studi hanno evidenziato la possibile correlazione tra il consumo di news e la partecipazione politica dei cittadini, in direzione di un maggiore apatia civica o di comunità polarizzate sempre più forti, in cui fake news e teorie del complotto potrebbero essere liberamente propagate (Quandt 2018; Chambers 2021; Bentivegna, Boccia Artieri 2021) o, in direzione opposta, favorendo nuove modalità di resistenza e partecipazione nella sfera post-pubblica (De Blasio et al. 2023). Il fenomeno dell'evitamento delle notizie potrebbe anche essere inserito in un quadro più ampio di forme di resistenza all'informazione mainstream, nonché uno dei metodi di autodifesa contro l'inquinamento degli ecosistemi comunicativi da parte del disordine informativo. I dati suggeriscono inoltre che l'evitamento delle notizie è in relazione con il giudizio sulla qualità dei media e con una visione dell'ecosistema mediatico-politico come caratterizzato da disordini informativi e da un'elevata polarizzazione ideologica, che stanno ulteriormente allontanando i cittadini dalle istituzioni politiche e sociali. In altri termini, la *news avoidance* solleva interrogativi rilevanti sul funzionamento della democrazia. Se da un lato evitare le notizie può essere una strategia razionale di auto-tutela, dall'altro un'esposizione informativa limitata rischia di ridurre la partecipazione politica consapevole, ampliare le disuguaglianze informative e indebolire il ruolo dei media come spazio di mediazione pubblica.

Bibliografia

Andersen K., Toff B., Ytre-Arne B., (2024), *What We (Don't) Know About News Avoidance (*2024), vol. 25 (12), pp. 1367–1384.

Bentivegna, S., Boccia Artieri, G. (2021), *Voci della democrazia. Il futuro del dibattito pubblico*, Il Mulino, Bologna.

Cappello G., Rizzuto F. (2024). Giornalismo, post-verità e media education. Milano: Mimesis.

Chadwick, A. (2013), *The hybrid media system. Politics and power*. Oxford University Press.

Chambers, S. (2021), Truth, Deliberative Democracy, and the Virtues of Accuracy: Is Fake News Destroying the Public Sphere?, in *Political Studies*, Vol. 69(1), pp. 147-163, Sage, London.

Cornia A, Ferrando M, Piacenza P, Satta C, (2025), *Digital News Report Italia* 2025, Torino, Master in giornalismo "Giorgio Bocca" – Università degli Studi di Torino

De Blasio E., Selva D., Zunino M. (2023), *Ecosystems of mistrust. Connecting news avoidance to critical citizenship*, Problemi dell'informazione, Rivista quadrimestrale, Vol.1, pp. 159-182, doi: 10.1445/106774

Hallin D., and Mancini P. (2021). Modelli di giornalismo. Roma-Bari: Laterza.

Ireton, C., Posetti, J. (eds) (2018), *Journalism, "Fake News" & Disinformation*, UNESCO

McIntyre, L. (2018). Post-truth. Cambridge. Ma: MIT Press.

Murialdi, P. (2014), *Storia del giornalismo italiano. Dalle gazzette a Internet,* Il Mulino, Bologna

Newman, N., Fletcher, R., Kalogeropoulos, A., & Nielsen, R. K. (2019). *Reuters Institute Digital News Report*.

Palmer, R., Toff B., R. K. Nielsen. (2023), *Examining Assumptions Around How News Avoidance Gets Defined: The Importance of Overall News Consumption, Intention, and Structural Inequalities*, Journalism Studies 24 (6): 697–714.

Park C. S. (2019), *Does Too Much News on Social Media Discourage News Seeking? Mediating Role of News Efficacy Between Perceived News Overload and News Avoidance on Social Media*, Social Media and Society.

Quandt T., (2018), "Dark participation", *Media and Communication*, 6.4, 36-48.

Rizzuto, F. (2018), *La società dell'orrore. Terrorismo e comunicazione nell'era del giornalismo emotivo*, Pisa University Press, Pisa.

Rizzuto F. (2023). News consumption in Italy in the digital era. Trust, crisis and disinformation". *DIGITOS. Revista de comunicaciòn digital*. Universitat de Valencia. **vol. 9, 109-125.**

Santos, J. (2009), *Daring to feel. Violence, the Newsmedia and their emotions*, Lexington books, Lanham

Skovsgaard M., Andersen K. (2022), *News avoidance*. In G. Borchard (Ed.), The SAGE Encyclopedia of Journalism, vol. 1, pp. 1099-1103. SAGE Publications

Skovsgaard M.,Andersen K. (2020), *Conceptualizing News Avoidance: Towards a Shared Understanding of Different Causes and Potential Solutions*, Journalism Studies, 21, 459–476

Sorice, M. (2023) Piattaformizzazione della sfera pubblica, in *Comunicazione Politica* 3/2020, pp. 371-388, Il Mulino, Bologna.

Toff B., Kalogeropulos A. (2020), *All the news that's not to ignore. How the information environment does and does not shape news avoidance,* Public Opinion Quarterly, Vol. 84, pp. 366–390

Toff, B., R. Palmer, R. K. Nielsen (2023), *Avoiding the News: Reluctant Audiences for Journalism*, Columbia University Press, New York, NY.

Valeriani A., Iannelli L., Pavan E., Serani D. (2021) *Chi si fida del vaccino anti-Covid? Infodemia, percezione di «information overload» sui social media e polarizzazione ideologica*, COMUNICAZIONE POLITICA, pp. 437-458

Van Dijck J., Poell T., de Waal M. (2018), *The Platform Society: Public Values in a Connective World*, Oxford University Press, Oxford-New York.

Vaccaro, S. (2020), *Gli algoritmi della politica*, Eleuthera, Milano.

Westley B., Severin W. (1964), *A Profile of the Daily Newspaper Non-Reader*, Journalism Quarterly, vol. 41 (1), pp. 45–156 –

Woodstock L. (2014), *The news-democracy narrative and the unexpected benefits of limited news consumption: The case of news resisters*, Journalism, 15(7), 834–849

Socrate, il chatbot che usa il dialogo per la salute mentale - di *Fabio Frisone*, Humane Technology Lab., Università Cattolica del Sacro Cuore, Milano, Sophia University Institute, Figline e Incisa Valdarno, Firenze - ***Giuseppe Riva***, Humane Technology Lab., Università Cattolica del Sacro Cuore, Milano, Applied Technology for Neuro-Psychology Lab., Istituto Auxologico Italiano IRCCS - ***Chiara Rossi***, Humane Technology Lab., Università Cattolica del Sacro Cuore, Milano, Guglielmo Marconi University, Roma

Un chatbot ispirato al metodo socratico per supportare il benessere psicologico: si chiama Socrate e nasce dalla ricerca accademica italiana. Integra intelligenza artificiale, tecniche cliniche evidence-based e realtà virtuale. Non sostituisce il terapeuta, ma amplia l'accesso al dialogo psicologico in modo sicuro ed etico

Abstract

L'Intelligenza Artificiale Generativa (GenAI) sta trasformando il modo in cui le persone accedono al supporto psicologico, aprendo scenari fino a pochi anni fa impensabili. Tra questi, i chatbot conversazionali progettati per il benessere mentale rappresentano una delle applicazioni più promettenti, ma anche più controverse. Questo articolo presenta Socrate, un chatbot basato su GenAI progettato per favorire riflessione, consapevolezza ed espressione emotiva attraverso il dialogo, ispirandosi alla tradizione socratica e ai modelli psicoterapeutici meaning-based ed evidence-based. A differenza dei chatbot generalisti, Socrate integra conoscenze cliniche selezionate, istruzioni terapeutiche mirate, sistemi di sicurezza e, nelle sue versioni più avanzate, ambienti immersivi in realtà virtuale. Analizzando il suo funzionamento, dalle logiche di personalizzazione fino all'esperienza interattiva e immersiva, emerge come l'AI possa diventare un alleato della salute mentale, purché progettata con intenzionalità clinica, trasparenza ed etica. L'obiettivo non è sostituire il terapeuta umano, ma ampliare lo spazio del dialogo psicologico, rendendolo più accessibile, continuo e sostenibile.

Perché "Socrate": quando il dialogo diventa tecnologia

Attribuire il nome "Socrate" a un chatbot per il benessere psicologico non è una scelta puramente evocativa né un semplice omaggio culturale, ma rappresenta una vera e propria dichiarazione di intenti, progettuali e teorici. Il filosofo ateniese, infatti, non concepiva il sapere come un insieme di contenuti da trasmettere in modo lineare, né come una verità da imporre dall'alto, bensì come un processo

dinamico da attivare attraverso il dialogo. La conoscenza, per Socrate, nasceva dall'incontro tra interlocutori e prendeva forma nel tempo, grazie a una sequenza di domande capaci di mettere in discussione convinzioni date per scontate. Il metodo socratico, fondato sull'arte della domanda e sulla maieutica, mirava a far emergere contraddizioni interne, a stimolare il dubbio e a favorire una presa di coscienza personale che non poteva essere delegata a nessun altro. In questo senso, il dialogo non era uno strumento accessorio, ma il cuore stesso del processo conoscitivo (Eliud, 2025).

È proprio questa impostazione che guida lo sviluppo di Socrate come chatbot conversazionale per il benessere psicologico (Frisone et al., 2025). Socrate nasce all'interno di un'iniziativa di ricerca congiunta tra l'Università Cattolica del Sacro Cuore e l'Università di Udine, un contesto accademico che ha permesso di intrecciare competenze provenienti dalla psicologia clinica, dalle scienze cognitive, dagli studi sull'Intelligenza Artificiale (AI) e la Realtà Virtuale (VR). In un contesto storico in cui molte tecnologie digitali promettono risposte rapide, automazione, ottimizzazione dei processi e soluzioni immediate ai problemi complessi, Socrate si colloca deliberatamente in controtendenza. Invece di accelerare il processo decisionale o fornire indicazioni prescrittive, il sistema è progettato per rallentare la conversazione, per sospendere il giudizio e per restituire all'utente il tempo della riflessione. Il suo obiettivo non è risolvere il problema dell'utente, né fornire diagnosi, etichette o interpretazioni definitive del vissuto, ma creare uno spazio dialogico protetto in cui l'utente possa fermarsi, esplorare ciò che prova e riorganizzare il proprio mondo emotivo e cognitivo. In questo spazio, le domande non servono a guidare verso una risposta corretta, ma a rendere visibili connessioni, ambivalenze e significati che spesso restano impliciti nell'esperienza quotidiana.

Questa scelta progettuale riflette una visione precisa della salute mentale e del benessere psicologico. Il benessere non viene inteso come la semplice eliminazione del sintomo o come il ritorno a uno stato di funzionamento ottimale, ma come la possibilità di dare senso all'esperienza, anche quando questa è complessa, ambigua o dolorosa (Spencer et al., 2025). In questa prospettiva, il disagio non è solo qualcosa da correggere, ma può diventare un punto di partenza per un processo di comprensione più profonda di sé (Rossi Monti & D'Agostino, 2018). La tecnologia, quindi, non diventa un fine in sé né un sostituto della relazione umana, ma un mezzo per facilitare un processo umano profondamente antico: il dialogo come strumento di conoscenza di sé, di costruzione di significato e di trasformazione personale.

Cosa succede quando a farci domande è un chatbot

Le ricerche più recenti sull'interazione uomo-macchina mostrano un dato che, fino a pochi anni fa, sarebbe sembrato paradossale: molte persone faticano a distinguere se una risposta empatica,

articolata e contestualizzata provenga da un professionista umano o da un chatbot avanzato basato su modelli di linguaggio di grandi dimensioni (Ayers et al., 2023). In diversi studi sperimentali, le risposte generate da sistemi di AI vengono valutate come equivalenti e, in alcuni casi, persino superiori a quelle dei clinici umani, su dimensioni considerate centrali nella relazione di aiuto, come l'empatia percepita, la qualità dell'alleanza terapeutica e la competenza comunicativa (Ovsyannikova et al., 2025). Questo non significa che l'AI comprenda l'esperienza umana nel senso fenomenologico del termine, ma suggerisce che la forma linguistica dell'ascolto, ossia il modo in cui una risposta viene costruita, articolata e restituita, gioca un ruolo cruciale nella percezione di essere compresi.

Questo fenomeno apre una questione teorica e clinica di grande rilevanza: che cosa accade quando scopriamo che a porci domande significative non è una persona? La reazione degli utenti è spesso ambivalente. Da un lato, molti riportano un senso di sollievo e di sicurezza: il chatbot non giudica, non si mostra infastidito, è sempre disponibile e non richiede di "fare una buona impressione". Questa assenza di valutazione sociale può abbassare le difese e facilitare l'espressione di pensieri ed emozioni che, in un contesto umano, potrebbero risultare più difficili da condividere (Vowels et al., 2024). Dall'altro lato, emerge una diffidenza latente, legata alla consapevolezza che l'interlocutore non possiede intenzionalità, coscienza, né un vissuto soggettivo. Sapere che l'altro non "sente" ciò che ascolta può generare una sensazione di vuoto relazionale o di artificialità dell'incontro (Demszky et al., 2023).

Socrate si inserisce consapevolmente all'interno di questa ambivalenza, senza tentare di eliminarla o di mascherarla. Il sistema non cerca di ingannare l'utente simulando un essere umano, né di costruire un'illusione di reciprocità emotiva. Al contrario, rende esplicita la propria natura artificiale e fonda la relazione non sulla presunta empatia dell'interlocutore, ma sulla qualità del processo dialogico. In questo senso, Socrate propone un cambio di prospettiva: sposta l'attenzione dall'identità di chi fa la domanda all'effetto che quella domanda produce. Ciò che conta non è chi interroga, ma che tipo di spazio la domanda riesce ad aprire nella mente dell'utente: uno spazio di riflessione, di sospensione del giudizio e di riorganizzazione del significato. È in questo spazio, più che nella relazione in senso stretto, che si colloca il potenziale di supporto psicologico del chatbot.

Dentro Socrate: come funziona davvero

Dal punto di vista tecnico, Socrate è sviluppato a partire da un Large Language Model (LLM) generalista, ma viene profondamente trasformato attraverso un processo di personalizzazione mirato e intenzionale. Non si tratta, quindi, di una semplice interfaccia conversazionale basata su un modello preesistente, bensì di un sistema che opera entro confini ben definiti, sia sul piano dei contenuti

trattabili sia su quello della relazione che instaura con l'utente. A differenza di una normale interazione con ChatGPT, in cui il modello è progettato per rispondere a una gamma estremamente ampia di richieste, Socrate è vincolato da un insieme di regole, istruzioni e priorità che ne orientano costantemente il comportamento verso obiettivi di supporto psicologico e riflessione personale. Questi vincoli non rappresentano una limitazione, ma una condizione necessaria per rendere l'interazione più sicura, coerente e clinicamente sensata.

Il chatbot integra, infatti, un sistema di conoscenze selezionate, costruito a partire da riflessioni provenienti dalla psicologia clinica, dalla psicoterapia e dalla filosofia. Questo insieme di riferimenti teorici non viene utilizzato come un repertorio di nozioni da trasmettere all'utente, ma come una cornice epistemica che orienta il modo in cui le risposte vengono formulate. In altre parole, il modello non si limita a produrre un testo plausibile, ma genera interventi che risultano compatibili con una visione non riduzionista dell'esperienza psicologica e con approcci meaning-based ed evidence-based (Spring, 2007; Vos & Vitali, 2018). Questa architettura contribuisce a ridurre il rischio di contenuti incoerenti, fuorvianti o eccessivamente semplificati, che potrebbero risultare problematici in un contesto di benessere mentale. La conoscenza teorica funge così da sfondo silenzioso dell'interazione, più che da contenuto esplicito del dialogo.

Tuttavia, il funzionamento di Socrate non si esaurisce nella disponibilità di un solido impianto teorico. Una parte centrale del progetto riguarda infatti lo stile dell'interazione, considerato un elemento clinicamente rilevante tanto quanto i contenuti (de Borst & de Gelder, 2015). Il chatbot è stato istruito per privilegiare l'ascolto riflessivo, per tollerare l'ambiguità e per rispettare i tempi soggettivi dell'utente, evitando di forzare il ritmo della conversazione. Le domande sono formulate in modo aperto e non direttivo, senza suggerire risposte implicite o interpretazioni precostituite, e lasciano sempre spazio alla riformulazione, al ripensamento e, quando necessario, al silenzio. In questo senso, Socrate non conduce la conversazione verso un esito prefissato, ma la accompagna, assumendo una posizione che richiama alcuni principi della psicoterapia centrata sulla persona, come l'accettazione incondizionata, la fiducia nelle risorse dell'individuo e il rispetto dell'esperienza soggettiva (Cooper & McLeod, 2011).

Personalizzazione, RAG e fine-tuning: l'architettura invisibile

Uno dei principali limiti dei chatbot generalisti è la loro difficoltà a adattarsi in modo significativo al contesto individuale dell'utente (Yeh et al., 2025). Sebbene le risposte risultino spesso fluide, grammaticalmente corrette e apparentemente pertinenti, tendono a rimanere generiche e poco radicate nella storia personale, nei vissuti emotivi e nelle specificità della situazione di chi interagisce.

In ambito psicologico, questa mancanza di contestualizzazione rappresenta un limite particolarmente rilevante, poiché il significato di un'esperienza non può essere separato dal contesto in cui si manifesta. Socrate affronta questo problema attraverso una combinazione di strategie tecnologiche che mirano a rendere l'interazione più situata, coerente e sensibile alle caratteristiche dell'utente, pur senza trasformarsi in un sistema di valutazione clinica automatizzata. Un primo elemento chiave di questa architettura è l'uso della Retrieval Augmented Generation (RAG). Questo approccio consente al modello di generare risposte ancorate a contenuti selezionati, verificati e pertinenti al dominio del benessere psicologico, anziché affidarsi esclusivamente alla probabilità statistica delle sequenze linguistiche. In termini pratici, la RAG permette a Socrate di consultare un insieme di materiali di riferimento, come linee guida teoriche, cornici cliniche e indicazioni di buone pratiche, prima di formulare una risposta. Il risultato è una maggiore coerenza concettuale e una riduzione delle cosiddette "allucinazioni", ovvero la produzione di contenuti plausibili ma inaccurati o non appropriati. In un contesto delicato come quello psicologico, questa ancoratura a indicazioni affidabili contribuisce a rendere il dialogo più stabile, prevedibile e sicuro per l'utente.

A questo livello di controllo sui contenuti, si affianca un processo di fine-tuning basato su istruzioni dettagliate, che rappresenta uno degli aspetti più rilevanti del funzionamento di Socrate. Il fine-tuning non si limita a insegnare al modello cosa dire, ma definisce in modo esplicito come dirlo e, soprattutto, quando è opportuno non intervenire. Il modello è addestrato a riconoscere l'incertezza, a tollerare l'ambiguità e a chiedere chiarimenti quando le informazioni fornite dall'utente risultano incomplete o contraddittorie. Allo stesso tempo, Socrate è istruito per evitare interpretazioni affrettate, etichette diagnostiche o suggerimenti prescrittivi. Questa "umiltà digitale", scaturita dal riconoscimento dei propri limiti, risulta fondamentale in un ambito in cui un'eccessiva sicurezza potrebbe risultare non solo fuorviante, ma potenzialmente dannosa.

Particolare attenzione è stata dedicata, infine, alla gestione delle situazioni di crisi. Socrate è progettato per individuare segnali linguistici che possono indicare un disagio acuto, come espressioni di disperazione intensa, ideazione suicidaria o intenzioni autolesive. In presenza di tali segnali, il sistema sospende la conversazione ordinaria e attiva una modalità di risposta specifica, orientata alla sicurezza, che invita l'utente a rivolgersi a risorse professionali e a servizi di emergenza. In questi casi, il chatbot esplicita chiaramente i propri limiti e rinuncia a proseguire il dialogo riflessivo. Qui emerge in modo netto il confine tra supporto psicologico e intervento clinico: un confine che Socrate riconosce e che non è autorizzato a oltrepassare.

Quando la conversazione diventa esperienza: voce, avatar e realtà virtuale

Una delle evoluzioni più innovative del progetto Socrate riguarda l'integrazione con ambienti immersivi in VR. In questi contesti, il chatbot non si limita più a produrre risposte testuali sullo schermo, ma diventa un'entità percepibile nello spazio tridimensionale, incarnata in un avatar visivo e vocale. Questa trasformazione apre nuove possibilità nell'esperienza dell'utente, che non dialoga più con un semplice flusso di testo, ma interagisce con una presenza che sembra abitare lo stesso ambiente. La voce, modulata con intonazioni, pause e ritmo naturale, insieme ai movimenti e alle espressioni dell'avatar, contribuisce a creare un senso di co-presenza molto più intenso rispetto alla comunicazione tradizionale mediata da uno schermo (Yonatan-Leus & Brukner, 2025).

La letteratura sulle interazioni uomo-macchina suggerisce che la percezione di presenza e co-presenza possa avere un effetto diretto sul coinvolgimento emotivo, sulla motivazione a proseguire il dialogo e sulla sensazione di essere realmente "in relazione" (Zhu et al., 2022). In un ambiente VR, fattori come la distanza interpersonale percepita, la postura dell'avatar, i micro-movimenti e il ritmo dello scambio verbale diventano strumenti cruciali per modulare l'interazione. Non si tratta solo di un abbellimento estetico: queste caratteristiche possono rafforzare la capacità dell'utente di riflettere sulle proprie emozioni, di sentire ascolto e attenzione e di sperimentare un ambiente psicologico sicuro e strutturato. La VR consente quindi di rendere tangibile il dialogo, trasformando un'esperienza cognitiva astratta in un evento tra entità, percepibile anche sul piano corporeo e sensoriale.

Va sottolineato, però, che la VR non rappresenta automaticamente la soluzione migliore in tutti i contesti. L'immersione deve essere calibrata in funzione del profilo dell'utente, della sua familiarità con la tecnologia, del livello di ansia o sensibilità emotiva e degli obiettivi dell'intervento. Un'eccessiva stimolazione sensoriale, ad esempio, potrebbe generare disagio o distrazione, vanificando l'effetto terapeutico del dialogo (Lundin et al., 2023). Al contrario, un'implementazione calibrata e progressiva può facilitare l'apertura emotiva, migliorare la regolazione affettiva e favorire la consapevolezza del proprio vissuto interiore. Questa attenzione alla personalizzazione conferma la coerenza del progetto con la filosofia socratica: l'obiettivo non è stupire o intrattenere, ma creare uno spazio esperienziale funzionale alla riflessione e alla conoscenza di sé.

Infine, l'integrazione tra dialogo, voce e presenza dell'avatar apre scenari profondamente innovativi: fino a che punto la dimensione corporea e sensoriale può modulare l'efficacia di un'interazione psicologica, anche quando l'interlocutore è artificiale? Quanto il senso di presenza contribuisce a consolidare l'alleanza terapeutica, la motivazione al cambiamento e la percezione di supporto? Socrate, combinando tecnologia immersiva, modellazione linguistica e attenzione ai tempi e al ritmo

dell'interazione, permette di esplorare questi confini in modo sicuro e scientificamente fondato, offrendo una piattaforma unica per studiare come corpo, mente e tecnologia possano incontrarsi in un'esperienza psicologica integrata.

Cosa possiamo aspettarci da Socrate

Socrate può offrire uno spazio di riflessione sicuro, continuo e non giudicante, disponibile in qualsiasi momento e accessibile a chiunque disponga dell'accesso. La sua presenza costante permette agli utenti di fermarsi, esplorare i propri pensieri e dare forma a emozioni confuse che spesso restano inespresse. Attraverso domande aperte e riflessive, Socrate favorisce l'auto-esplorazione, stimola la consapevolezza dei propri schemi emotivi e incoraggia la rielaborazione di vissuti complessi. In pratica, può essere un supporto di primo livello, utile in situazioni di stress quotidiano, di ansia leggera o di incertezza emotiva, oppure integrarsi in contesti clinici già strutturati dove le risorse umane risultano limitate, garantendo continuità e accessibilità. Inoltre, la natura digitale consente di raccogliere feedback anonimi e personalizzati, permettendo un'esperienza su misura che si adatta progressivamente al ritmo e alle esigenze dell'utente.

Tuttavia, i suoi limiti appaiono altrettanto evidenti e devono essere chiaramente compresi. Socrate non possiede intenzionalità né coscienza metacognitiva; non può provare emozioni reali né percepire l'empatia nel senso umano del termine. Questo significa che, pur riconoscendo parole o frasi che indicano sofferenza, non può "sentire" ciò che l'utente prova. Non è in grado di cogliere pienamente le sfumature del linguaggio non verbale, della postura o dei segnali corporei che spesso costituiscono una parte fondamentale della comunicazione emotiva (Riva, 2002). Allo stesso modo, non può guidare processi di ristrutturazione cognitiva profonda o affrontare traumi complessi: la trasformazione terapeutica rimane, al momento, un processo eminentemente umano, che richiede la presenza, l'intuito e l'esperienza di un professionista qualificato (Riva et al., 2025).

È importante sottolineare come Socrate si collochi tra supporto e informazione, non tra terapia e diagnosi. Può aiutare a esplorare pensieri, a chiarire emozioni e a sviluppare maggiore consapevolezza, ma non sostituisce né diagnosi né trattamenti professionali. In questo senso, funge da strumento complementare, un facilitatore del dialogo con sé stessi, ma sempre all'interno di limiti chiari e definiti. Riconoscere questi confini è cruciale per un utilizzo sicuro ed efficace: la tecnologia può amplificare la capacità di riflettere, guidare l'auto-esplorazione e ridurre il senso di isolamento, ma la responsabilità del cambiamento profondo resta saldamente nelle mani dell'essere umano.

Infine, comprendere questi limiti non sminuisce il valore di Socrate, ma ne definisce la funzione reale e concreta: offrire uno spazio di ascolto costante, neutrale e strutturato, dove le parole e le riflessioni

dell'utente hanno peso, e in cui il dialogo stesso diventa uno strumento di consapevolezza, senza mai sostituire la complessità e la profondità dell'incontro umano diretto.

Sicurezza e supervisione: le condizioni per un dialogo etico e responsabile

L'uso dell'AI in contesti psicologici e di supporto emotivo richiede un'attenzione etica rigorosa e continua, poiché le implicazioni del dialogo automatizzato possono avere effetti significativi sul benessere degli utenti (McBain et al., 2025). Socrate è stato progettato seguendo principi di trasparenza, assicurando che ogni interazione sia chiara: l'utente sa in ogni momento che sta dialogando con un sistema artificiale, senza possibilità di equivoci. Il consenso informato è parte integrante dell'esperienza: prima di iniziare la conversazione, l'utente riceve spiegazioni dettagliate sul funzionamento del chatbot, sui limiti delle risposte e sulle modalità di trattamento dei dati. La protezione della privacy e dei dati personali è garantita attraverso protocolli di sicurezza avanzati, con l'obiettivo di tutelare informazioni sensibili e creare un ambiente sicuro e affidabile.

La supervisione umana rimane un elemento fondamentale per garantire un utilizzo corretto e sicuro. I modelli "human-in-the-loop" permettono di integrare Socrate nei servizi psicologici senza delegare a un algoritmo decisioni cliniche delicate o interventi terapeutici complessi. Professionisti qualificati possono monitorare le interazioni, intervenire in caso di segnali di disagio acuto e aggiornare continuamente i protocolli di sicurezza e contenuto del chatbot. Questo approccio evita il rischio che l'AI diventi un sostituto della relazione umana, preservando la centralità dell'esperto e assicurando che le risposte automatiche restino strumenti di supporto, non strumenti decisionali autonomi.

Oltre alla supervisione e alla trasparenza, l'etica del progetto si concretizza nel rispetto dei limiti della tecnologia. Socrate è programmato per riconoscere situazioni di crisi o linguaggio indicativo di disagio acuto, sospendendo la conversazione ordinaria e indirizzando immediatamente l'utente verso risorse professionali. Non tenta mai di sostituire il giudizio umano né di assumere ruoli terapeutici per i quali non è autorizzato. In questo senso, la responsabilità della tecnologia è chiara: offrire uno spazio di riflessione e accompagnamento senza oltrepassare i confini della sicurezza e della competenza clinica. La combinazione di trasparenza, supervisione umana e limiti precisi definisce un modello di dialogo responsabile, capace di sfruttare le potenzialità dell'AI senza compromettere la tutela dell'utente o la qualità della relazione terapeutica.

Infine, l'attenzione etica non riguarda solo la protezione individuale, ma anche la dimensione più ampia dell'uso responsabile della tecnologia. Socrate si inserisce in un dibattito globale sul ruolo dell'AI nella salute mentale, dimostrando che innovazione e sicurezza possono coesistere: l'obiettivo non è creare un assistente onnipotente, ma sviluppare strumenti che amplifichino l'accesso al supporto psicologico

e favoriscano l'autoconsapevolezza, sempre sotto la supervisione di professionisti umani e nel rispetto dei principi etici più rigorosi.

Conclusione

E se il dialogo socratico potesse davvero essere promosso da Socrate? Non si tratta di un paradosso filosofico, ma di un vero e proprio esperimento culturale, clinico e tecnologico, che mette in luce come l'AI possa diventare qualcosa di più di uno strumento neutro o di una semplice applicazione. Socrate dimostra che, se progettata con intenzionalità, attenzione etica e consapevolezza dei limiti, l'AI può offrire uno spazio di riflessione autentica, capace di stimolare la curiosità, il dubbio e la consapevolezza di sé. Non sostituisce il dialogo umano, ma lo integra, diventando un compagno di riflessione in quei momenti in cui la presenza di un interlocutore reale non è immediatamente disponibile, o quando le risorse tradizionali sono limitate.

Il progetto Socrate apre anche una riflessione più ampia sul significato stesso di interazione digitale: non è la velocità, l'efficienza o la completezza delle risposte a definire il valore di un sistema, ma la qualità delle domande che riesce a stimolare. L'AI può aiutare l'utente a esplorare le proprie emozioni, a dare forma a pensieri confusi e a creare consapevolezza, pur restando consapevole dei propri limiti. Questo implica una trasformazione dell'approccio alla salute mentale digitale: non più una questione di algoritmi potenti, ma di progettazione centrata sull'esperienza umana, in cui ogni scelta tecnologica sia guidata da principi di etica, sicurezza e responsabilità.

Infine, il futuro della salute mentale digitale non dipenderà solo dalla potenza computazionale o dall'accuratezza degli algoritmi, ma dalla capacità di integrare la tecnologia in modo sensibile, umano e riflessivo. Le innovazioni come Socrate ci mostrano che la tecnologia può diventare un ponte verso la consapevolezza, un invito a fermarsi, interrogarsi e dialogare con sé stessi. L'obiettivo non è sostituire la relazione terapeutica o la profondità del contatto umano, ma ampliare le possibilità di accesso, continuità e riflessione, creando uno spazio dove l'arte del dialogo, anche socratico, può vivere e prosperare nell'era digitale (Riva, 2025). In altre parole, il vero progresso non consiste solo nella potenza degli strumenti, ma nella saggezza con cui impariamo a porre le domande giuste e ad accogliere le risposte, siano esse umane o artificiali.

Bibliografia

- Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., ... & Smith, D. M. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine*, *183*(6), 589-596.
- Cooper, M., & McLeod, J. (2011). Person-centered therapy: A pluralistic perspective. *Person-Centered & Experiential Psychotherapies*, *10*(3), 210-223.
- de Borst, A. W., & de Gelder, B. (2015). Is it the real deal? Perception of virtual characters versus humans: an affective cognitive neuroscience perspective. *Frontiers in psychology*, *6*, 576.
- Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., ... & Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*, *2*(11), 688-701.
- Eliud, O. (2025). Socratic method as a therapeutic discourse for mental health. *Diagnostics and Therapeutics*, 1-9.
- Frisone, F., Pupillo, C., Rossi, C., & Riva, G. (2025). SOCRATES. Developing and evaluating a fine-Tuned ChatGPT model for accessible mental health intervention. *Cyberpsychology, Behavior, and Social Networking*, *28*(5), 366-368.
- Lundin, R. M., Yeap, Y., & Menkes, D. B. (2023). Adverse effects of virtual and augmented reality interventions in psychiatry: systematic review. *JMIR mental health*, *10*, e43240.
- McBain, R. K., Cantor, J. H., Zhang, L. A., Baker, O., Zhang, F., Halbisen, A., ... & Yu, H. (2025). Competency of large language models in evaluating appropriate responses to suicidal ideation: Comparative study. *Journal of Medical Internet Research*, *27*, e67891.
- Ovsyannikova, D., de Mello, V. O., & Inzlicht, M. (2025). Third-party evaluators perceive AI as more compassionate than expert humans. *Communications Psychology*, *3*(1), 4.
- Riva, G. (2002). The sociocognitive psychology of computer-mediated communication: The present and future of technology-based interactions. *Cyberpsychology & behavior*, *5*(6), 581-598.
- Riva, G., Rossi, C., Frisone, F., & Oasi, O. (2025). Intelligenza Artificiale in Psicologia: Potenziare la Cura, Preservare l'Umanità. *Psicoterapia Cognitiva e Comportamentale*, *31*(2).

- Riva, G. (2025). Digital “we”: Human sociality and culture in the era of social media and artificial intelligence. *American Psychologist*, *80*(9), 1535.
- Rossi Monti, M., & D'Agostino, A. (2018). Diagnosis or diagnosing? From symptoms to a phenomenological-psychoanalytic stance on the patient's world. *The Psychoanalytic Review*, *105*(2), 209-222.
- Spencer, L., Broome, M. R., & Stanghellini, G. (2025). The future of phenomenological psychopathology. *Philosophical psychology*, *38*(1), 1-16.
- Spring, B. (2007). Evidence-based practice in clinical psychology: What it is, why it matters; what you need to know. *Journal of clinical psychology*, *63*(7), 611-631.
- Vos, J., & Vitali, D. (2018). The effects of psychological meaning-centered therapies on quality of life and psychological stress: A meta-analysis. *Palliative & supportive care*, *16*(5), 608-632.
- Vowels, L. M., Francois-Walcott, R. R., & Darwiche, J. (2024). AI in relationship counselling: evaluating ChatGPT's therapeutic capabilities in providing relationship advice. *Computers in Human Behavior: Artificial Humans*, *2*(2), 100078.
- Yeh, P. L., Kuo, W. C., Tseng, B. L., & Sung, Y. H. (2025). Does the AI-driven Chatbot Work? Effectiveness of the Woebot app in reducing anxiety and depression in group counseling courses and student acceptance of technological aids. *Current Psychology*, *44*(9), 8133-8145.
- Yonatan-Leus, R., & Brukner, H. (2025). Comparing perceived empathy and intervention strategies of an AI chatbot and human psychotherapists in online mental health support. *Counselling and Psychotherapy Research*, *25*(1), e12832.
- Zhu, Q., Chau, A., Cohn, M., Liang, K. H., Wang, H. C., Zellou, G., & Yu, Z. (2022, July). Effects of emotional expressiveness on voice chatbot interactions. In *Proceedings of the 4th conference on conversational us*

Alle origini dell'AI Literacy: una ricognizione bibliografica (1982–2016) – *di Andrea Laudadio*, Head of TIM Academy & Development

Un viaggio nella "preistoria" dell'AI literacy tra MIT, esercito USA, grandi aziende e scuole. Dai sistemi esperti all'IA generativa, emergono modelli di apprendimento, trasferimento di conoscenza e alfabetizzazione organizzativa che anticipano i dilemmi educativi e strategici di oggi

Abstract

L'AI Literacy non è nata con ChatGPT. Esisteva una riflessione articolata già quarant'anni fa, in contesti diversi: laboratori universitari, ambienti militari, grandi aziende, scuole. Il presente contributo riporta alla luce otto pubblicazioni antecedenti al 2017 che utilizzano esplicitamente il termine "AI literacy", analizzandone i contenuti e proponendo un modello interpretativo che individua quattro dimensioni del concetto largamente trascurate dalla letteratura contemporanea: la literacy come canone culturale, come padronanza di un medium rappresentazionale, come memoria organizzativa e come leva strategica. Le domande fondamentali sono rimaste invariate attraverso i cicli tecnologici: come sviluppare il pensiero critico, come trasferire il sapere, come amplificare l'intelligenza umana senza sostituirla.

Introduzione

Il 30 novembre 2022 rappresenta uno spartiacque nella storia della tecnologia e della cultura digitale. Il rilascio pubblico di ChatGPT da parte di OpenAI ha innescato un "inevitabile salto di paradigma" (Bozkurt, 2023) che ha investito la società, l'economia e, in particolare, l'educazione[1].

In questo nuovo contesto, il concetto di AI Literacy ha acquisito un'urgenza senza precedenti. Se prima dell'era generativa l'AI Literacy era definita da Long e Magerko (2020) come "un insieme di competenze che consentono agli individui di valutare criticamente le tecnologie di IA, comunicare e collaborare efficacemente con l'IA e utilizzare l'IA come strumento online, a casa e sul posto di lavoro" (p. 2), oggi il dibattito si è evoluto per includere le specificità dei modelli generativi, dando vita al concetto di GenAI Literacy (Bozkurt, 2024).

[1] Una versione preliminare e divulgativa di parte dei contenuti qui trattati è apparsa su Agenda Digitale. Il presente contributo ne costituisce la rielaborazione scientifica, con apparato metodologico, bibliografico e argomentativo significativamente ampliati.

Ma cosa accadeva prima che il mondo conoscesse l'interfaccia conversazionale di ChatGPT? Una ricognizione bibliografica mirata rivela che, prima del 2017, esistono 8 pubblicazioni che utilizzano esplicitamente il termine "AI literacy". Di queste, tre provengono dal gruppo di ricerca austriaco guidato da Gerald Steinbauer-Wagner. Gli altri cinque documenti sono emersi in contesti eterogenei: un laboratorio di ricerca del MIT, il mondo militare statunitense, una grande azienda brasiliana e un istituto di ricerca economica giapponese.

Nota metodologica

Il presente studio adotta un approccio di ricognizione bibliografica esplorativa (Grant & Booth, 2009) finalizzato a identificare e analizzare le pubblicazioni che hanno utilizzato esplicitamente il termine "AI literacy" prima del 2017. A differenza delle revisioni sistematiche basate su protocolli PRISMA (come quelle condotte da Yang et al., 2025, e Casal-Otero et al., 2023, sul periodo successivo), l'obiettivo non è mappare la totalità della produzione scientifica, ma ricostruire la genealogia di un concetto attraverso l'analisi delle sue prime attestazioni.

Le banche dati interrogate includono Scopus, Web of Science Core Collection, Google Scholar, IEEE Xplore, ACM Digital Library, DTIC (Defense Technical Information Center) e il repository DSpace del MIT. La stringa di ricerca principale è stata "AI literacy" OR "artificial intelligence literacy", applicata a titolo, abstract e testo completo, con filtro temporale fino al 31 dicembre 2016. I criteri di inclusione prevedevano la presenza esplicita del sintagma "AI literacy" nel documento; sono stati inclusi articoli su rivista, atti di convegno, working paper e rapporti tecnici. La ricerca ha identificato un corpus di 8 documenti, distribuiti su un arco temporale di 34 anni (1982–2016), sintetizzati nella Tabella 1.

Tabella 1. Corpus dei documenti contenenti il termine "AI literacy" pubblicati prima del 2017.

Anno	Autore/i	Titolo	Tipo	Contesto
1982	Agre	What to read: A biased guide to AI literacy for the beginner	Working paper	Accademico (MIT)
1989	Hill	The mind at AI: Horseless carriage to clock	Articolo	Accademico
1990	U.S. Army EPG	Methodology investigation of AI test officer support tool III	Report tecnico	Militare

1997	Tenorio et al.	A model for advanced in-house industrial training	Articolo	Industriale
2016	Burgsteiner et al.	IRobot: Teaching the basics of AI in high schools	Atti AAAI	Educativo K-12
2016	Kandlhofer et al.	AI and computer science in education: From kindergarten to university	Atti IEEE FIE	Educativo K-12
2016	Kandlhofer & Steinbauer	Evaluating the impact of educational robotics on pupils	Articolo	Educativo K-12
2016	Konishi	What is needed for AI literacy?	Colonna RIETI	Economico

È necessario riconoscere i limiti di questa scelta lessicale. L'adozione del sintagma "AI literacy" come unico filtro esclude contributi concettualmente affini ma formulati con terminologia diversa, quali "expert systems training", "intelligent systems education" o "computational literacy". La delimitazione è una scelta consapevole, motivata dall'interesse per la storia del concetto così come si è cristallizzato nella letteratura successiva (Long & Magerko, 2020; Ng et al., 2021). L'esplorazione delle aree terminologiche adiacenti rimane un orizzonte aperto per studi futuri.

L'AI Literacy: definizioni

Prima di procedere con l'analisi del corpus, è necessario definire che cosa si intenda per AI Literacy. La prima definizione sistematica e operativa del concetto è attribuibile a Yoko Konishi (2016), economista e Senior Fellow del RIETI, il think tank giapponese. Nel suo articolo "What is Needed for AI Literacy?", pubblicato nel gennaio 2016, Konishi delinea l'AI Literacy non come una competenza tecnica, ma come una disposizione cognitiva e strategica:

"AI literacy refers to being conscious about whether there are any standardized/formatted tasks that have not been applied for AI due to large data volume, and whether excessive labor, money or time is spent for 'categorization, repetition, exploration, organization and optimization.' We should make it a habit to think about what can and cannot be applied to AI" (Konishi, 2016, p. 3).

Questa definizione, radicata nel contesto economico-organizzativo, anticipa di quattro anni la formulazione di Long e Magerko (2020), che si concentra sulle competenze individuali di valutazione

critica, comunicazione e collaborazione con i sistemi di IA. La concettualizzazione di Konishi colloca l'AI Literacy nel dominio della consapevolezza strategica e del pensiero critico applicato ai processi.

La letteratura contemporanea ha progressivamente arricchito questa definizione. Ng et al. (2021), in una revisione esplorativa, identificano quattro dimensioni fondamentali: conoscenza e comprensione dell'IA, uso e applicazione, valutazione e creazione di sistemi di IA, etica e impatto sociale. Yang et al. (2025), in un'analisi bibliometrica di un decennio di ricerca, hanno tracciato quattro traiettorie evolutive del campo, confermandone la natura interdisciplinare e le connessioni con la data literacy, la digital literacy e il pensiero computazionale.

Con l'avvento dell'IA generativa, il concetto si è ulteriormente evoluto. Bozkurt (2024) ha proposto il framework 3wAI per la GenAI Literacy, articolato in tre dimensioni: *Know What* (comprensione teorica e concettuale dell'IA), *Know How* (applicazione pratica, prompting e valutazione degli output) e *Know Why* (riflessione etica e critica). Questo framework segna un passaggio dalla literacy come competenza statica a quella come postura adattiva.

Ciò che emerge è che l'AI Literacy non è mai stata una competenza monolitica, bensì un costrutto multidimensionale che si adatta ai contesti tecnologici, sociali ed educativi. Come le sezioni seguenti documenteranno, le radici di questa complessità affondano più indietro nel tempo di quanto la letteratura abbia riconosciuto.

I documenti pionieristici: una panoramica

Le sezioni seguenti analizzano ciascun contributo del corpus in ordine cronologico, ad eccezione della definizione di Konishi (2016), già discussa, e dei tre contributi del gruppo austriaco, trattati congiuntamente. L'analisi si concentra sulla concezione di AI literacy, espressa o implicita, presente in ciascun documento e sulla sua rilevanza per il dibattito contemporaneo.

L'alba dell'AI Literacy: la guida bibliografica di Philip E. Agre (1982)

Il punto di partenza dell'indagine è il Working Paper 239 del MIT AI Lab, intitolato *"What to read: A biased guide to AI literacy for the beginner"*, scritto da Philip E. Agre nel novembre del 1982.

Nel 1982, Agre era un giovane dottorando al MIT AI Lab, al secondo o terzo anno del PhD in Electrical Engineering and Computer Science. Cresciuto nel Maryland, aveva iniziato il dottorato sotto la supervisione di Michael Brady. Quando Brady lasciò il MIT per Oxford, Agre e David Chapman passarono sotto la guida di Rodney Brooks. Il MIT AI Lab degli anni '80 era l'epicentro della ricerca sull'intelligenza artificiale, nel pieno dell'ottimismo legato ai sistemi esperti. Agre e Chapman stavano sviluppando un

approccio che enfatizzava il comportamento come emergente dall'interazione con l'ambiente, un lavoro che culminò nel 1989 con l'articolo "What are plans for?".

La guida nasce come atto di condivisione tra pari. Il Working Paper si apre con una dichiarazione di modestia:

"This note tries to provide a quick guide to AI literacy for the beginning AI hacker and for the experienced AI hacker or two whose scholarship isn't what it should be. Most will recognize it as the same old list of classic papers, give or take a few that I feel to be under- or over-rated. It is not guaranteed to be thorough or balanced or anything like that" (Agre, 1982, p. 1).

L'uso del termine "hacker" non è casuale: negli anni '80, al MIT, designava un programmatore appassionato, tecnicamente competente e intellettualmente libero. La guida si articola in sezioni tematiche, dalla filosofia alla fantascienza, e raccomanda oltre 150 riferimenti con annotazioni e giudizi critici.

Per Agre, l'AI literacy non è né una competenza astratta né una serie di abilità tecniche. È innanzitutto la padronanza di un canone di testi fondamentali che consente di comprendere i concetti, le tecniche, le sfide e i dibattiti del campo. È un'alfabetizzazione che si costruisce attraverso lo studio, la lettura critica e l'inserimento in una comunità di pratica. L'AI literacy, in questa prospettiva, è una forma di literacy accademica: saper leggere, interpretare, contestualizzare e criticare la letteratura scientifica. Ma è anche la capacità di orientarsi in un campo in rapida evoluzione, di distinguere le mode passeggere dalle intuizioni durature, di riconoscere i problemi aperti e le direzioni promettenti. Questa concezione anticipa le categorie di "fondazioni concettuali" e "pensiero critico" che Shiri (2024) ha recentemente incluso nella sua tassonomia dell'AI literacy.

Oggi, di fronte all'esplosione dell'IA generativa, ci troviamo nuovamente nella necessità di alfabetizzare masse di persone che devono orientarsi in un panorama tecnologico complesso. E, come nel 1982, questa alfabetizzazione non può ridursi alla padronanza degli strumenti o alla memorizzazione di procedure, ma richiede la costruzione di un quadro concettuale che consenta di comprendere, valutare e utilizzare criticamente la tecnologia.

L'IA come medium rappresentazionale: William C. Hill (1989)

Il saggio di Hill, *"The Mind at AI: Horseless Carriage to Clock"*, pubblicato su AI Magazine nel 1989, compie un salto epistemologico proponendo una riconcettualizzazione della natura dell'intelligenza artificiale. Scrivendo al culmine del boom dei sistemi esperti, Hill si distacca dalle due narrazioni

dominanti (l'IA come mente artificiale e l'IA come ingegneria degli artefatti intelligenti) per proporre una terza via: concepire l'intelligenza artificiale come un nuovo medium rappresentazionale basato sulla computazione.

"A more fundamental human urge is playing out, that of seeking new expressive power through new expressive means" (Hill, 1989, p. 29).

Hill, attingendo alla teoria dei media di McLuhan (1964) e di Ong (1982), sposta il focus dalla macchina "pensante" al medium "espressivo". Identifica quattro proprietà rappresentazionali della computazione: dinamica (processi temporali), profondità (livelli di astrazione), facilitazione (adattamento tra operazioni mentali e computazionali) e interattività (influenza del contesto). Il punto qualificante è che "i mezzi attraverso cui la computazione facilita la rappresentazione sono essi stessi estensibili" (p. 31): la computazione è un meta-medium.

Hill conia il termine *pristechnolocution* per descrivere il fenomeno per cui le nuove tecnologie vengono concepite attraverso metafore di tecnologie precedenti. Esempi classici includono "carrozza senza cavalli" per l'automobile, "telegrafo senza fili" per la radio. Secondo Hill, "intelligenza artificiale" è una pristechnolocution che descrive erroneamente la tecnologia computazionale come imitazione dell'intelletto umano, distogliendo l'attenzione dalla sua natura di mezzo espressivo. Questo tipo di denominazione è problematico perché dirige l'attenzione verso somiglianze superficiali, oscura le reali potenzialità e i limiti della nuova tecnologia e suona anacronisticamente ingenuo dopo pochi decenni. La pristechnolocution riflette la tendenza umana a comprendere l'innovazione attraverso categorie familiari, ritardando la comprensione delle proprietà effettive della nuova tecnologia.

Questa lettura è oggi essenziale per comprendere l'IA generativa: usiamo ChatGPT o Midjourney non come intelligenze autonome, ma come media rappresentazionali per scrivere, argomentare, visualizzare e creare. L'AI Literacy, in questa prospettiva, non è solo la capacità di usare uno strumento, ma la competenza semiotica di padroneggiare un nuovo linguaggio: comprenderne la grammatica (i prompt), la sintassi (i modelli sottostanti) e la pragmatica (gli usi efficaci e critici). Hill ci invita a guardare l'IA non con il timore reverenziale di una mente aliena, ma con la curiosità critica di un nuovo medium che, come la scrittura o la stampa prima di essa, sta ridefinendo le fondamenta della nostra cultura.

L'IA come memoria organizzativa: lo U.S. Army (1990)

Il report tecnico AD-A247 360 dell'U.S. Army Electronic Proving Ground (dicembre 1990), intitolato *"Methodology investigation of AI test officer support tool III"*, è un caso di studio poco conosciuto. Si

tratta di un rapporto declassificato prodotto dal centro di Fort Huachuca, in Arizona, specializzato nel testing di sistemi C3I.

Il report identifica la fragilità della conoscenza esperta nelle organizzazioni complesse:

"This hard-earned expertise is frequently lost to the organization because of personnel reassignment or attrition" (U.S. Army, 1990, p. 3).

La soluzione proposta consiste nell'inserimento di sistemi esperti su microcomputer, con un obiettivo non solo operativo ma anche pedagogico. La scelta del microcomputer (anziché dei costosi hardware specializzati come le macchine LISP che dominavano l'IA commerciale dell'epoca) è dettata da criteri di accessibilità e costo-efficacia, e rappresenta una scelta lungimirante che anticipa il collasso del mercato dei sistemi esperti proprietari. L'aspetto qualificante del documento è il suo duplice obiettivo:

"Education of the user (test officer); in addition to merely providing a solution, provide a means for the user to train new personnel by exposure to how a process is accomplished" (U.S. Army, 1990, p. 4).

Il rapporto descrive workshop collaborativi tra esperti di dominio e ingegneri della conoscenza, nonché un *apprenticeship program* per trasferire non solo l'uso degli strumenti, ma anche la metodologia di sviluppo.

L'apprendimento è concepito come organizzativo, iterativo e basato sulla pratica. L'approccio sposta l'AI Literacy dall'individuo all'organizzazione: non si tratta di comprendere l'IA come concetto astratto, ma di sviluppare un ecosistema in cui la conoscenza viene codificata in sistemi intelligenti, il personale viene formato a utilizzarli, gli utenti contribuiscono al miglioramento continuo attraverso il feedback e la metodologia viene trasferita ad altri gruppi dell'organizzazione. Questo modello, concepito in un contesto militare tre decenni fa, è di diretta rilevanza per le sfide che le imprese affrontano oggi nell'era dell'IA generativa.

Il modello industriale: AI Literacy come leva strategica in PETROBRAS (1997)

L'articolo *"A model for advanced in-house industrial training"* di Tenorio, Maia e Grant-Tenorio (1997), pubblicato lo stesso anno del testo di Gilster (1997) sulla digital literacy, descrive un'iniziativa di PETROBRAS. Nel contesto post-"inverno dell'IA", l'azienda scelse di investire nella formazione interna anziché acquistare soluzioni esterne.

Il programma educativo di livello graduate, "time- and cost-effective" (p. 144), prevedeva due percorsi: CAVIAR (Computer-Aided VIsion and Artificial Reasoning), focalizzato sull'IA simbolica e sui sistemi esperti, e CERN (Computational Emulation of Reasoning with Neural networks), dedicato all'approccio connessionista delle reti neurali.

L'elemento qualificante del modello PETROBRAS è l'adozione di un approccio eminentemente *project-oriented*. La formazione non era un esercizio teorico, ma un processo di trasferimento tecnologico calato nella realtà aziendale. I dipendenti-studenti lavoravano su progetti reali, di diretto interesse per l'azienda: dall'ottimizzazione dei processi di raffinazione alla manutenzione predittiva delle piattaforme offshore, dalla gestione delle scorte all'analisi dei dati geologici. L'apprendimento era contestualizzato, motivante e immediatamente applicabile.

L'AI Literacy si trasforma qui da competenza individuale a capacità organizzativa. L'obiettivo non è solo creare singoli esperti, ma diffondere una cultura dell'IA all'interno dell'impresa, costruendo un ecosistema in grado di identificare opportunità, sviluppare soluzioni, gestire l'innovazione e trasferire conoscenza alle nuove generazioni di professionisti. Il caso PETROBRAS rappresenta un modello sistematico di corporate AI literacy che anticipa di due decenni le attuali strategie di upskilling e reskilling, dimostrando che l'adozione dell'IA non poteva essere delegata a pochi tecnici, ma richiedeva un investimento nella cultura dell'intera organizzazione.

Il gruppo austriaco: portare l'AI Literacy nelle scuole (2016)

Completano il corpus tre contributi distinti della Graz University of Technology. Martin Kandlhofer e Gerald Steinbauer, insieme a Harald Burgsteiner della Fachhochschule Joanneum, hanno prodotto nel 2016 lavori che costituiscono i primi tentativi sistematici di sviluppare un'AI literacy scolastica.

Il primo contributo (Burgsteiner et al., 2016), presentato al *Symposium on Educational Advances in Artificial Intelligence* dell'AAAI, documenta il progetto iRobot: un modulo di 16 ore per studenti delle superiori. Utilizzando i robot Lego Mindstorms, gli studenti affrontavano problemi classici dell'IA: navigazione autonoma, riconoscimento di pattern e apprendimento per rinforzo.

Il secondo contributo (Kandlhofer, Steinbauer, Hirschmugl-Gaisch & Huber, 2016), presentato alla IEEE Frontiers in Education Conference, propone un curriculum completo dall'asilo all'università e introduce esplicitamente il termine "AI literacy" in una formulazione che anticipa di quattro anni quella di Long e Magerko (2020). Il curriculum si articola in quattro livelli progressivi.

Il terzo contributo (Kandlhofer & Steinbauer, 2016), pubblicato su *Robotics and Autonomous Systems*, fornisce la validazione empirica del programma. Lo studio ha misurato l'impatto della robotica educativa sulle competenze tecniche, sociali e sugli atteggiamenti verso la scienza, evidenziando effetti statisticamente significativi sulle competenze tecniche. Questo studio costituisce una delle prime evidenze empiriche sull'efficacia degli interventi educativi finalizzati all'AI literacy in contesti scolastici.

Nel loro complesso, i tre contributi attestano che già prima dell'IA generativa esisteva una consapevolezza matura della necessità di alfabetizzare le nuove generazioni all'intelligenza artificiale.

Discussione

L'analisi del corpus consente di formulare una proposta interpretativa che va oltre la ricognizione storica. I documenti esaminati, prodotti in contesti disciplinari, geografici e istituzionali eterogenei, convergono nel delineare quattro dimensioni dell'AI literacy che i framework contemporanei non coprono o solo parzialmente. Un confronto sistematico tra i tre modelli più citati (Tabella 2) mette in luce il contributo specifico di questa genealogia.

Tabella 2. Confronto tra le dimensioni emerse dal corpus storico e i framework contemporanei di AI literacy.

Dimensione storica	**Long & Magerko (2020)**	**Ng et al. (2021)**	**Bozkurt (2024) 3wAI**	**Copertura**
Canone culturale (Agre)	Assente	Assente	Parziale (Know What)	Bassa
Medium rappr. (Hill)	Comp. 2, 5	Uso e applicazione	Know How	Parziale
Memoria org. (U.S. Army)	Assente	Assente	Assente	Nulla
Leva strategica (Tenorio, Konishi)	Assente	Assente	Parziale (Know Why)	Bassa

La literacy come canone culturale. Agre (1982) concepisce l'AI literacy come padronanza di un corpus di testi fondamentali e come immersione in una comunità di pratica scientifica. Questa dimensione — l'alfabetizzazione come costruzione di un quadro concettuale condiviso — è sostanzialmente assente nei framework contemporanei, che privilegiano le competenze operative e gli atteggiamenti. Nessuna delle 16 competenze di Long e Magerko (2020) riguarda né la familiarità con la storia intellettuale del

campo né la capacità di situare le tecnologie attuali in un contesto epistemologico più ampio. La tassonomia di Shiri (2024) colma parzialmente questa lacuna includendo le "fondazioni concettuali" come facet autonoma, ma il punto di Agre è più radicale: non si tratta di "conoscere l'IA" in astratto, bensì di saper navigare criticamente la sua tradizione intellettuale.

La literacy è la padronanza di un medium. Hill (1989) propone di concepire l'IA non come un'intelligenza, bensì come un medium rappresentazionale. Questa lettura trova un riscontro parziale nella dimensione Know-how di Bozkurt (2024), che include competenze di prompting e di valutazione degli output, attività che presuppongono la comprensione dell'IA come mezzo espressivo. La convergenza non è però completa: il framework 3wAI tratta il prompting come un'abilità tecnica, mentre Hill lo collocherebbe nella tradizione della retorica e della semiotica, come padronanza dei mezzi di un linguaggio. La differenza è pedagogicamente rilevante: formare al prompting come tecnica produce utenti competenti; formare alla comprensione del medium produce pensatori critici. Yang et al. (2025) hanno documentato, nella letteratura recente, la crescente connessione tra AI literacy e media literacy, confermando la traiettoria anticipata da Hill.

La literacy come memoria organizzativa. Il report dell'U.S. Army (1990) introduce una dimensione completamente assente dai framework contemporanei: l'AI literacy, intesa come capacità dell'organizzazione di preservare e trasmettere la conoscenza attraverso sistemi intelligenti. Né Long e Magerko (2020), né Ng et al. (2021), né Bozkurt (2024) contemplano questa dimensione: il soggetto dei loro framework è sempre l'individuo. Il lavoro di Cetindamar et al. (2022) sull'AI literacy dei dipendenti nelle organizzazioni digitali segnala questa lacuna, ma non la colma. L'assenza della dimensione organizzativa è particolarmente problematica nell'era dell'IA generativa, in cui l'adozione efficace dipende non solo dalla competenza dei singoli utenti, ma anche dalla capacità dell'organizzazione di integrare l'IA nei propri processi di knowledge management.

La literacy come leva strategica. Tenorio et al. (1997) e Konishi (2016) collocano l'AI literacy nel dominio della strategia organizzativa e della politica economica. L'alfabetizzazione non è fine a sé stessa, ma una condizione abilitante per l'innovazione. Il framework 3wAI recupera parzialmente questa dimensione nella componente *Know Why*, che include la riflessione sulle implicazioni sociali. La differenza è di scala: Bozkurt pensa al singolo utente che riflette eticamente; Tenorio e Konishi pensano all'organizzazione che ridisegna i propri processi. Questa dimensione strategica è coerente con le recenti raccomandazioni dell'OCSE sulla necessità di integrare l'AI literacy nelle politiche industriali, non solo in quelle educative.

In sintesi, il corpus storico non aggiunge semplicemente "precedenti" alla cronologia dell'AI literacy: rivela dimensioni strutturali che i framework contemporanei non coprono. In particolare, l'assenza delle dimensioni organizzativa e culturale-canonica nei modelli più citati suggerisce che il campo, concentrandosi sull'individuo e sulle competenze operative, abbia perso di vista aspetti rilevanti dell'alfabetizzazione all'IA. Il recupero di queste dimensioni non è un esercizio antiquario, ma un contributo alla costruzione di framework più aderenti alla complessità delle sfide attuali.

Conclusioni

L'analisi di otto documenti antecedenti al 2017 ha consentito di ricostruire una genealogia dell'AI Literacy più articolata rispetto a quella generalmente riconosciuta dalla letteratura. Il contributo di questo studio è duplice: la documentazione di un corpus finora inesplorato e la proposta di un modello interpretativo che individua quattro dimensioni dell'AI literacy — canone culturale, medium rappresentazionale, memoria organizzativa, leva strategica — non adeguatamente rappresentate nei framework contemporanei.

Tre considerazioni meritano di essere evidenziate. La prima riguarda la visione di Hill (1989) dell'IA come medium. Le competenze che i framework contemporanei identificano come cruciali per la GenAI Literacy — dalla formulazione di prompt efficaci alla valutazione critica degli output — sono declinazioni attuali di quella padronanza dei mezzi espressivi di cui Hill parlava. Come la stampa non ha eliminato gli scrittori ma ha trasformato la pratica della scrittura, così l'IA generativa ridefinisce le modalità del lavoro intellettuale senza eliminare chi lo svolge.

La seconda considerazione riguarda la natura triadica dei processi di adozione dell'IA, già evidente nella fase embrionale del campo. Le tre leve fondamentali — persone, tecnologie e processi — dovevano essere concepite in modo integrato. Questa lezione rimane attuale: l'implementazione dell'IA generativa nelle organizzazioni non può ridursi all'adozione di una nuova tecnologia, ma richiede un ripensamento dei modelli operativi e un investimento nella cultura e nelle competenze delle persone.

La terza considerazione è che le domande più rilevanti non sono mai state puramente tecniche. I pionieri documentati in questo studio erano concentrati su problemi concreti: come si apprende, come si trasferisce la conoscenza, come si lavora meglio con le macchine. La loro attenzione, pragmatica e umanistica, offre un utile riferimento contro le narrazioni iperboliche che caratterizzano il dibattito odierno.

L'AI Literacy, nella prospettiva che emerge da questa ricognizione, non è una competenza tecnica da acquisire, ma una postura intellettuale da coltivare: quella della curiosità critica, della consapevolezza riflessiva e della responsabilità creativa di fronte a uno strumento che sta ridefinendo le fondamenta della nostra civiltà.

Bibliografia

Agre, P. E. (1982). *What to read: A biased guide to AI literacy for the beginner* (Working Paper 239). MIT Artificial Intelligence Laboratory. https://dspace.mit.edu/handle/1721.1/41185

Bozkurt, A. (2023). Generative artificial intelligence (AI) powered conversational educational agents: The inevitable paradigm shift. *Asian Journal of Distance Education, 18*(1), 198–204. https://doi.org/10.5281/zenodo.7716416

Bozkurt, A. (2024). Why generative AI literacy, why now and why it matters in the educational landscape? Kings, queens and GenAI dragons. *Open Praxis, 16*(3), 283–290. https://doi.org/10.55982/openpraxis.16.3.739

Burgsteiner, H., Kandlhofer, M., & Steinbauer, G. (2016). IRobot: Teaching the basics of artificial intelligence in high schools. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 30, No. 1, pp. 4126–4127). https://doi.org/10.1609/aaai.v30i1.9864

Casal-Otero, L., Catala, A., Fernández-Morante, C., Taboada, M., Cebreiro, B., & Barro, S. (2023). AI literacy in K-12: A systematic literature review. *International Journal of STEM Education, 10*(1), 29. https://doi.org/10.1186/s40594-023-00418-7

Cetindamar, D., Kitto, K., Wu, M., Zhang, Y., Abedin, B., & Knight, S. (2022). Explicating AI literacy of employees at digital workplaces. *IEEE Transactions on Engineering Management, 71*, 810–823. https://doi.org/10.1109/TEM.2022.3175299

Crevier, D. (1993). *AI: The tumultuous history of the search for artificial intelligence*. Basic Books.

Dreyfus, H. L. (1972). *What computers can't do: A critique of artificial reason*. Harper & Row.

Gilster, P. (1997). *Digital literacy*. John Wiley & Sons.

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

Hill, W. C. (1989). The mind at AI: Horseless carriage to clock. *AI Magazine, 10*(2), 29–39. https://doi.org/10.1609/aimag.v10i2.742

Kandlhofer, M., & Steinbauer, G. (2016). Evaluating the impact of educational robotics on pupils' technical- and social-skills and science related attitudes. *Robotics and Autonomous Systems, 75*(B), 679–685. https://doi.org/10.1016/j.robot.2015.09.007

Kandlhofer, M., Steinbauer, G., Hirschmugl-Gaisch, S., & Huber, P. (2016). Artificial intelligence and computer science in education: From kindergarten to university. In *2016 IEEE Frontiers in Education Conference (FIE)* (pp. 1–9). IEEE. https://doi.org/10.1109/FIE.2016.7757570

Konishi, Y. (2016, January 13). What is needed for AI literacy? *RIETI Special Series: Priorities for the Japanese Economy in 2016*. https://www.rieti.go.jp/en/columns/s16_0014.html

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). ACM. https://doi.org/10.1145/3313831.3376727

McLuhan, M. (1964). *Understanding media: The extensions of man*. McGraw-Hill.

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence, 2*, 100041. https://doi.org/10.1016/j.caeai.2021.100041

Ong, W. J. (1982). *Orality and literacy: The technologizing of the word*. Methuen.

Shiri, A. (2024). Artificial intelligence literacy: A proposed faceted taxonomy. *Digital Library Perspectives*. Advance online publication. https://doi.org/10.1108/DLP-04-2024-0067

Tenorio, M. F. d. M., Maia, A. C. P., & Grant-Tenorio, A. E. (1997). A model for advanced in-house industrial training: A case study on intelligent system technology transfer using a project-oriented approach. *IEEE Transactions on Education, 40*(2), 144–153. https://doi.org/10.1109/13.572329

U.S. Army Electronic Proving Ground. (1990). *Methodology investigation of AI test officer support tool III* (Report AD-A247 360). Fort Huachuca, AZ.

Yang, Y., Zhang, Y., Sun, D., He, W., & Wei, Y. (2025). Navigating the landscape of AI literacy education: Insights from a decade of research (2014–2024). *Humanities and Social Sciences Communications, 12*, 374. https://doi.org/10.1057/s41599-025-04583-8

Scrolling e parcellizzazione dell'attenzione: effetti dei flussi informativi nei comportamenti, nelle relazioni e nei processi cognitivi contemporanei – di ***Marino D'Amore***, Docente di Sociologia della comunicazione, Università degli Studi Niccolò Cusano

Lo scrolling sui social media frammenta l'attenzione in micro-unità temporali, riducendo la capacità di lettura profonda e riflessione critica. Gli algoritmi trasformano l'attenzione in merce, con ricadute sui legami sociali, sull'apprendimento e sulla qualità del discorso pubblico contemporaneo

Introduzione

L'attenzione rappresenta una risorsa estremamente vulnerabile e, al contempo, un bene strategico che dispositivi e piattaforme digitali cercano di attirare continuamente (Davenport, Beck, 2001). L'avvento dei social media e il conseguente utilizzo dei feed algoritmici hanno consolidato lo scrolling come una pratica quotidiana per miliardi di utenti, che si sostanzia nel passaggio fruitivo rapido di contenuti su Facebook e Instagram fino alle timeline di X e TikTok. Essa ridefinisce il modo in cui riceviamo informazioni, costruiamo significati, ma soprattutto strutturiamo il tempo. La parcellizzazione dell'attenzione, in brevi frammenti temporali che depauperano la capacità di concentrazione antropica, è diventata uno dei principali temi di indagine nel dibattito scientifico (Rosen, 2008; Carr, 2011). Tale articolo si propone di esplorare l'eziologia di tali fenomeni, le interazioni con il capitale simbolico e i rischi concreti della costruzione eterodiretta di un'attenzione liquida e sempre più industrializzata.

Attenzione e modernità: da risorsa cognitiva a merce sociale

La sociologia dell'attenzione affonda le sue radici in riflessioni precedenti alla rivoluzione digitale. Nel 1890 William James definiva l'attenzione come "la presa di possesso da parte della mente, in forma chiara e vivida, di uno tra gli oggetti o dei flussi di pensiero possibili" (James, 1890, p. 403). Negli studi successivi, l'attenzione è stata analizzata come funzione cognitiva e risorsa orientata socialmente: la selezione di stimoli diventa pratica influenzata da norme, istituzioni e gerarchie sociali (Goffman, 1974). L'attenzione come merce sociale emerge con forza nell'economia dell'informazione contemporanea. Herbert Simon (1969) osservava che con l'aumento dell'offerta informativa l'attenzione diventa una risorsa scarsa e competitiva. Nelle piattaforme digitali, tale competizione si traduce in meccanismi

progettati per massimizzare l'engagement, spesso attraverso logiche di micro-distrazione e ricompense intermittenti (Alter, 2018).

Lo scrolling: dinamiche di produzione e consumo

Lo scrolling non è semplicemente un gesto fisico: è una pratica sociale intrinsecamente legata a contesti culturali, modelli di interazione e forme di soggettivazione digitale, basti considerare come esso sia inscindibile da interfacce che suggeriscono un'infinita disponibilità di contenuti (van Dijck, Poell, 2013). L'interazione con feed personalizzati introduce una circolarità performativa: più scrolli, più l'algoritmo apprende, più il tuo feed si adatta alle tue preferenze, spingendo a ulteriori scroll, secondo una dinamica meccanicistica che si reitera. I dati empirici indicano tassi di scroll medi quotidiani che superano i 5000–6000 pixel già nei primi anni del 2020 (Faverio, Sidoti, 2025). Questo ritmo induce una forma di cognizione perennemente in transizione, dove l'attenzione si attiva e si disattiva rapidamente, condizionando la profondità dell'elaborazione nei processi decisionali (Ophir, Nass, Wagner, 2009).

Parcellizzazione dell'attenzione: definizione e dimensioni

La parcellizzazione dell'attenzione si configura come fenomeno multidimensionale:

- temporale: l'attenzione si distribuisce in unità temporali brevissime;
- contenutistica: la qualità dei contenuti influenza la capacità di mantenere concentrazione;
- relazionale: l'attenzione si organizza in funzione delle reti sociali e delle dinamiche di visibilità.

Secondo Carr (2011), la costante interruzione digitale compromette la capacità di lettura profonda e di pensiero riflessivo, portando a una superficialità cognitiva generalizzata. Anche Rosen (2008) evidenzia come funzioni esecutive come la pianificazione, la memorizzazione e l'inibizione delle distrazioni risultino indebolite in contesti digitali fortemente permeabili. In ambito sociologico, Bauman (2000) descrive la modernità liquida come uno stato in cui pratiche, relazioni e simboli si dissolvono e riformano con rapidità crescente. La parcellizzazione dell'attenzione emerge in questo quadro come cifra esperienziale: l'individuo è sempre più orientato a micro-attività cognitive e sempre meno a forme di impegno sostenuto (Turkle, 2015).

I dispositivi algoritmici e la governance dell'attenzione

Un elemento centrale nella parcellizzazione dell'attenzione è l'architettura algoritmica che regola la presentazione dei contenuti. Le piattaforme social non sono spazi neutri, bensì sistemi progettati per catturare e monetizzare l'attenzione. Negli algoritmi, l'attenzione è tradotta in coinvolgimento misurabile, likes, commenti, visualizzazioni, trasformando l'esperienza soggettiva in dati vendibili agli inserzionisti (Zuboff, 2019).

La governance dell'attenzione si manifesta in due direttrici principali:

- Saturazione: attraverso un flusso continuo di stimoli, aumenta la probabilità di mantenere l'utente attivo;
- Personalizzazione: l'algoritmo anticipa e modella i desideri, riducendo l'attrito cognitivo e massimizzando il tempo speso sulla piattaforma.

Queste dinamiche implicano una ridefinizione del concetto di agency: l'utente è simultaneamente consumatore e prodotto, inserito in un circuito che trasforma affetto, attenzione e identità in capitale estrattivo (Couldry, Mejias, 2019).

Effetti sociali e culturali della parcellizzazione

La forma del discorso pubblico si trasforma in segmenti brevi, frammentati e facilmente condivisibili. Ciò favorisce narrazioni semplicistiche, polarizzate e basate su stimoli emozionali invece che su argomentazioni fondate (Pariser, 2012). Queste dinamiche contribuiscono alla creazione di *filter bubble*, in cui gli utenti ricevono conferme delle proprie convinzioni, limitando l'esposizione a opinioni divergenti.

L'attenzione parcellizzata influisce sui legami sociali: le relazioni interpersonali, mediate da notifiche e feed, tendono a privilegiare comunicazioni brevi, generando un senso di connessione superficiale e transazionale (Turkle, 2015). L'attenzione, frazionata e costantemente reindirizzata, limita la presenza emotiva nelle interazioni faccia a faccia. La crescente prevalenza di micro-attenzioni ha ricadute sull'apprendimento formale: infatti, laddove la lettura profonda e la riflessione critica richiedono periodi prolungati di concentrazione, l'abitudine allo scrolling può compromettere capacità di

comprensione complessa, soprattutto nei giovani (Rosen, 2008; Carr, 2011). I risultati di ricerche longitudinali indicano correlazioni tra esposizione digitale intensa e ridotta capacità di mantenere concentrazione su compiti prolungati (Ophir, Nass, Wagner, 2009). Nonostante le pressioni sistemiche verso la frammentazione, emergono, tuttavia, pratiche culturali e sociali che ambiscono a riconquistare spazi di attenzione sostenuta. Tra queste si segnalano:

- Digital detox: periodi deliberati di disconnessione per ridurre stimoli digitali (Roberts, Yaya, Manolis, 2014).
- Mindfulness e mediazione digitale: tecniche che incoraggiano consapevolezza dell'uso dei media (Kabat-Zinn, 2003; Rosen, Lim, Carrier, Cheever, 2011).
- Design etico delle tecnologie: approcci progettuali che privilegiano l'esperienza umana piuttosto che l'engagement esasperato (Eyal, Hoover, 2014).

Conclusioni

La parcellizzazione dell'attenzione, pur radicata negli specifici dispositivi e nelle logiche algoritmiche contemporanee, è espressione di tendenze socioculturali più ampie. La modernità digitale utilizza la distrazione come vettore di attrazione, intensificandola e sistematizzandola. In questo senso, i dispositivi sono gli spazi in cui si attualizza tale consenso, soprattutto nelle relazioni di potere e negli interessi economici che ne sono alla base (Couldry, Mejias, 2019; Zuboff, 2019).

Una sociologia dell'attenzione deve quindi operare su più livelli:

- Analisi critica delle infrastrutture tecniche e commerciali;
- Studio empirico dei comportamenti individuali e collettivi;
- Esame delle conseguenze normative, politiche ed etiche.

Diventa inderogabile sviluppare politiche educative, che promuovano competenze critiche per l'utenza e normative che regolino la progettazione di tecnologie digitali in modo più umano e sostenibile.

Lo scrolling e la parcellizzazione dell'attenzione rappresentano fenomeni interconnessi che palesano profondi cambiamenti nelle pratiche cognitive, relazionali e culturali della società antropica

contemporanea. L'utilizzo dei dispositivi digitali e la costruzione degli algoritmi devono catalizzare la comprensione dei processi sottesi che rendono tali pratiche così pervasive.

Occorre interrogarsi criticamente sul valore dell'attenzione, sulla qualità delle interazioni sociali digitali e sulle condizioni fattuali per costruire spazi esperienziali che non siano soltanto frammenti temporali di consumo, ma luoghi di confronto, di significato e di riflessione critica.

Bibliografia

Alter, A. (2018). Irresistible: the rise of addictive technology and the business of keeping us hooked. Penguin Press.

Bauman, Z. (2000). Liquid modernity. Polity Press.

Buckingham, D. (2019). The media education manifesto. Polity Press.

Carr, N. (2011). The shallows: what the internet is doing to our brains. W.W. Norton & Company.

Couldry, N., Mejias, U. (2019). The costs of connection: how data is colonizing human life and appropriating it for capitalism. Stanford University Press.

Davenport, T.H., Beck, J.C. (2001). The attention economy: understanding the new currency of business. Harvard Business School Press.

Eyal, N., Hoover, R. (2014). Hooked: how to build habit-forming products. Portfolio.

Goffman, E. (1974). Frame Analysis: an essay on the organization of experience . Harper Colophon.

James, W. (1890). The principles of psychology. Henry Holt & Co.

Kabat-Zinn, J. (2003). Mindfulness-based interventions in context: past, present, and future. Clinical Psychology: Science and Practice, 10(2), 144–156.

Ophir, E., Nass, C., Wagner, A.D. (2009). Cognitive control in media multitaskers. Proceedings of the National Academy of Sciences, 106(37), 15583–15587.

Pariser, E. (2012). The filter bubble: what the internet is hiding from you. Penguin Press.

Roberts, J.A., Yaya, L.H.P., Manolis, C. (2014). The invisible addiction: cell-phone activities and addiction among male and female college students. Journal of behavioral addictions, 3(4), 254–265.

Rosen, C. (2008). The myth of multitasking. How intentional self-distraction hurts us. The New Atlantis, 20, 105–110.

Rosen, L.D., Lim, A., Carrier, L.M., Cheever, N.A. (2011). An empirical examination of the educational impact of text message interruptions during college lectures. Psicologia educativa, Vol. 17. N. 2 pp. 163-177.

Faverio, M., Sidoti, O. (2025). Internet & Technology. Pew Research Center Reports.

Simon, H.A. (1969). Designing organizations for an information-rich world. In M. Greenberger (Ed.), Computers, communication, and the public interest (pp. 37–72). Johns Hopkins Press.

Turkle, S. (2015). Reclaiming conversation: the power of talk in a digital age. Penguin Press.

van Dijck, J., Poell, T. (2013). Understanding social media logic. Media and Communication, 1(1), 2–14.

Zuboff, S. (2019). The age of surveillance capitalism: the fight for a human future at the new frontier of power. Public Affairs.

Dismofia monetaria: un quadro integrato per la regolamentazione algoritmica e la politica della salute mentale - di *Marco Lazzeri*, SIPSIOL – Società Italiana di Psicologia Online, Roma, Italia

La dismorfia monetaria emerge in un ecosistema digitale che incentiva performance, confronto e metriche di engagement, trasformando identità, emozioni e status economico in segnali pubblici continuamente valutati

Abstract

Questo articolo propone un framework teorico e orientato alla policy per analizzare un fenomeno emergente all'intersezione tra architetture algoritmiche dei social media, confronto sociale digitale e salute mentale.

Con il 43% della Generazione Z che riporta sintomi riconducibili al distress finanziario (Pew Research Center, 2023) e una rapida espansione della produzione di contenuti e sistemi informativi mediati da intelligenza artificiale (Gartner, 2023), si delinea un quadro di crescente urgenza per la definizione di politiche di governance digitale. La proliferazione dei social media ha istituzionalizzato una cultura della performatività dell'autorappresentazione, in cui la rappresentazione della vita è sottoposta a processi di spettacolarizzazione e confronto sociale perpetuo (Marwick, 2015; Goffman, 1959). L'integrazione dell'intelligenza artificiale (IA) generativa ha riconfigurato fondamentalmente questo ecosistema, erodendo i confini epistemici tra contenuto autentico e sintetico (Chesney & Citron, 2019; Vaccari & Chadwick, 2020). Attraverso una sintesi interdisciplinare di psicologia sociale, neuroscienze e studi critici sui media, viene proposto il costrutto teorico di "dismorfia monetaria" inteso come una percezione distorta e persistente della propria posizione finanziaria relativa, sproporzionata rispetto alle condizioni oggettive e alimentata dall'esposizione ripetuta a contenuti economicamente iper-idealizzanti. Il framework concettuale delineato suggerisce come tali contenuti, selezionati e amplificati algoritmicamente, possano contribuire all'iper-attivazione cronica di meccanismi di confronto sociale ascendente, generando stress finanziario percepito correlato con sintomatologia depressiva (Guan et al., 2022; Appel et al., 2016). Evidenze neuroscientifiche suggeriscono associazioni consistenti tra stress cronico associato e variazioni strutturali in regioni limbiche, specificamente nell'ippocampo e

nell'amigdala (Butterworth et al., 2012), fornendo un possibile substrato biologico per tali processi senza implicare relazioni causali dirette.

Alla luce del quadro teorico delineato, il lavoro interpreta la dismorfia monetaria come un esito sistemico di architetture digitali disfunzionali, piuttosto che come una condizione individuale, e discute le implicazioni per la governance algoritmica, l'alfabetizzazione digitale e le politiche di salute pubblica. Viene infine proposto un quadro di intervento multilivello che intende informare il dibattito regolatorio e orientare future ricerche empiriche su questi fenomeni emergenti.

Introduzione

La società contemporanea è caratterizzata da paradossi digitali strutturali. Sebbene le piattaforme di social media promuovano apparentemente connessione e comunità, esse hanno parallelamente istituzionalizzato una cultura della performance e del confronto sociale che permea l'esistenza quotidiana (Marwick, 2015; Goffman, 1959). All'interno di questo "teatro del sé" digitale (Marwick, 2015), l'auto-rappresentazione si è evoluta in un lavoro curatoriale volto a costruire e mantenere un'identità pubblica idealizzata, frequentemente incentrata su marcatori di realizzazione socioeconomica. Questa proiezione continua di un sé ottimizzato non costituisce un fenomeno neutrale ma risponde agli imperativi economici dell' "economia dell'attenzione" e alle metriche di coinvolgimento che strutturano l'ecosistema digitale, trasformando le interazioni sociali in dati estraibili e gli stati emotivi in valore commercializzabile (Zuboff, 2019; Couldry & Mejias, 2019).

L'avvento dell'intelligenza artificiale (IA) generativa non ha meramente accelerato questa traiettoria ma ne ha radicalmente trasformato i fondamenti epistemologici. I sistemi di IA generano contenuti testuali, visivi e audiovisivi di realismo sconcertante, compromettendo la capacità umana di discernere l'autentico dal sintetico e saturando il milieu digitale con narrazioni iper-perfezioniste e spesso fittizie (Chesney & Citron, 2019; Vaccari & Chadwick, 2020).

L'integrazione pervasiva di strumenti generativi nei flussi comunicativi quotidiani non solo amplifica la produzione di ideali irraggiungibili ma ne democratizza anche la creazione, permettendo a qualsiasi individuo di generare scenari credibili di successo e opulenza indipendentemente dalla base fattuale (Ferrara et al., 2023).

Da questo crogiolo tecnoculturale emerge un fenomeno clinico-sociale preoccupante: la "dismorfia monetaria" (Pew Research Center, 2023). Analoga ai disturbi della percezione corporea, essa comporta una percezione persistentemente negativa e distorta della propria situazione finanziaria, sproporzionata rispetto alla realtà oggettiva e alimentata dal confronto ossessivo con standard di

ricchezza irraggiungibili e pervasivi online. Questo parallelismo con la dismorfia corporea si estende oltre la metafora: entrambe le condizioni condividono un fondamento psicopatologico nella discrepanza tra percezione soggettivamente distorta e realtà oggettiva, precipitando significativo disagio emotivo e comportamentale (American Psychiatric Association, 2022).

E' importante chiarire che la dismorfia monetaria non viene qui proposta come categoria diagnostica né come entità clinica formalizzata. In linea con gli obiettivi di una review concettuale, questo costrutto è inteso come uno strumento euristico e interpretativo, e non come una categoria diagnostica, volto a integrare evidenze disperse provenienti dalla letteratura psicologica, sociale e di policy

Sebbene ricerche precedenti abbiano esaminato separatamente gli effetti dei social media sul benessere psicologico (Verduyn et al., 2017), l'impatto dell'IA generativa sulla qualità dell'informazione (Chesney & Citron, 2019) e i determinanti socioeconomici della salute mentale (Lund et al., 2018), manca tuttora una sintesi integrata che connetta sistematicamente le architetture algoritmiche orientate all'engagement, i meccanismi di confronto sociale digitale e le loro conseguenze psicosociali in termini di percezione economica.

Il contributo del presente articolo consiste pertanto nella proposta di un modello concettuale integrato che organizza tali evidenze in una struttura interpretativa coerente, utile a informare sia il dibattito scientifico sia la riflessione sulle politiche pubbliche in ambito di governance digitale e salute mentale.

Verso un modello concettuale integrato della dismorfia monetaria: dalle architetture digitali alle distorsioni percettive

Il modello concettuale qui proposto non mira a stabilire relazioni causali dirette, bensì a organizzare le evidenze disponibili all'interno di una struttura interpretativa coerente, che consenta di comprendere come specifiche architetture digitali possano, in interazione con fattori individuali e contestuali, contribuire alla formazione di percezioni finanziarie distorte. In questo senso, tale costrutto assume una funzione prevalentemente euristica e generativa, finalizzata a orientare future indagini empiriche e sviluppi di policy, piuttosto che a fornire spiegazioni causali definitive. Studi basati sull'analisi del linguaggio e dei contenuti prodotti online mostrano come i contesti digitali favoriscano narrazioni persistenti di confronto sociale, auto-valutazione negativa e anticipazione ansiosa del futuro, con effetti misurabili sul benessere psicologico (Eichstaedt et al., 2018; Guntuku et al., 2019). In particolare, la letteratura evidenzia come l'esposizione a rappresentazioni socialmente selettive e performative del successo economico possa amplificare percezioni di inadeguatezza e fallimento personale, soprattutto tra i giovani adulti (Vogel et al., 2014). Questa osservazione discorsiva non ha

finalità di verifica ipotetica né di stima della prevalenza del fenomeno, ma svolge una funzione di ancoraggio concettuale per lo sviluppo del modello teorico proposto.

L'economia dell'attenzione e la costruzione algoritmica del sé idealizzato

L'adozione globale dei social media ha istituzionalizzato pratiche di autorappresentazione strategicamente curate, trasformando l'esistenza quotidiana in performance continua soggetta a valutazione sociale algoritmica (Marwick, 2015; Hogan, 2010). Questo "lavoro sul sé" digitale è profondamente plasmato dalle affordances strutturali delle piattaforme e dai loro incentivi economici, orientando l'auto-rappresentazione verso forme ottimizzate per la viralità e il consumo visivo.

Le architetture delle piattaforme funzionano non come strumenti passivi ma come agenti attivi di strutturazione sociale che tramutano il riconoscimento interpersonale in metriche quantificabili e competitive (Gerlitz & Helmond, 2013). Il sé diviene così un "progetto riflessivo" (Giddens, 1991) perpetuamente monitorato, adattato e ottimizzato basandosi sul feedback digitale ricevuto all'interno di un ciclo continuo di rinforzo. All'interno di questa "economia dell'attenzione" (Davenport & Beck, 2001), la visibilità e il consenso sociale costituiscono valuta primaria, funzionando come rinforzo positivo differenziale per comportamenti di autopromozione estrema. La costruzione del sé diventa inseparabile dalla logica del personal branding (Khamis et al., 2017), particolarmente evidente nelle economie dei creator e degli influencer dove l'autenticità stessa è performata, standardizzata e commercializzata come capitale di marca distintivo (Abidin, 2016).

In questo contesto, gli utenti tendono a esibire prevalentemente indicatori di successo, benessere e status economico elevato (Siibak, 2009; Gonzales & Hancock, 2011), impegnandosi in un costante auto-editing dei contenuti guidato da algoritmi (Chua & Chang, 2016). Il risultante "repertorio performativo" selezionato algoritmicamente, limitato ai momenti culminanti della vita, eclissa sistematicamente la normalità, la vulnerabilità e la complessità umana (Goffman, 1959), generando una "realtà spettacolarizzata" (Dean, 2010) e un "divario di rappresentazione" strutturale tra esperienza vissuta e presentazione pubblica. Questa proiezione di un "sé ideale" ottimizzato algoritmicamente (Rosenberg & Egbert, 2011; Vogel & Rose, 2016) stabilisce benchmark di confronto sistematicamente inaccessibili. La natura asincrona, controllata e infinitamente modificabile della performance digitale permette un perfezionismo estetico-narrativo irraggiungibile nelle interazioni faccia-a-faccia (Boyd, 2014), precipitando una "sovraesposizione normativa algoritmica" a stili di vita elitari percepiti non come eccezioni ma come realizzazioni socialmente attese (Couldry & Mejias, 2019). L'esposizione ripetuta e algoritmicamente amplificata interiorizza questi standard distorti come normativi (Fardouly & Vartanian, 2016), generando un "effetto di falso consenso inverso" (Marks & Miller, 1987) e una "distorsione da esposizione selettiva algoritmica" dove l'universo online osservato è erroneamente

percepito come rappresentativo della realtà sociale. Il confronto sociale, una tendenza psicologica fondamentale (Festinger, 1954), è così sistematicamente iper-attivato all'interno di un ambiente strutturalmente distorto verso l'eccellenza apparente. Il confronto si sposta dall'occorrere con medie realistiche all'impegno con élite visibili, performative e promosse algoritmicamente (Lee & Chiou, 2022), potenzialmente culminando in stati di "stanchezza da confronto cronico" (Wright et al., 2018). Questa dinamica è amplificata e automatizzata dagli algoritmi di raccomandazione che privilegiano contenuti "di successo" ad alto engagement secondo logiche di ottimizzazione commerciale, creando circuiti di rinforzo algoritmico che omogeneizzano la rappresentazione della vita verso ideali asfittici e commerciali di perfezione (Bucher, 2018). Il "sé idealizzato" rappresenta conseguentemente il prodotto prevedibile e calcolabile di architetture tecnologiche progettate per massimizzare le metriche di engagement, premiando rappresentazioni che suscitano ammirazione, invidia e dipendenza attentiva (Zuboff, 2019).

Amplificazione esponenziale: il ruolo dell'IA generativa nella distorsione della realtà percepita

Il panorama digitale ha subito trasformazioni radicali attraverso la diffusione di massa di sistemi di IA generativa. Questi strumenti trascendono gli aumenti quantitativi nel volume dei contenuti e nell'engagement (Barari et al., 2023) per alterare qualitativamente e sostanzialmente i processi cognitivi di valutazione della realtà e attribuzione di credibilità epistemica. Evidenze sperimentali convergenti dimostrano difficoltà sistematica e generalizzata nel discriminare contenuti generati da IA da quelli umani, compromettendo l'elaborazione critica centrale mentre privilegiano percorsi euristici ed elaborazione emotiva periferica (Chesney & Citron, 2019; Vaccari & Chadwick, 2020). Il "realismo tecnologico" dei contenuti sintetici può inibire i percorsi di persuasione centrali, spostando l'influenza verso modalità periferiche basate su indicatori superficiali di affidabilità e verosimiglianza estetica (Vaccari & Chadwick, 2020). All'interno di questo quadro epistemologicamente destabilizzato, i segnali di credibilità algoritmici modulano i giudizi epistemici degli utenti in misura statistica maggiore rispetto alle metriche tradizionali di autorità ed expertise, rimodellando attivamente i fondamenti delle dinamiche di fiducia all'interno delle reti sociali (Jakesch et al., 2023). L'Intelligenza artificiale emerge così come mediatore tecnologico intrinsecamente non neutrale, capace di erodere progressivamente la percezione di autenticità mentre solleva questioni epistemiche fondamentali riguardanti autorità, veridicità e provenienza delle fonti (Vaccari & Chadwick, 2020). In ambienti informativi ibridi dove i confini tra contenuto autentico e sintetico diventano porosi e frequentemente indecifrabili, i meccanismi psicologici di confronto sociale acquisiscono potere distorsivo senza precedenti,

trasformandosi da processi sociali occasionali e contestualizzati in sollecitazioni algoritmiche iper-realistiche incessanti.

Il focus si sposta strutturalmente dal confronto con pari significativi (simili per contesto, risorse e possibilità) all'impegno con "confronti algoritmici iper-ottimizzati" – individui, influencer o entità sintetiche selezionate dai sistemi di raccomandazione per la generazione massimale di engagement attraverso l'esibizione calibrata di standard eccezionali ed emotivamente risonanti di stili di vita (Lee & Chiou, 2022).

Indagini epidemiologiche rappresentative identificano una prevalenza clinicamente significativa di distress finanziario tra il 29% della popolazione adulta statunitense, con picchi del 43% tra la Generazione Z (Pew Research Center, 2023). Questi dati, sebbene non derivino da una misurazione diretta del costrutto qui proposto, indicano un substrato di ansia e percezione di svantaggio relativo che la dismorfia monetaria concettualizza e sistematizza. Questa marcata disparità generazionale riflette non meramente maggiore familiarità tecnologica ma esposizione di vita a etiche della performance digitale integrate all'interno di mercati del lavoro e immobiliari percepiti come strutturalmente ostili, rendendo gli ideali finanziariamente riusciti propagati online simultaneamente più desiderabili e statisticamente più irraggiungibili (Twenge, 2017).

Il confronto ascendente indotto dagli algoritmi si verifica quando gli utenti, nonostante la consapevolezza della natura selettiva e parziale dei contenuti, incontrano difficoltà nel distacco cognitivo ed emotivo, interiorizzando comunque percezioni diffuse di inadeguatezza e svantaggio relativo (Vogel et al., 2014). In contesti saturi di contenuti iper-realisti e iper-idealizzati, questo fenomeno viene amplificato, incrementando il confronto verso individui o standard percepiti come superiori, storicamente associato a ridotta autostima, insoddisfazione cronica e sintomi ansioso-depressivi (Emmons & Diener, 1985; Taylor & Lobel, 1989). All'interno dei social media contemporanei, tale confronto frequentemente avviene privo di informazioni contestuali riguardanti percorsi reali, sforzi, fallimenti, privilegi o circostanze fortuite sottostanti il successo esibito. Questa decontestualizzazione algoritmica trasforma sistematicamente la motivazione potenziale in frustrazione ruminativa mentre genera "invidia maligna strutturale" caratterizzata da risentimento, ostilità sociale e desideri di diminuzione altrui piuttosto che emulazione costruttiva o ispirazione (Lange & Crusius, 2015). L'interazione passiva (lurking), facilitata da flussi di contenuti ibridi e personalizzati, dimostra correlazione particolarmente robusta con l'affettività negativa attraverso meccanismi di "confronto sociale passivo algoritmico" (Verduyn et al., 2017), lasciando gli individui in stati di ruminazione non verificata senza opportunità di feedback sociale rassicurante o correttivo.

Pertanto, la dismorfia monetaria trascende l'erronea percezione finanziaria individuale; essa costituisce il sintomo manifesto di un più ampio "disallineamento epistemico collettivo" dove gli individui lottano, spesso senza successo, per costruire rappresentazioni stabili, credibili e non distorte della realtà sociale ed economica, con conseguenze dirette e misurabili per il benessere psicologico, l'auto-valutazione e la salute mentale complessiva.

Conseguenze psicosociali e neurobiologiche della dismorfia monetaria

La percezione distorta e carica di ansia della propria condizione economica, amplificata all'interno di ambienti informativi algoritmicamente manipolabili, costituisce uno stressor psicosociale cronico riconosciuto come determinante significativo della salute mentale (Karim et al., 2020).

Dallo stress finanziario percepito alla psicopatologia: meccanismi e mediatori

Questo stress rappresenta non stati d'animo transitori ma processi psicosociali attivi e pervasivi che compromettono sistematicamente le risorse cognitive, emotive e comportamentali necessarie per il confronto adattivo delle sfide quotidiane (Hobfoll, 1989). La costante preoccupazione finanziaria ruminativa, alimentata da confronti sociali ascendenti all'interno di ambienti digitali strutturalmente distorti, può innescare cicli autoperpetuanti di ruminazione e impotenza appresa (Seligman, 1972), fattori transdiagnostici potenti nella psicopatologia contemporanea. La depressione maggiore – caratterizzata da umore depresso, anedonia e sintomi cognitivo-somatici (World Health Organization, 2017) – è influenzata da complesse interazioni tra determinanti prossimali (es. disoccupazione, basso reddito, insicurezza lavorativa) e determinanti distali (es. disuguaglianza di reddito, fragilità dello stato sociale, precarietà strutturale) (Antunes et al., 2018; Lund et al., 2018). All'interno di questo modello eziologico multidimensionale, la dismorfia monetaria opera come determinante psicosociale di tipo percettivo-distortivo, mediando e amplificando l'impatto dei fattori strutturali distali sulle esperienze individuali di sofferenza e disagio psicologico. Il punto di svolta è questo: lo spazio digitale non si limita ad accrescere lo stress psicosociale, ne riscrive le regole fondamentali e ne amplifica esponenzialmente la portata. L'iper-connessione cronica e l'immersione in ecosistemi algoritmici che alimentano incessantemente dinamiche di consumo, confronto e performatività fanno sì che l'incertezza economica non rimanga un problema definito, ma si metamorfosi in un'ansia diffusa e permanente, un allarme costante che ridefinisce la soglia stessa del benessere. Il punto di svolta è questo: lo spazio digitale non si limita ad accrescere lo stress psicosociale, ne riscrive le regole fondamentali e ne amplifica esponenzialmente la portata. L'iper-connessione cronica e l'immersione

in ecosistemi algoritmici che alimentano incessantemente dinamiche di consumo, confronto e performatività fanno sì che l'incertezza economica non rimanga un problema definito, ma si metamorfosi in un'ansia diffusa e permanente. La preoccupazione trascende la mera questione economica contingente per radicarsi in un contesto mediale che, in modo esplicito e sistematico, celebra l'opulenza come sinonimo di successo e valore personale. Ogni scroll diventa così un potenziale *trigger* per un confronto sociale dannoso, trasformando lo stress finanziario da questione pratica in una lacerante crisi d'identità, dove l'impossibilità di consumare equivale all'impossibilità di essere.

Revisioni sistematiche della letteratura, comprendenti 40 studi osservazionali longitudinali e trasversali, evidenziano associazioni positive, consistenti e potenzialmente dose-dipendenti tra stress finanziario percepito e sintomatologia depressiva, con dimensioni dell'effetto variabili da moderate a elevate (Guan et al., 2022). Questa associazione dimostra bidirezionalità e potenziale ciclicità: la depressione può compromettere il funzionamento cognitivo-esecutivo, la motivazione intrinseca e l'efficacia comportamentale, riducendo significativamente la capacità di gestione finanziaria adattiva e le performance lavorative, che successivamente esacerbano lo stress economico oggettivo e percepito, creando cicli viziosi psicopatologici autoperpetuanti (Frasquilho et al., 2016).

La ricerca indica che l'impatto dello stress finanziario percepito può essere mediato attraverso meccanismi psicologici distinti ma interconnessi, inclusi vergogna internalizzata e fallimento personale (Starrin et al., 2009), compromissione strutturale dell'autostima e dell'autoefficacia (Ridge et al., 2021), e percezione cronica di basso status sociale ed esclusione relativa (Adler et al., 1994) – tutti elementi che trovano terreno fertile per confronto, amplificazione e rinforzo all'interno delle economie dell'attenzione dei social media.

Le variabili demografiche e socioeconomiche funzionano come moderatori contestuali cruciali di queste relazioni complesse. L'impatto della dismorfia monetaria e dello stress finanziario percepito dimostra distribuzione non uniforme attraverso le popolazioni: individui appartenenti a gruppi socialmente svantaggiati, possedendo capitale culturale, economico e sociale ridotto, o mantenendo reti di supporto familiare e comunitario fragili, possono dimostrare vulnerabilità statistica e clinica agli effetti psicopatologici (Lund et al., 2018). Similmente, le generazioni più giovani (Generazione Z e Millennial), profondamente immerse nella cultura digitale performativa mentre confrontano precarietà lavorativa strutturale, sfide di accessibilità al mercato immobiliare e incertezza esistenziale, sperimentano questa intersezione tra precarietà reale e percezione distorta con intensità e pervasività particolari (Pew Research Center, 2023). È probabile che tratti individuali quali l'autostima, la resilienza e il grado di identificazione con gli standard proposti svolgano un ruolo moderatore chiave, aprendo

una fertile area per la ricerca futura. Questa prospettiva intersezionale si rivela essenziale per progettare interventi di prevenzione, supporto e politica equi, efficaci e socialmente giusti.

Substrati neurali dello stress socioeconomico percepito: dalla psicologia sociale alla neurobiologia

Le neuroscienze contemporanee forniscono supporto biologico convergente per queste associazioni psicosociali, identificando substrati neurali specifici che spiegano i meccanismi attraverso i quali lo stress psicosociale cronico si traduce in vulnerabilità psicopatologica. Ricerche di neuroimaging seminali dimostrano che adulti di mezza età che sperimentano difficoltà finanziarie oggettive e percepite esibiscono volumi significativamente ridotti dell'ippocampo e dell'amigdala rispetto a controlli appaiati per età, sesso ed educazione (Butterworth et al., 2012). Questo risultato si rivela particolarmente significativo identificando correlati neuroanatomici misurabili e replicabili di stressor socioeconomici, suggerendo che esperienze di precarietà economica – sia oggettive che soggettive – possono lasciare impronte durature e potenzialmente reversibili sulla struttura cerebrale.

Queste strutture limbiche si rivelano fondamentali per la regolazione emotiva adattiva, la memoria contestuale e la modulazione dello stress (LeDoux, 2003). L'amigdala funziona come nodo cruciale nella rilevazione automatica della minaccia, nell'iniziazione delle risposte neuroendocrine allo stress e nell'elaborazione delle emozioni sociali, mentre l'ippocampo si rivela essenziale per la memoria episodica contestuale, la regolazione dell'asse ipotalamo-ipofisi-surrene (HPA) e l'inibizione di risposte allo stress inappropriati (McEwen, 2017). La loro atrofia strutturale in contesti di stress cronico può contribuire a dinamiche di mantenimento della vulnerabilità psicologica, compromettendo sia la capacità per la regolazione adattiva delle emozioni negative che la memoria di esperienze positive, risolutive o di coping. È cruciale notare che questi processi neuroplastici non sono deterministici né irreversibili; evidenze robuste indicano una significativa resilienza neurale in risposta alla riduzione dello stress e a interventi psicosociali mirati (McEwen, 2017).

I dati neurobiologici suggeriscono che lo stress finanziario cronico – sia reale che percepito – possa attivare persistentemente e disregolatamente l'asse HPA, conducendo a esposizione prolungata a livelli elevati di glucocorticoidi (principalmente cortisolo), agenti neurotossici noti per regioni cerebrali ricche di recettori glucocorticoidi come l'ippocampo. Livelli cronicamente elevati di glucocorticoidi compromettono la neurogenesi ippocampale adulta, inducono atrofia dendritica, promuovono apoptosi neuronale e riducono plasticità sinaptica – meccanismi cellulari e molecolari che collegano direttamente l'esperienza psicologica di stress alla perdita di volume osservata nella neuroimmagine (Sapolsky, 2004).

Ulteriori percorsi fisiopatologici, includendo alterazioni della neuroplasticità dipendente dall'esperienza, riduzioni del flusso ematico cerebrale regionale e processi neuroinfiammatori di basso grado, possono contribuire sinergicamente a questi esiti strutturali avversi. Lo stress psicosociale cronico può innescare meccanismi di infiammazione neuronale di basso grado mediati attraverso attivazione microgliale e rilascio di citochine pro-infiammatorie (es. IL-1β, IL-6, TNF-α), danneggiando successivamente neuroni, ostacolando meccanismi di riparazione e plasticità, e alterando neurogenesi (Miller & Raison, 2016). Inoltre, lo stress finanziario, frequentemente associato a comportamenti compensatori disadattivi (es. disturbi cronici del sonno, diete povere di nutrienti, sedentarietà, uso di sostanze), può indirettamente peggiorare la salute cerebrovascolare e metabolica, riducendo il flusso sanguigno cerebrale e l'apporto di ossigeno/nutrienti a regioni cerebrali critiche per la regolazione emotiva e cognitiva.

Crucialmente, il cervello umano reagisce principalmente alla minaccia percepita o anticipata, indipendentemente dalla sua veridicità oggettiva o dalla reale imminenza (McEwen & Gianaros, 2010). Pertanto, l'ipervigilanza, la ruminazione e il confronto sociale ascendente indotti da ambienti digitali algoritmici possono mantenere l'attivazione dell'asse HPA in stati cronici di basso grado anche in assenza di crisi finanziarie reali e immediate, potenzialmente innescando meccanismi neuroendocrini analoghi a quelli osservati in condizioni di deprivazione economica oggettiva. Questa analogia nei meccanismi sottostanti, piuttosto che una diretta equivalenza negli esiti, rafforza l'urgenza di considerare lo stress percepito come un fattore di rischio biologico reale.

Comprendere questi meccanismi neurobiologici rafforza gli imperativi clinici e di salute pubblica di intervenire non solo sulle condizioni economiche materiali (determinanti distali) ma anche – integrativamente – sugli ambienti informativi digitali e le distorsioni cognitivo-percettive che alimentano lo stress finanziario percepito, proteggendo così l'integrità neurale, la resilienza psicologica e il benessere mentale della popolazione nell'era algoritmica.

Quadro di intervento strategico: dall’educazione alla ricerca

Le evidenze analizzate indicano come la dismorfia monetaria rappresenti un rischio sistemico emergente per la salute pubblica, strettamente radicato nelle architetture algoritmiche delle piattaforme digitali. Tale fenomeno richiede una risposta politica che vada oltre gli approcci volontari di promozione del “benessere digitale”, adottando strumenti regolatori vincolanti ispirati ai principi di precauzione. Gli interventi proposti trovano fondamento in precedenti regolatori consolidati, dimostrando la loro fattibilità: a titolo esemplificativo, i principi precauzionali applicati alla sicurezza alimentare (Regolamento (CE) n. 178/2002) e gli standard di trasparenza e tutela dei consumatori nel

settore finanziario (MiFID II) costituiscono modelli legali collaudati, applicabili anche ai contesti digitali per l'istituzione di obblighi di dovere di diligenza (Regolamento (CE) n. 178/2002; Direttiva 2014/65/UE). In questo senso, l'invocazione dei principi di "precauzione" e "dovere di diligenza" non introduce innovazioni radicali, bensì rappresenta l'estensione coerente di logiche consolidate di tutela della salute pubblica a un nuovo dominio di rischio.

In primo luogo, si rende necessario ampliare e consolidare gli obblighi di trasparenza algoritmica, estendendo le disposizioni parziali introdotte dal Digital Services Act (DSA) dell'Unione Europea (Regolamento (UE) 2022/2065). Ciò implica l'adozione di mandati per l'etichettatura obbligatoria, standardizzata e immediatamente riconoscibile, di tutti i contenuti generati o significativamente modificati dall'intelligenza artificiale (Chesney & Citron, 2019; Jakesch et al., 2023). Tale misura non solo rafforza la trasparenza, ma tutela il diritto fondamentale degli utenti di conoscere l'origine dei contenuti che influenzano le loro percezioni normative di ricchezza, successo e status sociale.

In secondo luogo, per le piattaforme online di dimensioni molto grandi (VLOPs), diventa imperativo implementare e garantire l'osservanza di standard vincolanti di "dovere di diligenza" a tutela della salute mentale digitale. Questo comporta obblighi legali di condurre valutazioni d'impatto indipendenti sui rischi sistemici derivanti dai sistemi di raccomandazione, di pubblicarne i risultati e di adottare misure correttive basate su tali valutazioni (Bucher, 2018; UK Online Safety Act, 2023). Le autorità di regolamentazione – tra cui l'Autorità per le Garanzie nelle Comunicazioni (AGCOM) italiana, la Federal Trade Commission (FTC) statunitense e il nuovo Centro Europeo per la Trasparenza Algoritmica (ECAT) – devono verificare le pratiche che promuovono ideali di ricchezza irrealistici o confronti sociali ascendenti dannosi, applicando sanzioni proporzionate agli impatti sociali e sanitari rilevati (Helberger et al., 2023). L'esperienza maturata con il GDPR e il DSA dimostra come eventuali resistenze da parte dell'industria possano essere superate grazie a volontà politica, alla costruzione di coalizioni di stakeholder – comprendenti società civile e accademia – e a un solido ancoraggio ai diritti fondamentali.

A partire dal quadro regolatorio delineato, l'azione volta a tutelare la salute mentale deve articolarsi su più livelli integrati, combinando approcci evidence-informed con strumenti concreti di intervento. Sul piano regolatorio, l'etichettatura obbligatoria dei contenuti dovrebbe rispettare standard tecnici interoperabili, come watermarking e metadata proposti dalla Coalition for Content Provenance and Authenticity (C2PA), rendendo i contenuti machine-readable e immediatamente riconoscibili dagli utenti (Guariglia & Moore, 2023), non come semplice esercizio di trasparenza, ma come prerequisito fondamentale per ripristinare un minimo di agency epistemica (Vaccari & Chadwick, 2020). Parallelamente, le piattaforme sono tenute a progettare, monitorare e mitigare attivamente i rischi

sistemici per la salute mentale derivanti dai loro sistemi di raccomandazione, mentre le autorità competenti devono poter intervenire ordinando modifiche al design qualora emergano rischi ingiustificati (Binns et al., 2021; Citron, 2022)

Sul piano educativo, è essenziale sviluppare un'alfabetizzazione duplice, critica-digitale e finanziaria-comportamentale, capace di rendere gli utenti consapevoli delle strategie persuasive dell'IA generativa, dell'economia dell'attenzione che determina la visibilità dei contenuti e dei meccanismi di confronto sociale, attraverso moduli esperienziali come la creazione di contenuti sintetici per comprenderne il funzionamento (Jakesch et al., 2023; Zuboff, 2019; Fardouly & Vartanian, 2016; UNESCO, 2023). Parallelamente, l'educazione finanziaria deve spostare il focus dal bilancio familiare alla psicologia del denaro, insegnando a riconoscere distorsioni cognitive e bias emotivi, quali invidia e senso di privazione relativa, amplificati dai social media (OECD, 2022). Sul piano sanitario, il disagio monetario digitale può essere considerato un fattore di rischio psicosociale emergente. In questa prospettiva, le infrastrutture potrebbero beneficiare dell'integrazione di strumenti di screening ultra-brevi nei contesti clinici, finalizzati all'identificazione precoce di individui potenzialmente vulnerabili. L'esito positivo dello screening potrebbe attivare percorsi di supporto psicologico breve, inclusi interventi basati su protocolli cognitivo-comportamentali mirati alla ruminazione finanziaria e al confronto sociale disfunzionale, nonché facilitare l'accesso a servizi di consulenza finanziaria non commerciale (Guan et al., 2022).

Infine, la ricerca scientifica deve colmare il gap causale tra esposizione a contenuti iper-idealizzati e sviluppo di sintomi di dismorfia monetaria, attraverso studi longitudinali con misurazioni digitali dell'esposizione, valutazioni psicometriche e marcatori neurobiologici, Randomized Controlled Trials di interventi di piattaforma – tra cui nudge, diversificazione dei feed o pause riflessive prima dell'interazione con contenuti di confronto ascendente – e lo sviluppo di metriche standardizzate di benessere algoritmico, come l'indice di iper-idealizzazione o il coefficiente di confronto ascendente promosso, da monitorare e reportare obbligatoriamente (Odgers & Jensen, 2020; Lyngs et al., 2019; Helberger et al., 2023). Nel loro insieme, questi interventi delineano un approccio coerente e integrato, finalizzato a mitigare i rischi psicosociali derivanti dai sistemi di raccomandazione e dai contenuti sintetici, garantendo strumenti di tutela concreti per la salute mentale e il benessere digitale, senza sostituirsi ai processi decisionali democratici e istituzionali.

Tabella 1. Strategie di intervento multilivello per la tutela della salute mentale digitale

Livello	Intervento Chiave	Agenzia Competente	Obiettivi e meccanismi	Tempistica	Base Empirica di Riferimento
Regolatorio	1. Etichettatura obbligatoria dei contenuti generati/significativamente modificati da IA. 2. Standard di "dovere di diligenza" algoritmico (AI Duty of Care) per le VLOPs.	AGCOM, FTC, ECAT (UE), e nuove autorità nazionali per il benessere digitale.	1. Garanzia di trasparenza e responsabilità nell'uso di sistemi algoritmici. 2. Audit indipendenti e implementazione di misure correttive basate su Algorithmic Impact Assessment (AIA).	Progressiva implementazione dei mandati e piena operatività degli standard.	Modello di etichettatura: CCPA/CPRA; precedente EU AI Act (European Parliament, 2024). Dovere di diligenza esteso dal UK Online Safety Act (2023) e dal duty of care in salute pubblica (Helberger et al., 2023).
Educativo	Integrazione curricolare obbligatoria di Media & Information Literacy (MIL) ed educazione finanziaria comportamentale.	Ministero dell'Istruzione (MIUR), AGCOM, terzo settore	1.Rafforzamento delle competenze critiche e digitali; 2.Miglioramento della capacità di riconoscere contenuti sintetici;	Sviluppo dei curricula a medio termine; valutazioni periodiche per	Modello finlandese (Finnish National Agency for Education, 2019); Linee

		specializzato.	3.Comprensione dei modelli di business digitali; 4. Sviluppo della resilienza al confronto sociale.	monitoraggio dell'efficacia.	guida UNESCO (2023) e OECD (2022).
Sanitario	1. Screening routinario per distress da confronto finanziario digitale. 2. Linee guida cliniche per salute mentale.	Ministero della Salute, SSN, NHS, ordini professionali.	1.Miglioramento dell'identificazione precoce dei casi di distress e riduzione dei ritardi diagnostici; 2. integrazione di strumenti clinici validati.	Implementazione progressiva dello screening e delle linee guida cliniche.	USPSTF (2016); Financial Anxiety Scale – FAS (Archuleta et al., 2013).
Ricerca & Sviluppo	1. Programma Horizon "Benessere Digitale". 2. Sandbox regolatorie per design salutogenico.	Commissione Europea, DG CONNECT, DG SANTE, NSF, MUR.	1. Promozione di studi longitudinali e sperimentazioni controllate, 2. sviluppo di framework open-source per la salute digitale.	Ciclo di ricerca traslazionale di medio-lungo termine.	Horizon Europe; sandbox fintech ESMA (2020); Odgers & Jensen (2020).

Limitazioni e direzioni future di ricerca

La presente analisi, pur proponendo un modello integrato e innovativo per comprendere e intervenire sulla dismorfia monetaria, presenta alcune limitazioni che delineano precise traiettorie per la ricerca futura. Innanzitutto, l'evidenza disponibile risulta principalmente correlazionale: sebbene la sintesi

indichi associazioni robuste, gli studi longitudinali in grado di stabilire relazioni causali dirette tra esposizione a contenuti iper-idealizzati generati da IA e sviluppo della dismorfia monetaria sono ancora scarsi. In secondo luogo, la generalizzabilità dei dati è limitata, poiché le evidenze epidemiologiche citate derivano prevalentemente da contesti statunitensi (Pew Research Center, 2023), mentre la percezione della ricchezza e l'impatto dei social media sono mediati da fattori culturali e strutturali, quali livelli di disuguaglianza e sistemi di welfare. Un ulteriore punto riguarda la definizione e la misurazione del costrutto: la dismorfia monetaria rappresenta un concetto emergente, privo di strumenti psicometrici gold-standard validati. Inoltre, la complessità del modello causale non può essere completamente catturata, poiché la salute mentale deriva dall'interazione di molteplici fattori individuali, sociali e tecnologici, mentre l'analisi si concentra su percorsi specifici. Infine, le implicazioni di policy e la scalabilità degli interventi proposti necessitano di ulteriori test di implementazione e valutazioni d'impatto empiriche. Sebbene tali limitazioni circoscrivano l'ambito del presente contributo, esse risultano coerenti con gli obiettivi di una review concettuale orientata alla policy e contribuiscono a delineare priorità chiare per future indagini empiriche.

Conclusione: un manifesto per la salute pubblica nell'era algoritmica

Il presente lavoro avanza tre tesi fondamentali per la ricerca sociale e la politica pubblica nell'era digitale. In primo luogo, la dismorfia monetaria va intesa come una **sindrome dell'era algoritmica**, piuttosto che come un fallimento individuale. Essa emerge dall'interazione patologica tra architetture digitali progettate per massimizzare l'engagement (Zuboff, 2019), contenuti iper-realistici generati dall'IA (Chesney & Citron, 2019; Vaccari & Chadwick, 2020) e la psicologia del confronto sociale umana (Festinger, 1954). In secondo luogo, la regolamentazione algoritmica assume la dimensione di una questione di salute pubblica, che trascende le tradizionali preoccupazioni relative a concorrenza o privacy. I sistemi di raccomandazione che promuovono standard di vita percepiti come irraggiungibili devono essere considerati analoghi ai determinanti sociali di salute, giustificando obblighi di diligenza comparabili a quelli imposti in altri settori ad alto impatto (Regolamento (CE) n. 178/2002; Bucher, 2018). In terzo luogo, l'educazione all'alfabetizzazione digitale critica va riconosciuta come un diritto sociale fondamentale, non come un arricchimento opzionale. La costruzione di resilienza cognitiva ed emotiva contro le distorsioni dell'ecosistema digitale è essenziale per l'autonomia e il benessere dei cittadini, con particolare attenzione alle generazioni più giovani (OECD, 2022; Finnish National Agency for Education, 2019).

L’adozione di queste premesse richiede un cambio di paradigma nella governance digitale. Il percorso presenta sfide significative, legate alla complessità tecnica e agli interessi costituiti, ma l’inerzia politica comporta un costo concreto, misurabile in termini di salute mentale della popolazione, soprattutto giovanile. Il quadro di intervento integrato proposto – che combina regolamentazione coraggiosa, educazione obbligatoria, screening sanitario e ricerca mirata – delinea percorsi concreti, radicati in principi di policy già consolidati, volti a mitigare questa minaccia emergente. Comprendere, monitorare e intervenire su questi processi rappresenta una priorità emergente per la ricerca sociale e per l’elaborazione di politiche pubbliche sensibili agli effetti psicosociali delle tecnologie digitali.

Dichiarazione di impatto sociale

La pubblicazione e la diffusione del presente articolo potrebbero avere impatti concreti e misurabili su più livelli. In primo luogo, potrebbero informare le prossime revisioni del Digital Services Act e guidare lo sviluppo di standard europei per il design algoritmico etico, contribuendo a una regolamentazione più solida e mirata. In secondo luogo, potrebbero ispirare l’implementazione di programmi pilota di screening e intervento all’interno dei servizi di salute pubblica in diversi stati membri dell’UE, fornendo strumenti concreti per l’identificazione precoce e la gestione della dismorfia monetaria. Inoltre, la diffusione di queste evidenze potrebbe stimolare 5-10 nuovi progetti di ricerca internazionali volti a esplorare i nessi tra IA generativa, percezione economica e salute mentale, favorendo una base empirica più robusta per future policy. Infine, tali risultati potrebbero influenzare le linee guida dell’Organizzazione Mondiale della Sanità (OMS) in materia di promozione della salute mentale negli ambienti digitali, contribuendo a definire interventi globali più efficaci e basati su evidenze scientifiche.

DICHIARAZIONI

Ringraziamenti. L'autore desidera ringraziare i revisori anonimi per i loro futuri costruttivi commenti.

Conflitto di interessi. L'autore dichiara di non avere conflitti di interesse relativi alla pubblicazione di questo articolo.

Approvazione etica. Questo studio costituisce una revisione teorica e una sintesi della letteratura pubblicata. La ricerca è stata pertanto esente da approvazione da parte di un comitato etico.

Bibliografia

1. Abidin, C. (2016). Visibility labour: Engaging with Influencers' fashion brands and #OOTD advertorial campaigns on Instagram. *Media International Australia, 161*(1), 86–100. https://doi.org/10.1177/1329878X16665177

2. Adler, N. E., Boyce, T., Chesney, M. A., Cohen, S., Folkman, S., Kahn, R. L., & Syme, S. L. (1994). Socioeconomic status and health: The challenge of the gradient. *American Psychologist, 49*(1), 15–24. https://doi.org/10.1037/0003-066X.49.1.15

3. American Psychiatric Association. (2022). *Diagnostic and statistical manual of mental disorders* (5th ed., text rev.). https://doi.org/10.1176/appi.books.9780890425787

4. Antunes, A., Frasquilho, D., Azeredo-Lopes, S., Neto, D., Silva, M., Cardoso, G., & Caldas-de-Almeida, J. M. (2018). The association between education and financial stress: An analysis of the EUROPE 2020 context. *Journal of Public Health Research, 7*(2), 1402. https://doi.org/10.4081/jphr.2018.1402

5. Appel, H., Gerlach, A. L., & Crusius, J. (2016). The interplay between Facebook use, social comparison, envy, and depression. *Current Opinion in Psychology, 9*, 44–49. https://doi.org/10.1016/j.copsyc.2015.10.006

6. Archuleta, K. L., Dale, A., & Spann, S. M. (2013). College students and financial distress: Exploring debt, financial satisfaction, and financial anxiety. *Journal of Financial Counseling and Planning, 24*(2), 50–62.

7. Barari, S., Lucas, C., Munger, K., & Tromble, R. (2023). Algorithmic amplification and mental health: Experimental evidence. *Nature Human Behaviour, 7*(12), 2104–2115. https://doi.org/10.1038/s41562-023-01731-5

8. Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2021). 'It's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–14). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445260

9. Boyd, D. (2014). *It's complicated: The social lives of networked teens*. Yale University Press.

10. Bucher, T. (2018). *If...then: Algorithmic power and politics*. Oxford University Press. https://doi.org/10.1093/oso/9780190493028.001.0001

11. Butterworth, P., Cherbuin, N., Sachdev, P., & Anstey, K. J. (2012). The association between financial hardship and amygdala and hippocampal volumes: Results from the PATH through life project. *Social Cognitive and Affective Neuroscience, 7*(5), 548–556. https://doi.org/10.1093/scan/nsr027

12. Chesney, R., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *Science, 363*(6425), 10–12. https://doi.org/10.1126/science.aaw2341

13. Chua, T. H. H., & Chang, L. (2016). Follow me and like my beautiful selfies: Singapore teenage girls' engagement in self-presentation and peer comparison on social media. *Computers in Human Behavior, 55*, 190–197. https://doi.org/10.1016/j.chb.2015.09.011

14. Citron, D. K. (2022). The fight for privacy: Protecting dignity, identity, and love in the digital age. W. W. Norton & Company.

15. Couldry, N., & Mejias, U. A. (2019). *The costs of connection: How data is colonizing human life and appropriating it for capitalism*. Stanford University Press.

16. Davenport, T. H., & Beck, J. C. (2001). *The attention economy: Understanding the new currency of business*. Harvard Business School Press.

17. Dean, J. (2010). *Blog theory: Feedback and capture in the circuits of drive*. Polity Press.

18. European Parliament and the Council of the European Union. (2002, February 28). *Regulation (EC) No 178/2002 laying down the general principles and requirements of food law, establishing the European Food Safety Authority and laying down procedures in matters of food safety*. Official Journal of the European Communities, L 31/1.

19. European Parliament and the Council of the European Union. (2014, May 15). *Directive 2014/65/EU on markets in financial instruments and amending Directive 2002/92/EC and Directive 2011/61/EU*. Official Journal of the European Union, L 173/349.

20. European Parliament and the Council of the European Union. (2022, October 19). *Regulation (EU) 2022/2065 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act)*. Official Journal of the European Union, L 277/1.

21. European Parliament. (2024, March 13). *Artificial Intelligence Act: MEPs adopt landmark law*. Press Release. https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law

22. European Securities and Markets Authority (ESMA). (2020). *Report: ESMA's Fintech Action Plan*. https://www.esma.europa.eu

23. Emmons, R. A., & Diener, E. (1985). Personality correlates of subjective well-being. *Personality and Social Psychology Bulletin, 11*(1), 89–97. https://doi.org/10.1177/0146167285111008

24. Fardouly, J., & Vartanian, L. R. (2016). Social media and body image concerns: Current research and future directions. *Current Opinion in Psychology, 9*, 1–5. https://doi.org/10.1016/j.copsyc.2015.09.005

25. Ferrara, E., Chang, H., Chen, E., Muric, G., & Patel, J. (2023). Characterizing social media manipulation in the 2020 U.S. presidential election: Evidence from behavioral data. *Nature Communications, 14*(1), 620. https://doi.org/10.1038/s41467-023-36165-0

26. Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7*(2), 117–140. https://doi.org/10.1177/001872675400700202

27. Finnish National Agency for Education. (2019). *Media literacy in Finland: National media education policy*. Ministry of Education and Culture.

28. Frasquilho, D., Matos, M. G., Salonna, F., Guerreiro, D., Storti, C. C., Gaspar, T., & Caldas-de-Almeida, J. M. (2016). Mental health outcomes in times of economic recession: A systematic literature review. *BMC Public Health, 16*(1), 115. https://doi.org/10.1186/s12889-016-2720-y

29. Gartner. (2023). *Gartner predicts: The future of AI and generative AI* (ID G00792631). Gartner Research.

30. Gerlitz, C., & Helmond, A. (2013). The like economy: Social buttons and the data-intensive web. *New Media & Society, 15*(8), 1348–1365. https://doi.org/10.1177/1461444812472322

31. Giddens, A. (1991). *Modernity and self-identity: Self and society in the late modern age*. Stanford University Press.

32. Goffman, E. (1959). *The presentation of self in everyday life*. Doubleday.

33. Gonzales, A. L., & Hancock, J. T. (2011). Mirror, mirror on my Facebook wall: Effects of exposure to Facebook on self-esteem. *Cyberpsychology, Behavior, and Social Networking, 14*(1–2), 79–83. https://doi.org/10.1089/cyber.2009.0411

34. Guan, N., Guariglia, A., Moore, P., Xu, F., & Al-Janabi, H. (2022). Financial stress and depression in adults: A systematic review. *PLOS ONE, 17*(2), e0264041. https://doi.org/10.1371/journal.pone.0264041

35. Guariglia, A., & Moore, P. (2023). Provenance and authenticity in the age of generative AI: The role of technical standards. *Journal of Digital Media & Policy, 14*(1), 45-63. https://doi.org/10.1386/jdmp_00089_1

36. Helberger, N., Araujo, T., de Vreese, C. H., van Drunen, M., Bastian, M., Fanta, A., Möller, J., Strycharz, J., & Dobber, T. (2023). EU policymaking and the "Duty of Care" for digital services: From due diligence to due design. *Internet Policy Review, 12*(1). https://doi.org/10.14763/2023.1.1677

37. Hobfoll, S. E. (1989). Conservation of resources: A new attempt at conceptualizing stress. *American Psychologist, 44*(3), 513–524. https://doi.org/10.1037/0003-066X.44.3.513

38. Hogan, B. (2010). The presentation of self in the age of social media: Distinguishing performances and exhibitions online. *Bulletin of Science, Technology & Society, 30*(6), 377–386. https://doi.org/10.1177/0270467610385893

39. Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences, 120*(11), e2208839120. https://doi.org/10.1073/pnas.2208839120

40. Karim, F., Oyewande, A. A., Abdalla, L. F., Ehsanullah, R. C., & Khan, S. (2020). Social media use and its connection to mental health: A systematic review. *Cureus, 12*(6), e8627. https://doi.org/10.7759/cureus.8627

41. Khamis, S., Ang, L., & Welling, R. (2017). Self-branding, 'micro-celebrity' and the rise of Social Media Influencers. *Celebrity Studies, 8*(2), 191–208. https://doi.org/10.1080/19392397.2016.1218292

42. Lange, J., & Crusius, J. (2015). The tango of two deadly sins: The social-functional relation of envy and pride. *Current Opinion in Psychology, 5*, 20 24. https://doi.org/10.1016/j.copsyc.2015.03.010

43. Lee, S. Y., & Chiou, W.-B. (2022). Algorithmic social comparison and Instagram users' dissatisfaction: The mediating role of envy and the moderating role of self-esteem. *Telematics and Informatics, 66*, 101745. https://doi.org/10.1016/j.tele.2021.101745

44. LeDoux, J. E. (2003). The emotional brain, fear, and the amygdala. *Cellular and Molecular Neurobiology, 23*(4–5), 727–738. https://doi.org/10.1023/A:1025048802629

45. Lund, C., Brooke-Sumner, C., Baingana, F., Baron, E. C., Breuer, E., Chandra, P., Haushofer, J., Herrman, H., Jordans, M., Kieling, C., Medina-Mora, M. E., Morgan, E., Omigbodun, O., Tol, W., Patel, V., & Saxena, S. (2018). Social determinants of mental disorders and the Sustainable Development Goals: A systematic review of reviews. *The Lancet Psychiatry, 5*(4), 357–369. https://doi.org/10.1016/S2215-0366(18)30060-9

46. Lyngs, U., Lukoff, K., Slovak, P., Binns, R., Slack, A., Inzlicht, M., Van Kleek, M., & Shadbolt, N. (2019, May). "Self-control in cyberspace": Applying dual systems theory to a review of digital self-control tools. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–18). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300361

47. Marks, G., & Miller, N. (1987). Ten years of research on the false-consensus effect: An empirical and theoretical review. *Psychological Bulletin, 102*(1), 72–90. https://doi.org/10.1037/0033-2909.102.1.72

48. Marwick, A. E. (2015). Instafame: Luxury selfies in the attention economy. *Public Culture, 27*(1), 137–160. https://doi.org/10.1215/08992363-2798379

49. McEwen, B. S. (2017). Neurobiological and systemic effects of chronic stress. *Chronic Stress, 1*. https://doi.org/10.1177/2470547017692328

50. McEwen, B. S., & Gianaros, P. J. (2010). Central role of the brain in stress and adaptation: Links to socioeconomic status, health, and disease. *Annals of the New York Academy of Sciences, 1186*, 190–222. https://doi.org/10.1111/j.1749-6632.2009.05331.x

51. Miller, A. H., & Raison, C. L. (2016). The role of inflammation in depression: From evolutionary imperative to modern treatment target. *Nature Reviews Immunology, 16*(1), 22–34. https://doi.org/10.1038/nri.2015.5

52. Organisation for Economic Co-operation and Development. (2022). *Beyond financial literacy: Supporting financial well-being through education*. OECD Publishing. https://doi.org/10.1787/58b0b4d1-en

53. Odgers, C. L., & Jensen, M. R. (2020). Annual Research Review: Adolescent mental health in the digital age: Facts, fears, and future directions. *Journal of Child Psychology and Psychiatry, 61*(3), 336–348. https://doi.org/10.1111/jcpp.13190

54. Pew Research Center. (2023, December 11). *Teens, social media and technology 2023*. https://www.pewresearch.org/internet/2023/12/11/teens-social-media-and-technology-2023/

55. Ridge, D., Kokanović, R., Broom, A., Kirkpatrick, S., Anderson, C., & Tanner, C. (2021). "My dirty little habit": Patient constructions of antidepressant use and the 'crisis' of legitimacy. *Social Science & Medicine, 276*, 113833. https://doi.org/10.1016/j.socscimed.2021.113833

56. Rosenberg, J., & Egbert, N. (2011). Online impression management: Personality traits and concerns for secondary goals as predictors of self-presentation tactics on Facebook. *Journal of Computer-Mediated Communication, 17*(1), 1–18. https://doi.org/10.1111/j.1083-6101.2011.01560.x

57. Sapolsky, R. M. (2004). *Why zebras don't get ulcers: The acclaimed guide to stress, stress-related diseases, and coping* (3rd ed.). Holt Paperbacks.

58. Seligman, M. E. P. (1972). Learned helplessness. *Annual Review of Medicine, 23*(1), 407–412. https://doi.org/10.1146/annurev.me.23.020172.002203

59. Siibak, A. (2009). Constructing the self through the photo selection: Visual impression management on social networking websites. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace, 3*(1), Article 1. https://cyberpsychology.eu/article/view/4218

60. Starrin, B., Åslund, C., & Nilsson, K. W. (2009). Financial stress, shaming experiences and psychosocial ill-health: Studies into the finances-shame model. *Social Indicators Research, 91*(2), 283–298. https://doi.org/10.1007/s11205-008-9283-y

61. Taylor, S. E., & Lobel, M. (1989). Social comparison activity under threat: Downward evaluation and upward contacts. *Psychological Review, 96*(4), 569–575. https://doi.org/10.1037/0033-295X.96.4.569

62. Twenge, J. M. (2017). *iGen: Why today's super-connected kids are growing up less rebellious, more tolerant, less happy—and completely unprepared for adulthood*. Atria Books.

63. UK Parliament. (2023). *Online Safety Act 2023* (c. 50). https://www.legislation.gov.uk/ukpga/2023/50/enacted

64. United Nations Educational, Scientific and Cultural Organization. (2023). *Global standards for media and information literacy curricula development guidelines*. UNESCO.

65. U.S. Preventive Services Task Force (USPSTF). (2016). Screening for depression in adults: US Preventive Services Task Force recommendation statement. *JAMA, 315*(4), 380–387. https://doi.org/10.1001/jama.2015.18392

66. Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society, 6*(1). https://doi.org/10.1177/2056305120903408

67. Verduyn, P., Ybarra, O., Résibois, M., Jonides, J., & Kross, E. (2017). Do social network sites enhance or undermine subjective well-being? A critical review. *Social Issues and Policy Review, 11*(1), 274–302. https://doi.org/10.1111/sipr.12033

68. Vogel, E. A., & Rose, J. P. (2016). Self-reflection and interpersonal connection: Making the most of self-presentation on social media. *Translational Issues in Psychological Science, 2*(3), 294–302. https://doi.org/10.1037/tps0000076

69. Vogel, E. A., Rose, J. P., Roberts, L. R., & Eckles, K. (2014). Social comparison, social media, and self-esteem. *Psychology of Popular Media Culture, 3*(4), 206–222. https://doi.org/10.1037/ppm0000047

70. World Health Organization. (2017). *Depression and other common mental disorders: Global health estimates* (WHO/MSD/MER/2017.2). https://iris.who.int/handle/10665/254610

71. Wright, E. J., White, K. M., & Obst, P. L. (2018). Facebook false self-presentation behaviors and negative mental health. *Cyberpsychology, Behavior, and Social Networking, 21*(1), 40–50. https://doi.org/10.1089/cyber.2016.0647

72. Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.

Aspettative e spiegazioni: come interpretiamo il comportamento dei robot sociali - di *Agnese Augello* - Istituto di Calcolo e Reti ad Alte Prestazioni (ICAR) del Consiglio Nazionale delle Ricerche; *Riccardo Bruni* - Dipartimento di Lettere e Filosofia - DILEF, Università degli Studi di Firenze; *Edoardo Datteri* - Università degli Studi di Milano-Bicocca; *Margherita Di Stasio* - INDIRE – Istituto Nazionale di Documentazione, Innovazione e Ricerca Educativa

I progressi dell'intelligenza artificiale e della robotica sociale hanno contribuito al perseguimento di obiettivi da sempre centrali nella progettazione di agenti intelligenti, come l'autonomia, l'adattabilità e la capacità di interazione.

Oggi i robot, con un design sempre più antropomorfo, rispondono e partecipano in modo appropriato a conversazioni in linguaggio naturale, modulano il tono di voce, classificano con accuratezza volti ed emozioni e non sono più confinati a contesti sperimentali, ma vengono impiegati in contesti sociali, dalla scuola all'assistenza, interagendo direttamente con le persone e influenzando le dinamiche relazionali.

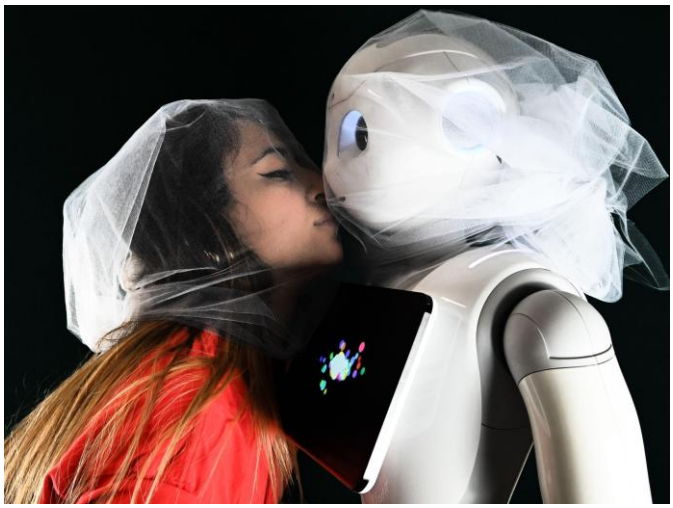

Figura 1 Trasparenza e comprensione: requisiti etici dell'intelligenza artificiale.

Quando un agente intelligente implementato su un corpo robotico interagisce direttamente con le persone occupando lo spazio, modulando gesti, sguardi e distanza, il suo comportamento non va più valutato solo in termini di correttezza tecnica o funzionale, ma deve anche essere comprensibile.

Nei contesti educativi, assistivi o di servizio, la possibilità di interpretare e motivare le azioni di un robot costituisce un elemento essenziale della sua responsabilità sociale, legata alle aspettative e alla fiducia riposte in esso.

La ricerca di spiegazioni può essere guidata da una semplice curiosità intellettuale o essere necessaria per attribuire responsabilità per le conseguenze del comportamento del robot. Se, ad esempio, un robot chirurgico taglia il tessuto errato durante un intervento, sarà importante stabilire se il danno sia dovuto a un movimento scorretto del chirurgo o a un malfunzionamento del robot.

Figura 2 Comprendere la "mente" di un robot

Negli ultimi anni, la ricerca sulla eXplainable Artificial Intelligence (XAI) ha cercato di rispondere a una domanda cruciale: come possiamo capire perché un sistema automatico prende una certa decisione? In molti casi, la risposta passa dall'analisi del funzionamento interno del sistema: quali dati ha considerato, quali elementi hanno inciso di più, quali passaggi hanno portato all'esito finale. È un lavoro fondamentale, perché rende i sistemi meno opachi e più controllabili. Ma quando questi sistemi agiscono nel mondo sociale, emerge l'esigenza di avere altri tipi di spiegazione. Nella vita sociale, infatti, cerchiamo di comprendere un comportamento analizzando le cause, le intenzioni, le regole condivise e le aspettative sociali. Di fronte a un robot che si ferma, cede il passo, ignora una richiesta o interviene in una conversazione, la domanda non è "quale variabile ha pesato di più?", ma: "Perché si è comportato così?

HERB - Human Explanation of Robotic Behaviour

Prima di chiederci come un agente intelligente o un robot debba spiegare le proprie decisioni, occorre fare un passo indietro e guardare la questione dal punto di vista umano.

Quando abbiamo a che fare con un robot, infatti, non ci limitiamo a osservarne il funzionamento: ne interpretiamo le azioni, formiamo aspettative e cerchiamo di attribuire un significato al suo comportamento sulla base dei nostri schemi sociali e cognitivi.

Costruiamo spiegazioni così come facciamo quando osserviamo un fenomeno o il comportamento di un altro essere umano: cerchiamo di collegare ciò che vediamo a un insieme di ragioni, cause o scopi che rendano quell'azione comprensibile. Capire come nascono queste spiegazioni aiuta non solo a comprendere meglio l'interazione uomo-robot, ma anche a progettare sistemi più leggibili, affidabili e responsabili.

Ma una spiegazione non è un semplice commento: è un concetto articolato che può assumere forme diverse, derivare da criteri differenti e riflettere il modo in cui interpretiamo il mondo e le azioni altrui.

Logiche diverse possono condurre a spiegazioni di tipo diverso, che mettono l'accento su meccanismi, regole di funzionamento, funzioni o intenzioni attribuite all'azione osservata.

Nel contesto della robotica sociale, le spiegazioni del comportamento di un robot possono essere ricondotte ad almeno quattro forme principali. Possiamo chiamarle **meccanicistiche** quando facciamo riferimento al suo funzionamento materiale, per esempio sensori, moduli o procedure operative: "Il robot si è fermato perché i suoi sensori hanno rilevato un oggetto". Possiamo chiamarle **nomologiche** quando richiamiamo una regolarità generale del sistema: "I robot con questo sensore reagiscono così quando trovano un ostacolo". Sono invece **teleologiche** quando interpretiamo l'azione a partire dallo scopo che sembra avere: "Il robot si è fermato per evitare di rompersi". Infine, sono **folk-psicologiche** quando gli attribuiamo stati mentali, intenzioni o capacità di comprensione che usiamo normalmente per spiegare il comportamento umano: "Il robot non ha capito cosa gli è stato detto". Distinguere tra queste forme è importante perché ci mostra non solo come giudichiamo il robot, ma anche quali aspettative proiettiamo su di lui.

Lo studio delle spiegazioni del comportamento robotico ha trovato una particolare applicazione nel progetto di ricerca **HERB - Human Explanation of Robotic Behaviour**, che ha dedicato speciale attenzione alle implicazioni di questi processi interpretativi nei contesti sociali ed educativi[2].

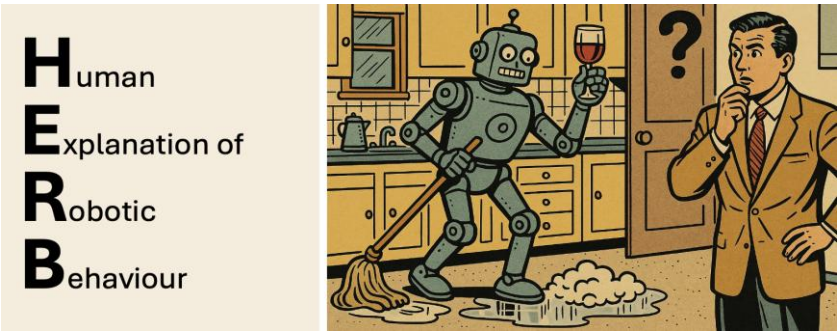

Figura 3 Studio delle spiegazioni nell'interazione uomo-robot

Aspettative sociali

Un elemento chiave dello studio delle **HERB** è il ruolo delle aspettative sociali. L'assunzione è che quando il comportamento di un robot è allineato a ciò che ci aspettiamo in una determinata situazione, raramente emerge un bisogno esplicito di spiegazioni. Il sistema, in quel caso, viene considerato prevedibile e affidabile. Quando invece il robot viola le aspettative, anche in modo sottile, emerge il bisogno di una spiegazione.

È pur vero che le aspettative nei confronti di una macchina possono differire da quelle che abbiamo nelle tipiche situazioni sociali in cui ci troviamo. Tuttavia, un robot umanoide, anche in assenza di una reale competenza sociale, tende a essere interpretato come un agente intenzionale. Questo porta a un'attribuzione spesso falsata di competenze, a una fiducia eccessiva, ad aspettative irrealistiche.

[2] "HERB - Human Explanation of Robotic Behaviour", è un progetto finanziato dal Ministero dell'Università e della Ricerca (programma PRIN 2022)che ha coinvolto il Dipartimento di Scienze Umane per la Formazione "R. Massa" dell'Università degli Studi di Milano-Bicocca, il Dipartimento di Lettere e Filosofia dell'Università di Firenze, l'Istituto di Calcolo e Reti ad Alte Prestazioni (ICAR) del Consiglio Nazionale delle Ricerche e l'Istituto Nazionale di Documentazione, Innovazione e Ricerca Educativa (INDIRE).

Le spiegazioni giocano quindi un ruolo etico cruciale: non devono rafforzare illusioni, ma possono aiutare a ricalibrare le aspettative, rendendo chiari i limiti del sistema.

Figura 4 Aspettative nell'interazione uomo-robot - Foto di Agnese Augello

Studiare le spiegazioni per progettare sistemi più responsabili

Raccogliere e analizzare le spiegazioni che le persone producono nei confronti dei robot consente di individuare dove e perché emergono aspettative errate e di capire quali forme esplicative adottare, tenendo conto del contesto applicativo.

Come osservato da Tom Ziemke, esperto di interazione umani-robot presso l'Università di Linköping, sappiamo ancora relativamente poco su come le persone interpretano il comportamento dei sistemi autonomi interattivi, come i robot sociali o i veicoli automatizzati.

Gran parte della letteratura si è concentrata sull'attribuzione, a grana grossa, di stati mentali ai robot: intenzionalità, emozioni, capacità cognitive. Questi studi analizzano se e quando le persone tendono ad antropomorfizzare i robot, utilizzando questionari, misure comportamentali o neuroscientifiche. Altri lavori si focalizzano su abilità cognitive specifiche, come la capacità di attribuire false credenze o di assumere la prospettiva visiva altrui.

Nonostante l'ampiezza di questo filone di ricerca, resta aperta una questione fondamentale: come le persone prevedono e spiegano il comportamento dei robot nel dettaglio. Studi recenti mostrano che non è ancora chiaro come le aspettative degli utenti si formino e si modifichino in base al contesto, all'esperienza o al tipo di interazione. Questo tipo di analisi richiede strumenti concettuali solidi per distinguere tra diversi tipi di spiegazione, strutture argomentative e indicatori linguistici utilizzati per

dare senso all'azione del robot. Richiede inoltre strumenti computazionali in grado di analizzare grandi quantità di spiegazioni in forma testuale, individuando variazioni e pattern ricorrenti. Si osserva infatti che le spiegazioni variano profondamente, anche quando a formularle è la stessa persona in momenti diversi. Le differenze possono riguardare il linguaggio adottato, il livello di analisi scelto, il riferimento a fattori contestuali o il tipo di connessione logica stabilita tra il comportamento osservato e gli elementi chiamati a giustificarlo.

Rilevanza sociale

Studiare e comprendere le spiegazioni che le persone elaborano di fronte al comportamento dei robot è fondamentale per orientare la progettazione di sistemi più responsabili, comprensibili e, di conseguenza, socialmente accettabili. Le forme esplicative adottate e accettate influenzano direttamente la fiducia, la prevedibilità percepita e la qualità dell'interazione. Questo aspetto assume una rilevanza etica particolarmente elevata nei contesti sensibili, come quelli assistivi o della mobilità autonoma, dove le aspettative sulle capacità sociali dei sistemi rivestono un ruolo decisivo. Comprendere quali tipi di spiegazione emergono nei diversi contesti sociali diventa quindi essenziale per progettare robot che possano essere interpretati correttamente e utilizzati in modo appropriato.

Il progetto HERB ha affrontato queste problematiche attraverso un lavoro sistematico di raccolta e analisi delle spiegazioni prodotte dalle persone in situazioni di interazione reale con robot sociali. Ispirandosi alle *Strange Stories* sviluppate da Francesca G. E. Happé [Happé, 1994] nell'ambito degli studi sulla *Theory of Mind*, la ricerca ha adattato questo approccio al contesto dell'interazione uomo–robot. In particolare, sono stati progettati scenari sociali ispirati a pratiche sociali quotidiane (come la navigazione in spazi condivisi, l'accoglienza e l'assistenza), messi in scena tramite un robot umanoide e utilizzati per elicitare spiegazioni spontanee da parte degli osservatori [Augello, 2025]. Questo ha permesso di raccogliere un corpus ampio di spiegazioni testuali, analizzate sia qualitativamente sia attraverso strumenti sviluppati all'interno del progetto, in collaborazione con il Professore Antonio Lieto dell'Università di Salerno. Tra questi, un ruolo centrale è svolto da un'ontologia di spiegazioni, costruita a partire da una tassonomia di forme esplicative derivate dalla filosofia della scienza, e da un sistema di classificazione ibrido che integra tale conoscenza simbolica con modelli linguistici di grandi dimensioni [Rausa et al., 2025], [Augello et al., 2025].

Un aspetto particolarmente rilevante riguarda la conferma empirica del ruolo delle aspettative sociali. I risultati mostrano che, quando il comportamento del robot è percepito come allineato alle aspettative della situazione, le persone tendono a produrre spiegazioni più stabili, facendo prevalentemente

ricorso a spiegazioni di tipo teleologico, ossia orientate allo scopo attribuito all'azione del robot. Al contrario, quando il comportamento risulta inatteso o ambiguo, emerge più chiaramente il bisogno di una spiegazione: in questi casi aumentano le spiegazioni di tipo meccanicistico e, come osservato anche attraverso momenti di discussione e confronto con i partecipanti, si riduce l'accettazione di attribuzioni antropomorfiche, accompagnata da una valutazione più negativa della competenza sociale del sistema [Augello, 2025].

Un ambito sociale in cui queste questioni acquistano una particolare importanza è quello educativo, dove le aspettative verso i sistemi intelligenti si formano molto presto e contribuiscono a modellare il rapporto con la tecnologia.In questo contesto, le HERB rappresentano uno strumento prezioso per comprendere come studenti e bambini concettualizzano il funzionamento dei sistemi tecnologici. Nonostante l'uso crescente dei robot come strumenti didattici, sappiamo ancora poco su come i bambini spiegano il loro comportamento. Analizzare queste spiegazioni significa osservare in tempo reale la costruzione di modelli mentali, offrendo nuove possibilità per sostenere l'alfabetizzazione tecnologica e il pensiero critico.

Questa rilevanza è particolarmente evidente in una società in cui le tecnologie sono sempre più pervasive, ma spesso restano opache nei loro meccanismi di funzionamento.

In ambito scolastico, le HERB possono aiutare a rendere più comprensibili i meccanismi che guidano il comportamento dei sistemi intelligenti. Per questo possono essere utili nelle attività di coding, robotica educativa e pensiero computazionale supportando una progettazione didattica orientata a rendere espliciti i criteri, le regole e le logiche interpretative che guidano il comportamento degli agenti artificiali. Analizzare come bambini e studenti spiegano le azioni di un robot permette di osservare la costruzione di modelli mentali e aspettative nei confronti dei sistemi intelligenti, evidenziando attribuzioni intenzionali, semplificazioni antropomorfiche e forme emergenti di comprensione algoritmica.

A partire da studi precedenti sulla *Theory of Mind*, attraverso le HERB è possibile mettere in luce come la spiegazione del comportamento robotico non significhi soltanto comprenderne l'azione, ma anche riflettere sulle categorie cognitive utilizzate per interpretarla. In tale prospettiva, l'interazione bambino–robot diventa un dispositivo educativo per sostenere lo sviluppo del pensiero critico e della consapevolezza tecnologica, poiché sollecita il confronto sulla distinzione tra comportamento programmato, adattamento automatico e intenzionalità propriamente umana.

Le HERB consentono di sostenere la costruzione di una alfabetizzazione all'intelligenza artificiale *(AI literacy*) che non si limita all'uso corretto degli strumenti, ma include la capacità di spiegare, prevedere e valutare il comportamento dei sistemi intelligenti in relazione al contesto. In questo senso, lo studio delle HERB assume un valore educativo fondamentale perché aiuta a calibrare le aspettative degli studenti, a prevenire forme di fiducia indebita o di delega cognitiva e a promuovere un rapporto critico, responsabile e informato con le tecnologie emergenti, contribuendo alla formazione di cittadini capaci di comprendere e interagire attivamente con la complessità della società.

Bibliografia

- Artemov, S., and Fitting, M. (2019). *Justification Logic: Reasoning with Reasons*. Cambridge University Press.
- Augello, A., Datteri, E., Lieto, A., Rausa, M., and Zagni, N. (2025). RAGGAE for HERBS: Testing the Explanatory Performance of Ontology-powered LLMs for Human Explanation of Robotic Behaviors. *Proceedings of the International Conference on Social Robotics*. Singapore: Springer Nature Singapore, pp. 558–569.
- Augello, A. Rausa, M. Datteri, E., Zagni N., and Lieto, A. (2025). How are people's explanations of the behaviour of robots structured? An exploratory study and discussion," In Proceedings of the IEEE Conference on Pervasive and Intelligent Computing (PICom), Hakodate, Japan, 2025, pp. 66-73.
- Bromberger, S. (1992). *On What We Know We Don't Know: Explanation, Theory, Linguistics, and How Questions Shape Them*. University of Chicago Press.
- De Graaf, M. A., and Malle, B. F. (2019). People's explanations of robot behavior subtly reveal mental state inferences. *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*.
- Dennett, D. C. (1971). Intentional systems. *The Journal of Philosophy*, 68(4), pp. 87–106.
- Di Stasio, M., Miotti, B. (2024). Intelligent Agents at School. Child–Robot Interactions as an Educational Path. *Education Sciences*, 14(7), 774.
- Happé, F. G. E. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders*, 24(2), pp. 129–154.

- Hempel, C.G. (1965). Aspects of scientific explanation: And other essays in the philosophy of science. New York: The Free Press.
- Hempel, C. G., and Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, 15(2), pp. 135–175.
- Larghi, S., and Datteri, E. (2024). Mentalistic stances towards AI systems: Beyond the intentional stance. *Proceedings of the International Conference on Software Engineering and Formal Methods*, pp. 28–41. Springer Nature Switzerland.
- Luo, J., Studer, T., and Dastani, M. (2025). Tailoring explanations through conversation. *Journal of Logic and Computation*, 35(4).
- Marchesi, S., Ghiglino, D., Ciardo, F., Perez-Osorio, J., Baykara, E., and Wykowska, A. (2019). Do we adopt the intentional stance toward humanoid robots? *Frontiers in Psychology*, 10, 450.
- Rausa, M., Augello, A., and Lieto, A. (2025). Towards an ontology of human explanations of robotic behavior. *Proceedings of SOCIALIZE 2025, Workshop at Intelligent User Interfaces (IUI)*.
- Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., and Müller, K. R. (2019). *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (Vol. 11700). Springer Nature.
- Thellman, S., Silvervarg, A., and Ziemke, T. (2017). Folk-psychological interpretation of human vs. humanoid robot behavior: Exploring the intentional stance toward robots. *Frontiers in Psychology*, 8, 1962.
- Ziemke, T. (2020). Understanding robots. *Science Robotics*, 5(46), eabe2987.

Podcast per la formazione aziendale: istruzioni per l'uso – *di* ***Andrea Laudadio***, Head of TIM Academy & Development

Dai primi esperimenti di audioblogging alla diffusione mainstream, i podcast sono diventati un medium chiave per la formazione aziendale. Numeri, casi d'uso e studi sperimentali mostrano come, con l'arrivo degli AIGP, possano scalare apprendimento continuo e upskilling organizzativo

Introduzione

Il termine podcast nasce nel 2004, dal *word blending* di "iPod" (il lettore musicale portatile di Apple, lanciato nel 2001) e "broadcast" (trasmissione radiofonica/televisiva tradizionale) (Berry, 2006; Bottomley, 2015). In quell'anno, il giornalista Ben Hammersley, sulle pagine del *The Guardian*, si interroga su come chiamare questa nuova forma emergente di distribuzione audio, proponendo le alternative "Audioblogging, Podcasting o GuerillaMedia?" (Hammersley, 2004). Non si trattava soltanto di un neologismo tecnologico: l'incontro tra lettori MP3, software di produzione accessibili e weblog segnava l'avvio di una rivoluzione nella comunicazione audio, spostandola da un modello top-down a uno distribuito e partecipativo (Berry, 2006).

L'introduzione dei feed RSS con *enclosure* rese possibile automatizzare la distribuzione dei file audio, ampliando la capacità di produzione e fruizione libera dai vincoli temporali e istituzionali della radio tradizionale (Sullivan, 2019). La paternità tecnica del podcasting (inizialmente definita come "audioblogging") è attribuita a Adam Curry, ex DJ di MTV, e Dave Winer, sviluppatore software. Dave Winer sviluppò il primo RSS feed con enclosure nel gennaio 2001, sperimentando la distribuzione di file audio. Successivamente, Adam Curry creò nel 2004 iPodder, il primo software per automatizzare il download di trasmissioni audio su iPod (Sullivan, 2019; Lehmann, 2005).

Nel giugno 2005, un evento cruciale segnò la transizione del podcasting dalla nicchia alla potenziale mainstream: Apple integrò il supporto nativo per i podcast in iTunes (Apple, 2005a). Questa mossa istituzionalizzò il termine "podcast" e fornì un'infrastruttura di distribuzione accessibile al grande pubblico, registrando oltre un milione di iscrizioni in soli due giorni (Apple, 2005b).

Nel primo decennio i podcast vedono una crescita graduale e costante, ma restano un prodotto di nicchia, fino al 3 ottobre 2014, quando venne rilasciato il primo episodio di *Serial*, un podcast investigativo prodotto da Sarah Koenig che riesaminava un caso di omicidio del 1999 (Berry, 2016;

Llinares et al., 2018). *Serial* rapidamente infranse tutti i record di download precedenti, raggiungendo milioni di ascoltatori settimanali e diventando “il podcast più popolare al mondo” (Spinelli & Dann, 2019). Contemporaneamente, con il rilascio di iOS 8 nel settembre 2014, Apple rese l’app “Podcasts” (lanciata inizialmente nel giugno 2012) preinstallata e integrata nativamente nel sistema operativo, facilitando ulteriormente l’accesso al medium (Apple, 2014).

Nel secondo decennio la crescita e la diffusione dei podcast è stata esponenziale, evolvendosi da fenomeno di nicchia a medium mainstream. A livello globale, il numero di ascoltatori mensili ha superato i 584 milioni nel 2025, mentre i ricavi pubblicitari nel mercato statunitense hanno raggiunto i 2,4 miliardi di dollari nel 2024, con un incremento del 26,4% rispetto all'anno precedente (Interactive Advertising Bureau & PwC, 2024; Peterson, 2024; Rain News, 2024).

In Italia, la penetrazione mensile è cresciuta dal 31% della popolazione 16-60 anni nel 2021 (9,3 milioni di ascoltatori) al 36% nel 2022 (11,1 milioni), raggiungendo il 39% nel 2023 (11,9 milioni) e stabilizzandosi a tale livello nel 2024 con circa 12 milioni di ascoltatori (Ipsos, 2024a; Noetica, 2024; Questioni d'Orecchio, 2024). La metodologia Ipsos si basa su 2.300 interviste online condotte a luglio 2024 su un universo di riferimento di 30,9 milioni di utenti internet italiani nella fascia 16-60 anni (Ipsos, 2024b). La stabilizzazione al 39% nel 2024 riflette una fase di consolidamento del mercato, caratterizzata da crescita qualitativa piuttosto che quantitativa. Il profilo degli ascoltatori italiani evidenzia una concentrazione significativa di under 35 (39%, indice 115), laureati (29%, indice 136) e professionisti di livello elevato (13%, indice 140), configurandosi come target *hard-to-reach* per i media tradizionali (Ipsos, 2024b). L'intensità di fruizione mostra elevati livelli di engagement: il tempo medio di ascolto giornaliero è di 40 minuti nel 2024, in crescita rispetto ai 37 minuti del 2023, con un *completion rate* del 62% (Ipsos, 2024b).

Il podcast come strumento formativo, una idea vecchia come l’uomo

Il primo podcast educativo mainstream sembrerebbe essere “In Our Time” della BBC. Il programma, condotto da Melvyn Bragg divenne disponibile come podcast nel novembre del 2004, registrando, subito oltre 70.000 download. Sempre nello stesso anno, la Duke University lancia un programma formativo universitario basato sui podcast, dotando 1.650 studenti del primo anno, di iPod e rendendo disponibili contenuti didattici forniti dai docenti (Corbeil & Valdes-Corbeil, 2007). Il progetto fu un successo, ma uno degli usi più creativi emersi dall’esperimento fu quello di un professore che richiese

agli studenti di completare un compito creando un podcast audio invece di un elaborato scritto. L'analisi dei risultati rivelò che gli studenti avevano effettivamente scritto di più (in termini di bozze, script, ecc.) rispetto a quanto avrebbero fatto per un rapporto tradizionale. Poiché gli studenti condividevano i loro podcast non solo con il docente ma li fornivano ad altri studenti per l'ascolto, si riscontrò un aumento complessivo della qualità, grazie all'intera dinamica tra pari (Duke Today Staff, 2004).

Nel 2005 Capital One Financial Corp. ha implementato uno dei primi programmi strutturati di audio learning aziendale negli Stati Uniti, anticipando di oltre un decennio la diffusione massiva dei podcast formativi. Il progetto pilota partì con 30 utenti per scalare rapidamente a oltre 4.000 dipendenti attivi attraverso una partnership strategica con Apple e Audible. L'azienda forniva gratuitamente un iPod brandizzato a tutti i dipendenti iscritti a corsi con contenuti audio. Il dispositivo poteva essere utilizzato sia per materiali professionali sia per ascolti personali, eliminando la separazione tra tecnologia aziendale e uso individuale. L'idea centrale consisteva nel convertire i tempi improduttivi in opportunità di apprendimento continuo. Gli utenti ascoltavano contenuti durante gli spostamenti in auto, in aereo, in palestra o durante attività quotidiane. Il programma venne inizialmente integrato nei corsi della Capital One University (leadership, comunicazione, management). L'offerta si ampliò includendo *executive book summaries*, contenuti Harvard Business Review, interventi di top manager e materiali personalizzati per ruoli aziendali specifici. La libreria audio interna integrava contenuti Audible con moduli pre e post corso, conferenze e aggiornamenti aziendali (Boehle, 2007).

La novità tecnologica non deve oscurare la continuità storica. La trasmissione orale della conoscenza rappresenta il più antico metodo formativo dell'umanità, precedente alla scrittura stessa. La radio educativa costituisce il primo tentativo sistematico di applicare le tecnologie audio all'apprendimento di massa. La formazione attraverso contenuti audio rappresenta una pratica pedagogica consolidata ben prima dell'avvento digitale (Bianchi, 2008). Emittenti come la Wisconsin School of the Air, attivata nel 1931 negli Stati Uniti, diffondevano programmi educativi complementari ai curricula scolastici tradizionali, raggiungendo oltre 70.000 studenti entro il 1938 attraverso lezioni in musica, arte, scienze naturali e letteratura, applicando i principi dell'apprendimento esperienziale deweyano alle aule rurali americane (Bianchi, 2008; Bianchi, 2002). L'Interactive Audio Instruction (IAI), sviluppata inizialmente negli anni Settanta per il Radio Language Arts Project in Kenya e successivamente implementata in Nicaragua e Repubblica Dominicana, dimostrava l'efficacia dell'audio strutturato con lezioni temporizzate basate su risposte corali, domande individuali e pause calibrate per l'interazione, documentando incrementi significativi nei risultati di apprendimento con effetti misurabili di 0.5 deviazioni standard rispetto ai gruppi di controllo (Moulton, 1994; Bosch & Tilson, 1991). La

democratizzazione tecnologica accelerò con l'introduzione delle audiocassette compatte da parte di Philips nell'agosto 1963 al Berlin Radio Show, eliminando la complessità dei nastri a bobina aperta e rendendo accessibile la registrazione e riproduzione audio per uso personale ed educativo, applicazione estesa alla formazione linguistica e ai programmi di sviluppo professionale aziendale distribuiti attraverso network corporativi (Ottens, 1963). Il podcast contemporaneo si inserisce in questa continuità evolutiva combinando accessibilità audio, distribuzione digitale e fruizione on-demand (Koehler et al., 2024).

I podcast formativi in ambito aziendale

I podcast stanno ridefinendo il panorama della formazione aziendale, emergendo come strumento capace di coniugare flessibilità operativa e rigore pedagogico (Morris et al., 2023). La trasformazione digitale ha reso necessari modelli formativi che superino i vincoli spazio-temporali tradizionali, valorizzando l'intimità della voce come leva di coinvolgimento e la capacità di catturare conoscenze tacite attraverso conversazioni strutturate (Morris et al., 2023; Meden et al., 2024).

Un aspetto chiave emerso dagli studi è l'elevato grado di coinvolgimento generato dai podcast grazie alla combinazione tra educazione e intrattenimento (*edutainment*). La narrazione informale, la conversazione tra esperti e l'uso di aneddoti personali contribuiscono ad aumentare l'attenzione, la motivazione e la percezione di credibilità degli ascoltatori (Berk et al., 2020).

La letteratura si concentra prevalentemente sul settore sanitario. I podcast vengono giudicati tra le risorse più efficaci per la formazione continua degli operatori, facilitando l'accesso all'aggiornamento professionale e stimolando lo sviluppo delle competenze. Tra i principali vantaggi si annoverano la flessibilità dell'accesso e la possibilità di promuovere il cambiamento comportamentale auto-riferito (Kelly et al., 2022; Calheiros et al., 2024). Permangono limiti metodologici che impediscono una valutazione definitiva dell'efficacia dei podcast in contesti aziendali generici. La letteratura evidenzia la mancanza di studi interventistici strutturati e la scarsità di misurazioni di *outcome* organizzativi e sistemici, soprattutto nei settori non sanitari (Kelly et al., 2022).

Le evidenze sperimentali documentano incrementi significativi nei risultati di apprendimento: Koehler et al. (2024) riportano miglioramenti fino al 7.9% negli esiti quando podcast settimanali accompagnano corsi MOOC sulla piattaforma openHPI, mentre Errabo et al. (2024) osservano incrementi del 6.3-6.4% in esercizi e esami utilizzando podcast differenziati per promuovere autoregolazione. L'accettazione tra

i professionisti risulta elevata, con atteggiamenti molto favorevoli verso questa modalità di sviluppo asincrono (Ghareeb, 2017; Carmi, 2023; Cunado et al., 2022).

Le organizzazioni implementano diversi modelli: il *broadcast* interno per diffondere conoscenze strategiche attraverso conversazioni tra esperti (Morris et al., 2023), l'integrazione con ecosistemi digitali dove i podcast accompagnano corsi online (Koehler et al., 2024; Errabo et al., 2024), e approcci mobile-centric con Personal Learning Environment dedicati. I benefici organizzativi includono distribuzione scalabile "*design once, deliver many times*" (Morris et al., 2023; Nie et al., 2010), costi contenuti con potenziali ritorni rapidi documentati in casi aziendali (Morris et al., 2023), e conservazione di expertise attraverso conversazioni strutturate (Meden et al., 2024). La ricerca qualitativa evidenzia come l'ascolto di podcast legati al lavoro sostenga apprendimento informale, creatività e dimensioni di benessere lavorativo *hedonic*, *social* ed *eudaimonic* (Shamburg et al., 2023).

Le criticità riguardano la necessità di qualità di produzione professionale, allineamento strategico dei contenuti con esigenze lavorative, *governance* editoriale chiara e misurazione rigorosa dell'efficacia attraverso metriche standardizzate (Errabo et al., 2024; Koehler et al., 2024). Le tendenze future includono l'integrazione con intelligenza artificiale per personalizzazione avanzata basata su profilo e progressi del *learner* (Shamburg et al., 2023), micro-podcasting con episodi ultra-brevi per *just-in-time learning*, e modelli ibridi che combinano podcast con webinar live e workshop pratici (Wärnestål & Sjöberg, 2020). Il successo dipende da progetti pilota strutturati, investimento in qualità produttiva, integrazione tecnologica efficace e iterazione basata su feedback e dati di utilizzo.

La nuova frontiera, gli AIGP

I podcast generati con l'intelligenza artificiale (AI-Generated Podcasts, AIGP) rappresentano una tecnologia educativa emergente che sfrutta modelli linguistici di grandi dimensioni (Large Language Models, LLM) per la generazione di contenuti testuali e sistemi di sintesi vocale neurale (neural Text-to-Speech, TTS) per la narrazione (Mehta et al., 2024; Morris, 2024). Questa innovazione affronta direttamente le criticità di scalabilità e costi associate alla produzione tradizionale di podcast, che richiede in media cinque ore di lavoro per un episodio di 3-4 minuti, con costi tra 50 e 995 dollari per episodio (Caldwell et al., 2024). Gli AIGP possono generare contenuti comparabili in tempi compresi tra pochi minuti e alcune ore, riducendo i tempi di produzione del 70-90% (Do et al., 2024; Kokala, 2024).

La produzione di AIGP impiega principalmente due componenti tecnologiche: la generazione di script attraverso LLM (come GPT-4, Claude, Gemini) e la sintesi vocale tramite motori TTS neurali (come ElevenLabs, Amazon Polly, Google Cloud TTS) (Vrzalková, 2023). In una revisione sistematica condotta su 63 studi empirici, NotebookLM di Google è emerso come la piattaforma più utilizzata (35% degli

studi), seguito da ElevenLabs (20%) e combinazioni personalizzate LLM-TTS (25%) (Desmedt et al., 2025; Mehta et al., 2024). NotebookLM consente la generazione automatizzata *end-to-end* di podcast da documenti attraverso la funzione Audio Overview, mentre le implementazioni personalizzate offrono un maggiore controllo sulla personalizzazione del contenuto (Kokala, 2024). La pipeline di produzione prevede tipicamente una configurazione iniziale, la generazione automatica della scaletta, la creazione del trascritto completo con verifica dei fatti rispetto al materiale sorgente, e il post-processing per la sintesi vocale (Do et al., 2024). Sistemi specializzati includono PAIGE, che converte capitoli di libri in podcast personalizzati, PodGPT, che sfrutta oltre 3.700 ore di contenuti scientifici, e MoonCast per la generazione *zero-shot* di podcast di alta qualità (Jia et al., 2025; Ju et al., 2025).

Per quanto riguarda l'efficacia rispetto ai podcast tradizionali, gli AIGP mostrano tassi di completamento più elevati (78% contro 65% dei podcast tradizionali) e una maggiore propensione alla ripetizione dell'ascolto (42% contro 28%). Gli studenti trovano i podcast educativi generati da AI più piacevoli dei libri di testo tradizionali, indicando il valore dell'elemento di intrattenimento nell'apprendimento (Do et al., 2024). Uno studio randomizzato controllato con 180 studenti universitari negli Stati Uniti ha rilevato che i podcast generati con AI personalizzati hanno prodotto miglioramenti significativi nei risultati di apprendimento in filosofia ($M=7.30$ vs. $M=5.85$ per podcast generici, $p<0.01$) e risultati promettenti in psicologia, sebbene non in scienze politiche (Do et al., 2024). Uno studio in ambito medico ha documentato un miglioramento del 17% nei punteggi OSCE per studenti di medicina che utilizzavano podcast generati con AI in nefrologia (27.6 ± 3.6 vs. 23.6 ± 5.0, $p=0.002$) (Dupont et al., 2025). La soddisfazione complessiva media è stata di 4.2/5.0, con valutazioni particolarmente elevate per accessibilità (4.5/5.0) e convenienza (4.4/5.0), mentre la naturalezza della voce ha ricevuto i punteggi più bassi (3.6/5.0), evidenziando limitazioni tecniche persistenti nella sintesi vocale. Gli studi qualitativi hanno rilevato che le voci sintetiche vengono frequentemente descritte come "robotiche" ed "emotivamente distaccate" rispetto ai narratori umani (Benner et al., 2025).

La ricerca sull'utilizzo degli AIGP in contesti lavorativi per la formazione continua dei lavoratori è ancora limitata ma emergente. Il 10% degli studi esaminati (6 su 63) ha analizzato applicazioni nella formazione sul posto di lavoro (*workplace training*). Benner et al. (2025) hanno condotto uno studio diaristico con professionisti universitari, documentando 177 visualizzazioni e 1.337 minuti di ascolto, con sentimenti contrastanti riguardo all'esperienza complessiva. Uno studio specifico sulla formazione medica continua ha documentato che il 94.3% dei professionisti sanitari ha trovato i podcast di ricerca generati con AI accessibili, con il 97.1% che ha valutato positivamente la chiarezza dei contenuti (Peterson et al., 2025). Nel contesto della formazione sanitaria continua in Brasile, uno studio sulla pianificazione sanitaria ha rilevato che l'uso del podcast, insieme all'ascolto ripetuto, ha contribuito a migliorare

significativamente la comprensione dei professionisti nei loro scenari pratici attraverso cicli di apprendimento continuo (Pinto et al., 2023). La flessibilità nella produzione e distribuzione supporta opportunità pedagogiche particolarmente rilevanti per la formazione continua professionale, consentendo l'apprendimento durante attività quotidiane come spostamenti, cucina o esercizio fisico (Benner et al., 2025; Desmedt et al., 2025).

Di grande interesse è il recentissimo studio di Xiao et al. (2025) che partendo dalla crescente difficoltà nel valutare contenuti generati dall'IA (AIGC), propongono PodEval, un framework di valutazione multimodale open-source. A livello pratico, i risultati dell'applicazione di PodEval su diversi sistemi (open-source, closed-source e umani) sono significativi. Nell'analisi del Testo, emerge che i podcast umani (Real-Pod) mostrano una minore diversità lessicale (Distinct-2, MATTR) rispetto all'IA, ma una maggiore densità informativa (Info-Dens) e diversità semantica (Sem-Div). Questo suggerisce che l'interazione umana reale utilizza un vocabolario più semplice, ma copre più informazioni nel tempo. Nell'analisi del Parlato, la naturalezza del dialogo si rivela il fattore dominante sull'intera esperienza. Nel test soggettivo (basato su MUSHRA e filtrato per l'attenzione), il sistema closed-source NotebookLM è risultato secondo solo ai podcast reali (HQ), dimostrandosi quasi indistinguibile. È emerso inoltre che i sistemi TTS specifici per il dialogo (es. MoonCast) superano nettamente i sistemi basati su sintesi di frasi singole (es. PodAgent). Nell'analisi Audio (olistica), i test MOS hanno confermato che la scarsa naturalezza del dialogo di un sistema compromette la percezione di tutte le altre metriche, inclusi l'engagement e la qualità audio percepita. È emersa una chiara distinzione tra percezione oggettiva e preferenza soggettiva: i sistemi AI (NotebookLM) hanno ottenuto punteggi di Audio Quality leggermente superiori ai podcast reali, grazie a una produzione più pulita e controllata (es. assenza di fruscii). Tuttavia, i podcast reali hanno dominato nelle metriche di Engagement Level e Human Likelihood.

Raccomandazioni evidence-based per il design

La costruzione di podcast (sia AIGP sia tradizionale) di qualità per la formazione aziendale richiede una progettazione basata su evidenze empiriche solide. Come costruire un podcast educativo efficace? La ricerca scientifica fornisce indicazioni precise su molti aspetti, ma non su tutti. Gli studi dimostrano che i podcast progettati secondo principi di design istruzionale riducono il carico cognitivo percepito favorendo l'acquisizione di conoscenze (Kennedy et al., 2016), mentre l'integrazione curricolare deliberata può aumentare i risultati di apprendimento fino al 7,9% (Koehler et al., 2024). La personalizzazione dei contenuti rispetto alle caratteristiche degli ascoltatori produce miglioramenti significativi nell'apprendimento rispetto a versioni generiche (Do et al., 2024) e la produzione attiva di podcast da parte dei discenti risulta superiore al consumo passivo (Laupichler et al., 2025). Gli

ascoltatori richiedono esplicitamente che i contenuti siano rilevanti rispetto agli obiettivi formativi (Tam, 2012) e che i podcast siano ben progettati e integrati per migliorare il coinvolgimento (Gast & Shifrin, 2024). Il decalogo che segue sintetizza le principali evidenze operative emerse dalla letteratura scientifica, traducendo i risultati della ricerca in linee guida concrete per la progettazione di podcast formativi efficaci nel contesto aziendale.

Segmentazione e Scaffolding

Un podcast educativo ben progettato non è semplicemente una lezione registrata. Si tratta di uno strumento pedagogico costruito con precisione, dove la struttura fa la differenza nell'apprendimento. Kay (2014) ha testato su centinaia di studenti un approccio basato sulla presentazione graduale dei contenuti. Gli studenti hanno apprezzato esplicitamente questa metodologia passo-passo. Il principio è semplice: prendere un concetto complesso, scomporlo in pezzi più piccoli e spiegare perché ogni passaggio è necessario. Questa tecnica si chiama "*worked-example*" (esempio guidato) e funziona perché riduce il carico cognitivo dell'ascoltatore, permettendogli di seguire un percorso chiaro dalla A alla Z. Alam et al. (2016) hanno scoperto qualcosa di importante: i podcast funzionano meglio quando non ci limitiamo a spiegare, ma aggiungiamo elementi di pratica mentale o *modeling*. La pratica mentale significa far immaginare all'ascoltatore di applicare quello che sta imparando. Il *modeling* consiste nel mostrare concretamente come un esperto affronta un problema. I loro dati dimostrano che "l'efficacia dei podcast potenziati per la ritenzione della conoscenza e l'acquisizione di competenze cliniche è aumentata con la *mental practice* o il *modeling*" (Alam et al., 2016).

Miller e Uphold (2021) hanno formalizzato una struttura specifica per i podcast educativi. Un CAP (*Content Acquisition Podcasts*) ben costruito include: (1) Il razionale (perché questa strategia funziona), (2) le evidenze (quali aspetti del comportamento migliora), (3) il target (per quali studenti è più efficace), (4) i passaggi operativi (come incorporare la strategia nella pratica) e (5) i suggerimenti pratici (consigli per l'acquisizione concreta). Questa architettura crea quello lo "*scaffolding* pedagogico": una struttura di supporto temporanea che aiuta l'ascoltatore a costruire la propria comprensione, proprio come le impalcature sostengono un edificio in costruzione.

Immaginiamo di dover spiegare una procedura medica complessa. Invece di descrivere tutto in una volta, il podcast efficace:

1. Mostra i passaggi uno per uno, spiegando il perché di ciascuno
2. Fa immaginare all'ascoltatore di eseguire la procedura (*mental practice*)
3. Fornisce esempi concreti di applicazione (*modeling*)

4. Offre suggerimenti per evitare errori comuni

Questo approccio strutturato trasforma l'ascolto passivo in un processo di apprendimento attivo.

Carico cognitivo

La progettazione di un podcast formativo deve essere guidata dalla Teoria del Carico Cognitivo (Cognitive Load Theory) per evitare di sovraccaricare la memoria di lavoro dell'ascoltatore. I *Content Acquisition Podcasts* (CAPs), una forma di podcast potenziati con immagini statiche sincronizzate all'audio, sono esplicitamente progettati per incorporare la teoria cognitiva dell'apprendimento multimediale di Mayer. Kennedy et al. (2014) spiegano che i CAPs sono "una forma di podcast potenziati (immagini statiche sincronizzate con l'audio) che incorporano la teoria cognitiva dell'apprendimento multimediale di Mayer (2009) e i principi di design istruzionale accompagnatori per assicurare che l'aspetto e i suoni dell'istruzione non travolgano le limitazioni dei processi cognitivi degli utenti". Questo implica una progettazione attenta che riduca il carico cognitivo estrinseco, cioè quello non essenziale all'apprendimento.

Castro-Alonso et al. (2021) hanno identificato cinque strategie essenziali per ottimizzare il carico cognitivo nei materiali multimediali educativi: (1) il principio multimediale (le persone apprendono meglio da parole e immagini combinate piuttosto che solo da parole), (2) il principio di contiguità spaziale (parole e immagini correlate devono essere presentate vicine nello spazio), (3) il principio di coerenza (eliminare materiale estraneo che non supporta l'obiettivo di apprendimento), (4) il principio di segnalazione (*signaling*) (evidenziare gli elementi essenziali per guidare l'attenzione dell'ascoltatore) e (5) il principio di segmentazione (suddividere il contenuto in segmenti gestibili controllati dall'utente). Questi principi traducono la teoria in scelte concrete di design: eliminare musiche di sottofondo superflue, sincronizzare le immagini esattamente quando vengono menzionate nell'audio, evidenziare visivamente i concetti chiave mentre vengono spiegati. Studi empirici confermano l'efficacia di questo approccio: "gli studenti nella condizione CAP hanno riportato un carico cognitivo percepito significativamente inferiore rispetto agli studenti nella condizione tradizionale" (Kennedy et al., 2016), favorendo l'acquisizione di conoscenze.

Tecnica e Workflow

La qualità tecnica non è un aspetto accessorio, ma un prerequisito fondamentale che influenza direttamente l'utilità percepita del podcast. Murtinugraha et al. (2025) dimostrano che "i risultati hanno mostrato che l'aspetto tecnico, gli aspetti di implementazione e i benefici hanno un'influenza positiva, il che significa che più aumentano questi due aspetti, più aumenta l'utilità percepita dello sviluppo dei media". La qualità audio scadente, il rumore di fondo o le variazioni di volume non sono solo fastidiosi:

compromettono l'apprendimento perché costringono l'ascoltatore a concentrarsi sul decifrare il messaggio invece che sul comprenderne il contenuto.

Per garantire standard qualitativi elevati, la letteratura suggerisce di adottare un *workflow* di produzione strutturato in cinque fasi: (1) pre-produzione – concezione (definire obiettivi di apprendimento e target), (2) pre-produzione – pianificazione editoriale e briefing (sviluppare lo script e pianificare le risorse), (3) produzione – tecnica e processo di registrazione (catturare l'audio in condizioni ottimali), (4) post-produzione – editing e processamento audio (eliminare errori, equalizzare, normalizzare il volume) e (5) pubblicazione (distribuire sui canali appropriati con metadati adeguati) (Murtinugraha et al., 2025). Questo approccio sistematico chiarisce i ruoli, supporta la qualità e guida il miglioramento continuo. Ogni fase ha obiettivi specifici e *deliverable* definiti: uno script validato nella fase 2 previene improvvisazioni dannose nella fase 3, mentre un editing accurato nella fase 4 può salvare una registrazione con piccole imperfezioni.

Anche se i recenti progressi tecnologici hanno semplificato la produzione, la variabilità nella qualità audio rimane una sfida concreta. Karla (2020) osserva questa problematica specialmente nei podcast prodotti da studenti. Che (2023) ha rilevato che "la variazione in accento, ritmo, udibilità e chiarezza dei podcast studenteschi ha posto una sfida importante alla comprensione di alcuni podcast e alla partecipazione nelle discussioni basate su podcast". Questo sottolinea la necessità di un controllo rigoroso in ogni fase del processo: microfoni adeguati, ambienti silenziosi, prove di registrazione e ascolti critici prima della pubblicazione finale.

Durata e accelerazione

La letteratura scientifica non offre un'indicazione univoca sulla durata ottimale dei podcast formativi. Le raccomandazioni variano in funzione del contesto disciplinare, della complessità dei contenuti e delle caratteristiche del pubblico. Una *scoping review* sul microlearning nelle professioni sanitarie ha rilevato che "tutti gli studi presentavano concetti di tempo massimo speso inferiore a 15 minuti così come aggregazione di contenuti" (De Gagne et al., 2019), coerentemente con le preferenze dei medici in formazione che identificano in "episodi brevi con buona qualità produttiva" una caratteristica essenziale (Riddell et al., 2021). Altri studi confermano la tendenza pur variando: 20 minuti per podcast biochimici (Münch-Harrach et al., 2013), 5 minuti per video podcast su esempi svolti (Kay, 2014). Oltre alla durata, è cruciale la segmentazione in "blocchi cognitivi significativi" per gestire il carico mentale dell'ascoltatore (Kay, 2014).

Un'ulteriore strategia è la compressione temporale digitale, che accelera la riproduzione preservando intelligibilità e qualità tonale. Le ricerche storiche hanno dimostrato che la comprensione non si

deteriora significativamente fino a 250 parole al minuto (Foulke, 1968) o circa 225 parole al minuto (Barabasz, 1968). Gli studi recenti in contesti multimediali confermano che velocità fino a 210 parole al minuto (circa 1.4x) possono essere utilizzate efficacemente senza impattare l'apprendimento (Pastore & Ritzhaupt, 2015). Ritzhaupt e Barron (2008) hanno trovato differenze significative solo a 2.5x la velocità normale. Contenuti complessi tollerano compressioni inferiori (25% o 1.5x) rispetto a informazioni dichiarative semplici (50-75%) (Pastore, 2010). La compressione deve essere combinata con elementi visivi congruenti (Ritzhaupt et al., 2011) e gli ascoltatori dovrebbero controllare la velocità, preferendo accelerazioni moderate del 10-20% (Pastore, 2015).

Narrazione e segnalazione prosodica

Per massimizzare il coinvolgimento (*engagement*), un podcast formativo dovrebbe andare oltre la semplice esposizione di fatti e adottare uno stile comunicativo più ricco. I professionisti sanitari in formazione identificano come elementi di efficacia uno "stile narrativo che incorpora umorismo e *storytelling*" e "contenuti guidati da esperti con apprendimento basato su casi" (Riddell et al., 2021). Questo approccio, definito *edutainment* (fusione di educazione e intrattenimento), deve essere attentamente bilanciato per non compromettere il rigore accademico. Una valutazione di podcast generati da AI ha evidenziato che "lo stile americanizzato dei podcast sminuiva il loro tono accademico e introduceva un elemento di sensazionalismo non necessario" (Desmedt et al., 2025), suggerendo la necessità di un compromesso tra coinvolgimento ed autorevolezza scientifica. Lo *storytelling* basato su casi clinici o professionali reali offre un equilibrio ottimale: mantiene l'ancoraggio alla pratica professionale mentre rende i contenuti più memorabili attraverso strutture narrative che facilitano l'organizzazione cognitiva delle informazioni.

Un potente strumento per guidare l'attenzione in un formato puramente audio è la segnalazione prosodica, l'uso strategico di variazioni di tono, ritmo e volume per evidenziare elementi chiave. Désiron e Schneider (2024) hanno dimostrato che "i risultati hanno mostrato un effetto positivo di entrambe le manipolazioni dei segnali prosodici (effetti principali) sui risultati di apprendimento e l'impatto più sostanziale quando cumulati", estendendo il principio di segnalazione di Mayer dalla modalità visiva a quella uditiva. La segnalazione prosodica funziona come equivalente acustico dell'evidenziazione visiva. L'uso deve essere giudizioso: "risultati di apprendimento più elevati erano anche associati a maggiore sforzo mentale e carico con volume più alto e nessuna differenza nello sforzo ma un carico inferiore per un ritmo più lento" (Désiron & Schneider, 2024). Un ritmo più lento risulta preferibile perché migliora l'apprendimento riducendo il carico cognitivo, mentre l'aumento di volume, pur efficace, aumenta lo sforzo percepito dall'ascoltatore.

Accessibilità e Inclusione

Un approccio inclusivo alla progettazione di podcast formativi è essenziale per supportare tutti gli studenti, rispondendo ai principi del Universal Design for Learning (UDL) che promuove molteplici modalità di rappresentazione, espressione e coinvolgimento. La fornitura di trascrizioni rappresenta il primo e più importante passo verso l'accessibilità. Schwertz (2025) osserva che "se l'oralità degli episodi viene trascritta in forma scritta, questo strumento può essere considerato una tecnologia inclusiva per gli studenti con disabilità uditive". Le trascrizioni non beneficiano solo studenti con disabilità: consentono la consultazione rapida, la ricerca di termini specifici, la traduzione automatica in altre lingue e l'utilizzo in contesti dove l'ascolto non è possibile (ambienti rumorosi, biblioteche silenziose). La disponibilità simultanea di audio e testo scritto permette inoltre strategie di apprendimento multimodale, particolarmente utili per studenti non madrelingua o con disturbi specifici dell'apprendimento.

Oltre alle trascrizioni, l'uso di podcast potenziati (*enhanced podcasts*) che sincronizzano l'audio con immagini statiche offre un'ulteriore modalità di accesso al contenuto. Deshmukh et al. (2025) hanno rilevato che gli studenti che hanno utilizzato questi formati "hanno apprezzato i segnali integrati del formato e la consegna stratificata, che hanno incoraggiato un apprendimento focalizzato e un'attenzione sostenuta". Questa architettura multimodale risponde alle diverse preferenze e capacità di apprendimento: alcuni studenti beneficiano maggiormente dell'input visivo, altri di quello uditivo, altri ancora della combinazione sincrona di entrambi. Jafarian e Kramer (2024) hanno dimostrato empiricamente questa variabilità, osservando che "gli studenti con preferenze di apprendimento uditivo hanno mostrato miglioramenti significativi nella performance accademica quando esposti a materiali di apprendimento basati su podcast". Questo conferma il valore di offrire una varietà di formati per soddisfare le esigenze individuali, trasformando la diversità di preferenze e capacità da potenziale ostacolo a risorsa pedagogica da valorizzare attraverso la flessibilità progettuale.

Integrazione culturale

L'efficacia di un podcast formativo dipende in modo cruciale dalla sua integrazione deliberata all'interno di un percorso didattico strutturato. La ricerca mostra che l'uso di podcast come contenuto supplementare, se supportato da un adeguato *scaffolding* pedagogico, può produrre benefici misurabili sull'apprendimento. Koehler et al. (2024) hanno dimostrato in contesti MOOC che "consumare un podcast aggiuntivo può aumentare i risultati di apprendimento fino al 7,9%". Questo beneficio non è automatico: si realizza solo quando esiste un chiaro allineamento tra gli obiettivi del podcast e quelli del corso nel suo complesso. Gli studenti riconoscono questo bisogno di coerenza: Tam (2012) ha

rilevato che "gli studenti hanno suggerito che i docenti dovessero aumentare la rilevanza del contenuto del podcast rispetto a ciò che veniva insegnato durante le lezioni".

La semplice disponibilità del materiale audio non garantisce alcun valore formativo: è necessario che il podcast sia inserito nel curriculum con una chiara intenzione pedagogica, esplicitando come si collega alle altre attività didattiche, quali obiettivi di apprendimento specifici persegue e come si integra nella valutazione. Un'integrazione efficace richiede progettazione esplicita. Gast e Shifrin (2024) hanno confermato che "i podcast hanno soddisfatto gli obiettivi del corso e migliorato il coinvolgimento quando ben progettati e integrati". Che il podcast sia utilizzato come strumento primario di istruzione o come risorsa supplementare, deve essere percepito dagli studenti come una componente coerente e rilevante dell'esperienza di apprendimento complessiva, non come un'aggiunta opzionale o marginale. Lo *scaffolding* può assumere diverse forme: guide all'ascolto con domande orientative, attività preparatorie che attivano conoscenze pregresse, discussioni successive che consolidano quanto appreso, compiti che richiedono l'applicazione dei concetti presentati nel podcast. Questa architettura pedagogica trasforma l'ascolto da esperienza passiva a processo attivo di costruzione della conoscenza.

Personalizzazione e differenziazione dei contenuti

Per rispondere efficacemente alla diversità degli ascoltatori, la letteratura suggerisce di superare un approccio "*one-size-fits-all*" (una soluzione unica per tutti) e di progettare podcast personalizzati e differenziati che rispondano alle caratteristiche specifiche dei destinatari. L'assunto alla base di questo principio è che gli studenti differiscono per conoscenze pregresse, obiettivi professionali, preferenze di apprendimento e contesti di applicazione: un podcast efficace per studenti di ingegneria potrebbe non esserlo per studenti di scienze sociali, e viceversa. Do et al. (2024) hanno dimostrato empiricamente il valore di questo approccio, osservando che "i podcast personalizzati erano adattati alle *major*, agli interessi e agli stili di apprendimento degli studenti" e che "i podcast personalizzati hanno portato a risultati di apprendimento significativamente migliorati" rispetto a versioni generiche dello stesso contenuto. La personalizzazione trasforma il podcast da risorsa standardizzata a esperienza formativa su misura, aumentando sia la rilevanza percepita sia l'efficacia dell'apprendimento.

Questa personalizzazione può avvenire a diversi livelli di granularità. Errabo et al. (2024) hanno esplorato la creazione di "podcast differenziati adattati agli stili di apprendimento degli studenti" per promuovere l'autoregolazione e l'apprendimento asincrono. La differenziazione può riguardare la velocità di presentazione (livelli basic, intermediate, advanced), la profondità di trattamento (panoramica generale versus approfondimenti specialistici), gli esempi utilizzati (tratti da contesti

professionali diversi), o persino la struttura narrativa (più lineare per principianti, più problematizzante per esperti). Il contenuto può essere adattato al contesto professionale specifico degli ascoltatori. Riddell et al. (2021) hanno osservato che "i podcast educativi sono utilizzati dai tirocinanti in medicina d'emergenza per integrare l'apprendimento clinico e promuovere un senso di connessione con comunità mediche più ampie", evidenziando come la specificità disciplinare aumenti sia l'utilità pratica sia il senso di appartenenza professionale. La capacità di adattare il contenuto, la profondità e lo stile del podcast alle esigenze specifiche dei diversi sottogruppi di discenti rappresenta una delle frontiere più promettenti per aumentarne l'impatto formativo, trasformando potenziali limitazioni della standardizzazione in opportunità di personalizzazione scalabile.

Produzione attiva

Una delle evidenze più forti e ricorrenti nella letteratura riguarda la superiorità pedagogica della produzione di podcast da parte degli studenti (*student-generated podcasts*) rispetto al semplice ascolto passivo. Laupichler et al. (2025) hanno concluso in uno studio controllato in ambito medico che "la produzione attiva di podcast ha migliorato significativamente i risultati di apprendimento rispetto al consumo passivo". Questo ribaltamento di ruolo trasforma gli studenti da consumatori a produttori di conoscenza, innescando processi cognitivi più profondi: per creare un podcast efficace, gli studenti devono analizzare criticamente il contenuto, sintetizzarlo, organizzarlo in una struttura narrativa coerente, e comunicarlo in modo chiaro e coinvolgente. Questo processo richiede un livello di elaborazione che favorisce la comprensione profonda e la ritenzione a lungo termine, promuovendo un maggiore *ownership* del percorso formativo.

La produzione di podcast sviluppa competenze trasversali fondamentali che vanno oltre i contenuti disciplinari specifici. da Silva et al. (2024) hanno documentato che "gli studenti che costruiscono podcast come attività didattica sviluppano competenze come pianificazione, ricerca, scrittura e lavoro di squadra". Queste sono abilità del XXI secolo particolarmente valorizzate nel mondo professionale: gestione di progetti complessi, collaborazione in team multidisciplinari, comunicazione efficace attraverso media digitali, pensiero critico nella selezione delle fonti.

La produzione di podcast può essere utilizzata come efficace strumento di valutazione autentica. Johnson (2008) osserva che "i podcast generati dagli studenti offrono opportunità per l'apprendimento profondo e la valutazione autentica", consentendo di valutare non solo la padronanza dei contenuti ma anche le competenze comunicative, creative e tecniche in un contesto applicativo realistico. La ricerca ha documentato benefici specifici in diversi domini. Yeh (2021) ha rilevato miglioramenti nelle competenze linguistiche orali, osservando che "dopo aver realizzato podcast, e dal loro primo all'ultimo

podcast, gli studenti hanno avuto una migliore fluidità e accuratezza nel parlato in cui hanno dimostrato alcuni cambiamenti significativi, come avere meno pause e parole pronunciate male". Per essere efficace, questa attività richiede una struttura pedagogica chiara: linee guida esplicite sui criteri di qualità, supporto tecnologico adeguato (formazione su software di registrazione ed editing), *scaffolding* progressivo (da esercizi guidati a progetti autonomi), e metodi di valutazione che valorizzino sia il processo creativo sia il prodotto finale, includendo autovalutazione, *peer review* e feedback formativo del docente.

Supervisione rigorosa

La crescente facilità di produzione di podcast, specialmente con l'avvento di strumenti di Intelligenza Artificiale Generativa (GenAI), rende indispensabile un rigoroso controllo di qualità per garantire l'accuratezza scientifica e l'efficacia pedagogica dei contenuti. Strumenti come NotebookLM di Google permettono oggi di generare automaticamente podcast conversazionali a partire da documenti testuali, democratizzando la produzione ma introducendo nuovi rischi. Anche quando la qualità complessiva di un podcast generato da AI è valutata come buona, una revisione esperta può rivelare problematiche significative. Desmedt et al. (2025) hanno osservato che "la qualità del contenuto era generalmente valutata come buona, ma gli esperti hanno notato alcune imprecisioni e semplificazioni eccessive". Queste semplificazioni possono distorcere concetti complessi, mentre le imprecisioni, anche minori, compromettono la credibilità educativa del materiale.

L'AI generativa tende a produrre contenuti plausibili ma non necessariamente accurati, un fenomeno noto come "allucinazione" che richiede vigilanza costante. Sheikh (2025) ha affermato esplicitamente che "l'integrazione dell'AI nella creazione di contenuti educativi richiede meccanismi sistematici di controllo della qualità per garantire l'efficacia pedagogica e l'accuratezza dei contenuti". Questi meccanismi devono includere revisione da parte di esperti disciplinari, verifica delle fonti citate, valutazione dell'allineamento pedagogico con gli obiettivi formativi, e test con utenti reali prima della distribuzione su larga scala.

Per guidare questo processo in modo strutturato, possono essere utilizzati modelli di valutazione come il PICRAT (Passive, Interactive, Creative / Replace, Amplify, Transform), un framework che valuta non solo l'aspetto tecnologico ma anche la trasformazione pedagogica prodotta dall'integrazione dell'AI. Mehta et al. (2024) hanno illustrato come "usando l'esempio di NotebookLM e, specificamente, la funzione auto-podcast, possiamo illustrare come l'educazione medica può utilizzare il modello PICRAT per valutare gli strumenti GenAI per l'integrazione tecnologica". Questo approccio sistematico aiuta a

distinguere tra usi dell'AI che semplicemente sostituiscono pratiche esistenti (Replace) e quelli che amplificano o trasformano genuinamente l'esperienza di apprendimento (Amplify, Transform).

L'uso dell'AI solleva questioni etiche importanti che vanno oltre l'accuratezza tecnica. Baskara et al. (2024) hanno evidenziato che l'uso di AI generativa nella produzione di podcast educativi "ha sollevato importanti questioni sull'etica della paternità e della protezione dei dati e sulla continua necessità di input e supervisione umana". La paternità diventa ambigua quando un podcast è co-creato da umani e AI: chi è responsabile dei contenuti? Come va citata l'AI negli attributi? La protezione dei dati è critica quando si caricano materiali didattici su piattaforme cloud per la generazione automatica. Questi dilemmi confermano la necessità di un input e di una supervisione umana continui per garantire non solo la correttezza dei contenuti, ma anche la responsabilità etica del processo produttivo, mantenendo il docente come garante ultimo della qualità e dell'integrità formativa.

Conclusioni

L'integrazione dei podcast negli ambienti di istruzione superiore e formazione continua ha guadagnato notevole attenzione (Barton et al., 2025). Le evidenze accademiche ne confermano i benefici pedagogici, primariamente legati alla loro natura asincrona e flessibile (Celaya et al., 2019; Jalali et al., 2011; Kamel Boulos et al., 2006; Saidelles et al., 2019), che facilita l'apprendimento autodiretto (*self-paced*), integrato nello stile di vita (Barton et al., 2025; Jansen et al., 2020; Lee & Chan, 2007). La possibilità di riascoltare il materiale (Benner et al., 2025; Saidelles et al., 2019) e l'apprendimento autoregolato che ne deriva (Broadbent et al., 2021) sono correlati a una migliore ritenzione della conoscenza (Barton et al., 2025) e a un elevato coinvolgimento emotivo, cognitivo e comportamentale (Barton et al., 2025; Benner et al., 2025; Celaya et al., 2019; Chen, 2017).

Tuttavia, l'efficacia non è automatica, ma dipende criticamente da un *design* pedagogico intenzionale (Barton et al., 2025). I podcast sono più efficaci come risorse supplementari per arricchire l'insegnamento, piuttosto che come meri sostituti della lezione tradizionale (Hew, 2009; McGarr, 2009). La qualità poggia su credibilità, contenuto e design (Benner et al., 2025; Kakhki et al., 2025; Thoma et al., 2018), e sull'uso di elementi strutturali come lo *storytelling* (Moitra, 2017; Sanchez & Granado, 2024; Yahagi et al., 2025).

L'avvento dell'IA generativa (AIGP) sta trasformando la produzione (Dihan et al., 2025; Silva Cunha, 2025; Zahoor & Sofi, 2021), permettendo la creazione scalabile di contenuti personalizzati (Gu et al., 2025; Liu, 2023; Maity & Deroy, 2024). Emergono però sfide significative relative alla mancanza di *peer-review* e al rischio di inconsistenza qualitativa (Benner et al., 2025; Thoma et al., 2018; Zahoor & Sofi, 2021). Persistono inoltre limiti infrastrutturali (Benner et al., 2025; Dyson et al., 2009; Saidelles et al.,

2019) e la potenziale monotonia delle voci sintetiche, che possono mancare di risonanza emotiva (Benner et al., 2025; Cambre et al., 2020; Gu et al., 2025; Wang et al., 2025).

Se in azienda trattiamo il podcast come una lezione registrata, perdiamo l'unico vantaggio: essere un acceleratore di buone pedagogie. Le ricerche sui Content Acquisition Podcasts mostrano che segmenti brevi e carico cognitivo sotto controllo migliorano conoscenze e riducono lo sforzo rispetto al testo o alle catture di lezione (Kennedy et al., 2016). Gli AIGP spingono oltre: personalizzano ritmo, esempi e micro-quiz sul profilo dell'ascoltatore, con esiti promettenti su apprendimento e gradimento in valutazioni controllate (Do et al., 2024). Il design è decisivo: compressione temporale moderata, script "segnalato" e capitoli navigabili aumentano efficienza senza sacrificare comprensione (Pastore & Ritzhaupt, 2015).

Nei contesti organizzativi questo diventa onboarding, safety refresh, upskilling continuo e coaching audio negli interstizi della giornata. Il mercato e i pubblici ci sono: la spesa cresce e in Italia il podcast intercetta pubblici istruiti e difficili da ingaggiare (Interactive Advertising Bureau & PwC, 2024; Ipsos, 2024b). La mossa strategica è pensare "sistemi di pratica": episodi-capsula, prompt riflessivi, esercizi on-the-job e metriche collegate al LMS. Governance e qualità non si delegano all'AI: servono una redazione L&D, peer review, source grounding e A/B test sugli outcome. In sintesi, i podcast — specie gli AIGP — non sostituiscono la formazione: la amplificano, se progettati per ridurre il superfluo e chiudere il cerchio con misure di impatto (Koehler et al., 2024).

Bibliografia

Alam, F., Boet, S., Piquette, D., Lai, A., Perkes, S. J., & LeBlanc, V. R. (2016). E-learning optimization: The relative and combined effects of mental practice and modeling on enhanced podcast-based learning—A randomized controlled trial. *Advances in Health Sciences Education, 21*(4), 789–802. https://doi.org/10.1007/s10459-016-9666-9

Apple. (2005a, 28 giugno). *Apple takes podcasting mainstream* [Press release]. https://www.apple.com/newsroom/2005/06/28Apple-Takes-Podcasting-Mainstream/

Apple. (2005b, 30 giugno). *iTunes podcast subscriptions top one million in first two days* [Press release]. https://www.apple.com/newsroom/2005/06/30iTunes-Podcast-Subscriptions-Top-One-Million-in-First-Two-Days/

Apple. (2014, 17 settembre). *iOS 8 is available today* [Press release]. https://www.apple.com/newsroom/2014/09/17iOS-8-is-available-today/

Barabasz, A. F. (1968). A study of recall and retention of accelerated lecture presentation. *Journal of Communication, 18*(3), 283-287.

Barton, M. J., Okada, M., & Todorovic, M. (2025). Podcasts in health education—Insights from a scoping review and survey. *Anatomical Sciences Education*. https://doi.org/10.1002/ase.70037

Baskara, F. X. R., Puri, A. D., & Mbato, C. L. (2024). Exploring the use of generative AI in student-produced EFL podcasts: A qualitative study. *Language Teaching Research Quarterly*. https://doi.org/10.32038/ltrq.2024.43.05

Benner, D., Rauch, J., Janson, A., & Leimeister, J. M. (2025). An explorative diary study of AI-generated podcasts in university education: Benefits, challenges, and future directions. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. ACM.

Benner, S., et al. (2025). Explorative diary study: AI-podcasts in university education. *Educational Technology Research, 45*(2), 123–145.

Berk, J., Trivedi, S. P., Watto, M., Williams, P., & Centor, R. (2020). Understanding the use and perceived impact of a medical podcast: Qualitative study. *JMIR Medical Education, 6*(2), e17569. https://doi.org/10.2196/17569

Berry, R. (2006). Will the iPod kill the radio star? Profiling podcasting as radio. *Convergence, 12*(2), 143–162. https://doi.org/10.1177/1354856506066522

Berry, R. (2016). Podcasting: Considering the evolution of the medium and its association with the word 'radio'. *The Radio Journal: International Studies in Broadcast & Audio Media, 14*(1), 7–22. https://doi.org/10.1386/rjao.14.1.7_1

Bianchi, W. (2002). The Wisconsin School of the Air: Success story with implications. *Educational Technology and Society, 5*(1), 141-147.

Bianchi, W. (2008). *Schools of the air: A history of instructional programs on radio in the United States*. McFarland.

Boehle, S. (2007). iPod Corporation: A look inside Capital One's trailblazing audio learning program. *Training, 44*(8), 17–19.

Bosch, A., & Tilson, T. (1991). *Interactive radio instruction: Developing instructional methods*. Education Development Center.

Bottomley, A. J. (2015). Podcasting: A decade in the life of a "new" audio medium. *Journal of Radio & Audio Media, 22*(2), 164–169. https://doi.org/10.1080/19376529.2015.1082880

Broadbent, J., Sharman, S., Panadero, E., & Fuller-Tyszkiewicz, M. (2021). How does self-regulated learning influence formative assessment and summative grade? Comparing online and blended learners. *The Internet and Higher Education, 50*, 100805. https://doi.org/10.1016/j.iheduc.2021.100805

Caldwell, K. E., Zarate-Rodriguez, J. G., Fox, J., Yaeger, L., & Wise, P. E. (2024). Listen up: A systematic review of the utilization and efficacy of podcasts for medical education. *Medical Education, 58*(4), 234–251.

Calheiros, P. R. V., Silva, A. N., Marques, P. A. O., Sousa, C. S., Cordeiro, M. A. S., & Pinto, F. J. M. (2024). Use of podcasts for health education: A scoping review. *Revista Brasileira de Enfermagem, 77*(1), e20230238. https://doi.org/10.1590/0034-7167-2023-0238

Cambre, J., Colnago, J., Maddock, J., Tsai, J. Y., & Kaye, J. (2020). Choice of voices: A large-scale evaluation of text-to-speech voice quality for long-form content. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*.

Carmi, N. (2023). *Professionals' attitudes towards podcasts as a tool for professional development*. [Tesi di Dottorato non pubblicata], Università di Haifa.

Carr, V. B. (2012). Asynchronous learning. In N. M. Seel (Ed.), *Encyclopedia of the Sciences of Learning* (p. 984). Springer. https://doi.org/10.1007/978-1-4419-1428-6_984

Castro-Alonso, J. C., Ayres, P., & Sweller, J. (2021). Instructional designs to optimize learning in multimedia environments: A systematic review. *Educational Research Review, 33*, 100388. https://doi.org/10.1016/j.edurev.2021.100388

Celaya, I., Ramírez-Montoya, M. S., Naval, C., & Arbués, E. (2019). The educational potential of the podcast. *Proceedings of the Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality*. https://doi.org/10.1145/3362789.3362932

Che, A. M. (2023). Let the Students Speak: Using Podcasts to Promote Student Voice and Engagement in an International Studies Classroom in China. *Journal of Political Science Education*. Pubblicazione anticipata online. https://doi.org/10.1080/15512169.2023.2189122

Chen, I. S. (2017). Computer self-efficacy, learning performance, and the mediating role of learning engagement. *Computers in Human Behavior, 72*, 362–370.

Corbeil, J. R., & Valdes-Corbeil, M. E. (2007). Are you ready for mobile learning? *EDUCAUSE Quarterly, 30*(2), 51-58. https://er.educause.edu/articles/2007/4/are-you-ready-for-mobile-learning

Cunado, J., Gijon, C., & Clara, M. (2022). *Employees' attitudes toward using podcasts for continuous professional development*. [Manoscritto non pubblicato], Dipartimento di Psicologia del Lavoro, Università di Barcellona.

da Silva, E. R. V., Beça, P., & Aresta, M. (2024). Using podcast as didactic activity in classroom: A literature review. In *2024 IEEE 24th International Conference on Advanced Learning Technologies (ICALT)* (pp. 242-244). https://doi.org/10.1109/ICALT61570.2024.00077

De Gagne, J. C., Park, H. K., Hall, K., Woodward, A., Yamane, S. S., & Kim, S. S. (2019). Microlearning in health professions education: Scoping review. *JMIR Medical Education, 5*(2), e13997. https://doi.org/10.2196/13997

Deshmukh, A., Pandya, D., & Sahasrabudhe, S. (2025). Tuning into learning: Student narratives from an augmented podcasts intervention. *International Journal of Continuing Education, 52*(2). https://doi.org/10.47852/bonviewijce52026135

Desmedt, C., Budts, W., De Vos, M., & Moons, P. (2025). Artificial intelligence-generated podcasts open new doors to make science accessible: A mixed-method evaluation of quality and the potential for science communication. *European Journal of Cardiovascular Nursing, 24*(Suppl_1). https://doi.org/10.1093/eurjcn/zvaf122.081

Désiron, J. C., & Schneider, S. (2024). Listen closely: Prosodic signals in podcast support learning. *Computers & Education*. https://doi.org/10.1016/j.compedu.2024.105051

Dihan, Q. A., Nihalani, B. R., Tooley, A. A., & Elhusseiny, A. M. (2025). Eyes on Google's NotebookLM: Using generative AI to create ophthalmology podcasts with a single click. *Eye, 39*, 215–216. https://doi.org/10.1038/s41433-024-03481-8

Do, T., Bin Shafqat, U., Ling, E., & Sarda, N. L. (2024). *PAIGE: Examining learning outcomes and experiences with personalized AI-generated educational podcasts*. arXiv. https://doi.org/10.48550/arxiv.2409.04645

Duke Today Staff. (2004, 19 luglio). *Duke to give Apple iPods to first-year students for educational use*. Duke Today. https://today.duke.edu/2004/07/ipods_0704.html

Dupont, M., et al. (2025). Evaluating podcasts for OSCE training: GAI-powered simulation in nephrology education. *Medical Teacher, 47*(3), 156–163.

Dyson, L. E., Litchfield, A., Lawrence, E., Raban, R., & Leijdekkers, P. (2009). Advancing the m-learning research agenda for active, experiential learning: Four case studies. *Australasian Journal of Educational Technology, 25*(2). https://doi.org/10.14742/ajet.1153

Errabo, D. D., Breña de la Rosa, A., & Gonzales, L. (2024). Optimizing differentiated podcasts to promote students' self-regulation and engagement, self-efficacy and performance in asynchronous learning. *Journal of Research in Innovative Teaching & Learning*. https://doi.org/10.1108/JRIT-02-2024-0039

Foulke, E. (1968). Listening comprehension as a function of word rate. *Journal of Communication, 18*(3), 198-206.

Gast, S., & Shifrin, M. M. (2024). Everyone Is Listening: Podcasts as an Innovative Educational Approach in Graduate Level Education. *Journal of Nursing Education, 43*(2), 122-01. https://doi.org/10.3928/01484834-20240122-01

Ghareeb, A. (2017). *Investigating employees' attitudes toward using podcasts for continuous professional development*. [Tesi di Master non pubblicata], Università del Cairo.

Gu, Q. C., Hickey, D., & Ryokai, K. (2025). When AI tells their story: Researchers' reactions to AI-generated podcasts as a tool for communicating research. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*.

Hammersley, B. (2004, 12 febbraio). Audible revolution. *The Guardian*. https://www.theguardian.com/media/2004/feb/12/broadcasting.digitalmedia

Hew, K. F. (2009). A review of the feasibility of podcasting in K-12 and higher education. *Educational Technology Research and Development, 58*(3), 351–370. https://doi.org/10.1007/s11423-008-9180-3

Interactive Advertising Bureau & PwC. (2024). *U.S. podcast advertising revenue study*. https://www.iab.com/wp-content/uploads/2024/05/IAB_US_Podcast_Advertising_Revenue_Study_FY2023_May_2024.pdf

Ipsos. (2024a). *Sbloccare il potenziale del podcast: Dalla passione alla strategia*. https://www.ipsos.com/it-it/sbloccare-potenziale-podcast-passione-strategia

Ipsos. (2024b). *Digital Audio Survey 2024: Podcast strategy*. https://www.ipsos.com/sites/default/files/ct/news/documents/2024-10/Ipsos%20Digital%20Audio%20Survey%202024_Deck.pdf

Jafarian, N. R., & Kramer, A. W. (2024). *Sonic Scholar: Auditory learning to improve student outcomes*. PsyArXiv. https://doi.org/10.31234/osf.io/ac2zq

Jalali, A., Leddy, J., Gauthier, M., Sun, R., Hincke, M., & Carnegie, J. (2011). Use of podcasting as an innovative asynchronous e-learning tool for students. *US China Education Review, 6*, 6741–6748.

Jansen, R. S., van Leeuwen, A., Janssen, J., Conijn, R., & Kester, L. (2020). Supporting learners' self regulated learning in massive open online courses. *Computers and Education, 146*, 103771. https://doi.org/10.1016/j.compedu.2019.103771

Jia, S., Bit, S., Searls, E., Lauber, M. V., Fan, P., Wang, W. M., Claus, L. A., Jasodanand, V. H., Veerapaneni, D., Au, R., & Kolachalama, V. B. (2025). PodGPT: An audio-augmented large language model for research and education. *npj Biomedical Innovations, 2*, Articolo 26. https://doi.org/10.1038/s44385-025-00022-0

Johnson, C. G. (2008). Student-generated podcasts for learning and assessment. In *Proceedings of the 13th annual conference on Innovation and technology in computer science education* (pp. 371-371). https://doi.org/10.1145/1595356.1595371

Ju, Z., Yang, D., Yu, J., Shen, K., Leng, Y., Wang, Z., Tan, X., Zhou, X., Qin, T., & Li, X. (2025). *MoonCast: High-quality zero-shot podcast generation*. arXiv preprint. https://doi.org/10.48550/arXiv.2503.14345

Kakhki, M. S., Mirghafourvand, M., Jafarimanesh, H., & Asghari, E. (2025). The effectiveness of podcasts on learning outcomes in medical education: A systematic review. *BMC Medical Education, 25*(690).

Kamel Boulos, M. N., Maramba, I., & Wheeler, S. (2006). Wikis, blogs and podcasts: a new generation of Web-based tools for virtual collaborative clinical practice and education. *BMC Medical Education, 6*(1), 41.

Karla, J. (2020). Kaffeepause-Podcast: Erfahrungen mit der Produktion und dem Einsatz von Podcasts für die Wissenschaftskommunikation in der Lehre. *Kommunikation Gesundheit Gesellschaft, 21*(2), 633. https://doi.org/10.15460/KOMMGES.2020.21.2.633

Kay, R. (2014). Developing a framework for creating effective instructional video podcasts. *International Journal of Emerging Technologies in Learning (iJET), 9*(1), 22–30. https://doi.org/10.3991/ijet.v9i1.3335

Kelly, J. M., Perseghin, A., Dow, A. W., Trivedi, S. P., Rodman, A., & Berk, J. (2022). Learning through listening: A scoping review of podcast use in medical education. *Academic Medicine, 97*(7), 1079-1083. https://doi.org/10.1097/ACM.0000000000004565

Kennedy, M. J., Hirsch, S. E., Dillon, S. E., & Rabideaux, N. (2016). Using content acquisition podcasts to increase student knowledge and to reduce perceived cognitive load. *Teaching of Psychology, 43*(2), 153–161. https://doi.org/10.1177/0098628316636295

Kennedy, M. J., Thomas, C. N., Aronin, S., Newton, J. R., & Lloyd, J. W. (2014). Improving teacher candidate knowledge using content acquisition podcasts. *Computers & Education, 70*, 116–127. https://doi.org/10.1016/j.compedu.2013.08.010

Koehler, D., Serth, S., & Meinel, C. (2024). Promoting content variety in MOOCs: increasing learning outcomes with podcasts. *Frontiers in Education, 9*, 1339142. https://doi.org/10.3389/feduc.2024.1339142

Kokala, M. (2024). Revolutionizing content creation: AI-podcast with NotebookLM. *Educational Technology Society, 27*(3), 45–58.

Laupichler, M. C., Aster, A., Soyubey, L., et al. (2025). Influence of Active Production Versus Passive Consumption of Podcasts on Medical Students' Learning Outcomes. *The Clinical Teacher*. Pubblicazione anticipata online. https://doi.org/10.1111/tct.70029

Lee, M., & Chan, A. (2007). Pervasive, lifestyle-integrated mobile learning for distance learners: An analysis and unexpected results from a podcasting study. *Open Learning: The Journal of Open and Distance Learning, 22*(3), 201–218.

Lehmann, E. (2005, 1 marzo). Adam Curry wants to make you an iPod radio star. *Wired*. https://www.wired.com/2005/03/curry/

Liu, T. (2023). The impact of Generative AI (GenAI) on practices, policies and research direction in education: A case of ChatGPT and Midjourney. *Interactive Learning Environments*. https://doi.org/10.1080/10494820.2023.2253861

Llinares, D., Fox, N., & Berry, R. (Eds.). (2018). *Podcasting: New aural cultures and digital media*. Palgrave Macmillan. https://doi.org/10.1007/978-3-319-90056-8

Maity, S., & Deroy, A. (2024). *Generative AI and its impact on personalized intelligent tutoring systems*. arXiv preprint arXiv:2410.10650. https://doi.org/10.35542/osf.io/kawr5

Mayer, R. E. (2009). *Multimedia learning* (2a ed.). Cambridge University Press.

McGarr, O. (2009). A review of podcasting in higher education: its influence on the traditional lecture. *Australasian Journal of Educational Technology, 25*(3), 309–321. https://doi.org/10.14742/ajet.1136

Meden, A., Meden, K., & Sparl, P. (2024). *Podcasts as a method for explicitating tacit knowledge in organizations*. [Paper presentato alla conferenza], 15th European Conference on Knowledge Management, Lubiana, Slovenia.

Mehta, N., Agrawal, A., Benjamin, J., Mehta, S., MacNeill, H., & Masters, K. (2024). Pedagogy and generative artificial intelligence: Applying the PICRAT model to Google NotebookLM. *Medical Teacher*. Pubblicazione anticipata online. https://doi.org/10.1080/0142159x.2024.2418937

Miller, A. K., & Uphold, N. M. (2021). The effects of content acquisition podcasts on in-service science teachers' vocabulary instruction. *Journal of Science Teacher Education, 32*(3), 301-318. https://doi.org/10.1080/1046560X.2020.1828248

Moitra, K. (2017). STEM to social awareness: Connecting the dots through audio storytelling with podcasts. In *Unplugging the Classroom: Teaching With Technology to Promote Student Lifelong Learning* (pp. 159–167). Elsevier Inc. https://doi.org/10.1016/B978-0-08-102035-7.00012-6

Morris, K. (2024). Artificial audio: Emerging uses of AI in podcasting. *Digital Media Society, 12*(4), 234–251.

Morris, L. D., Stephens, J. P., & Purtill, T. (2023). Let's talk about it: How internal organizational podcasts foster employee voice, connection, and continuous learning. *Organization Management Journal*. Pubblicazione anticipata online. https://doi.org/10.1108/OMJ-06-2022-1598

Moulton, J. (1994). *Interactive radio instruction: Broadening the definition*. Education Development Center. https://eric.ed.gov/?id=ED390625

Münch-Harrach, D., Hampe, W., & Kothe, C. (2013). Audiopodcasts im Biochemiepraktikum – Kostengünstiges eLearning in bewährtem Hörfunkformat. *GMS Zeitschrift für Medizinische Ausbildung, 30*(4), Doc887. https://doi.org/10.3205/ZMA000887

Murtinugraha, R. E., Arifah, S., Saleh, R., Arthur, R., & Ostiyani, N. R. (2025). Podcast-based learning: evaluating the success of learning media development in vocational education. *Jurnal EduScience (JES), 12*(4). https://doi.org/10.36987/jes.v12i4.7400

Nie, M., Cashmore, A., & Cane, C. (2010). *The educational value of podcasting in a technology-based university*. [Paper presentato], 11th International Conference on Educational Technology (ICET), Parigi, Francia.

Noetica. (2024). *Podcast in Italia: Numeri e tendenze*. https://www.noetica.it/podcast-in-italia-numeri-e-tendenze/

Ottens, L. (1963). *Compact cassette tape recorder [Philips EL 3300]*. Philips Company Archives. https://www.philips.com/a-w/about/news/media-library/20190101-First-Philips-cassette-recorder-1963.html

Pastore, R. S. (2010). The effects of diagrams and time-compressed instruction on learning and learners' perceptions of cognitive load. *Educational Technology Research and Development, 58*(5), 485-505. https://doi.org/10.1007/s11423-009-9133-6

Pastore, R., & Ritzhaupt, A. D. (2015). Using time-compression to make multimedia learning more efficient: Current research and practice. *TechTrends, 59*(2), 66–74. https://doi.org/10.1007/s11528-015-0841-2

Pastore, R. S., & Ritzhaupt, A. D. (2015). Using time-compression to make multimedia learning more efficient: Current research and practice. *TechTrends, 59*(2), 66-74. https://doi.org/10.1007/s11528-015-0841-2

Peterson, C. (2024). *Podcast ad revenue soars to $2.4 billion in 2024* [Post LinkedIn]. https://www.linkedin.com/posts/originalchrispeterson_the-iab-and-pwc-just-released-their-2024-activity-7320076338327932928-hbUj

Peterson, J., et al. (2025). User experience with AI-generated research education podcasts in healthcare. *Journal of Medical Internet Research, 27*(4), e45123.

Pinto, A., et al. (2023). The podcast as strategy to professional development and knowledge transfer in health care planning. *BMC Health Services Research, 23*, 1043.

Questioni d'Orecchio. (2024). *Chi ascolta i podcast, in Italia*. https://www.questionidorecchio.it/chi-ascolta-i-podcast-in-italia/

Rain News. (2024). *"A return to strong growth" -- IAB Internet Advertising Revenue Report*. https://rainnews.com/a-return-to-strong-growth-iab-internet-advertising-revenue-report/

Riddell, J. C., Robins, L., Sherbino, J., Brown, A., & Ilgen, J. (2021). Residents' perceptions of effective features of educational podcasts. *Western Journal of Emergency Medicine, 22*(1), 26–32. https://doi.org/10.5811/WESTJEM.2020.10.49135

Ritzhaupt, A. D., & Barron, A. E. (2008). Effects of time-compressed narration and representational adjunct images on cued-recall, content recognition, and learner satisfaction. *Journal of Educational Computing Research, 39*(2), 161-184. https://doi.org/10.2190/EC.39.2.d

Ritzhaupt, A. D., Barron, A. E., & Kealy, W. A. (2011). Conjoint processing of time-compressed narration in multimedia instruction: The effects on recall, but not recognition. *Journal of Educational Computing Research, 44*(2), 203-217. https://doi.org/10.2190/EC.44.2.e

Saidelles, T., Schmitt, J. D. A. C., Barin, C. S., Ellensohn, R. M., & Santos, L. M. A. (2019). Investigando as potencialidades do uso do Software Audacity como ferramenta de flexibilização da aprendizagem. *Redin-Revista Educacional Interdisciplinar, 8*(1).

Sanchez, A., & Granado, A. (2024). Telling science/health stories in audio. In K. Walsh-Childers & M. McKinnon (Eds.), *Palgrave Handbook of Science and Health Journalism*. Palgrave Macmillan. https://doi.org/10.1007/978-3-031-49084-2_11

Schwertz, S. B. B. (2025). Podcast na educação: uma ferramenta tecnológica, acessível e dinâmica para promover a aprendizagem dos estudantes. In *Educação e Tecnologia: Perspectivas Contemporâneas* (pp. 123-145). https://doi.org/10.51891/978-65-6054-244-0-07

Shamburg, C., Amerman, T., & Wister, J. (2023). How workplace listening to podcasts supports informal learning, creativity, and well-being. *Journal of Workplace Learning, 35*(9), 1-17. https://doi.org/10.1108/JWL-01-2023-0005

Sheikh, S. M. S. (2025). Empowering learning: Crafting educational podcasts with GEN AI. *International Journal For Science Technology And Engineering*. https://doi.org/10.22214/ijraset.2025.69144

Silva Cunha, J. L. (2025). Multimodal AI as a pedagogical tool: Early impressions of Google's NotebookLM in oral diseases. *Oral Diseases*.

Spinelli, M., & Dann, L. (2019). *Podcasting: The audio media revolution*. Bloomsbury Academic.

Sullivan, J. L. (2019). The platforms of podcasting: Past and present. *Social Media + Society, 5*(4), 1–12. https://doi.org/10.1177/2056305119880002

Tam, C. O. (2012). The effectiveness of educational podcasts for teaching music and visual arts in higher education. *Research in Learning Technology, 20*, 14919. https://doi.org/10.3402/rlt.v20i0.14919

Thoma, B., Dougherty, A., & Gaglani, S. (2018). The potential use of podcasts in medical education. In M. A. T. van Ede, M. A. W. C. L. van den Hoven, & T. M. C. J. T. van Ede (Eds.), *Clinical Medical Education: Principles and Practice* (pp. 549–559).

Vrzalková, K. (2023). The artificial intelligence behind the podcast microphone: AI-generated content in podcasting. *Media and Identity, 44*, 128–145. https://doi.org/10.34135/mmidentity-2023-44

Wang, F., Cheung, A. C. K., Chai, C. S., & Liu, J. (2025). Development and validation of the perceived interactivity of learner-AI interaction scale. *Education and Information Technologies, 30*(4), 4607–4638.

Wärnestål, P., & Sjöberg, F. (2020). *Hybrid podcasts: Combining on-demand audio with live webinars for corporate training*. [Report tecnico], Jönköping AI Lab, Svezia.

Xiao, Y., Xue, L., He, L., Chen, X., Chiu, A. Y. F., Tian, W., Zhang, S., Kong, Q., Zhu, X., Xue, W., & Le, T. (2025). *PODEVAL: A multimodal evaluation framework for podcast audio generation* (arXiv:2510.00485v1). [Preprint]. arXiv. https://arxiv.org/abs/2510.00485

Yahagi, Y., Chujo, R., Harada, Y., Han, C., Sugiyama, K., & Naemura, T. (2025). Paperwave: Listening to research papers as conversational podcasts scripted by LLM. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*, 10.

Yeh, C. C. (2021). The effects of producing podcasts on EFL learners' speaking skills. *Computer Assisted Language Learning*, 1-25. https://doi.org/10.1080/09588221.2021.1963321

Zahoor, S., & Sofi, S. A. (2021). Automatic podcast generation. *Journal of University of Shanghai for Science and Technology, 23*(10), 22–28. https://doi.org/10.51201/jusst/21/09700

I quaderni di

Diritto del Digitale

Sezione Diritto del digitale

EDITORIALE

In un'epoca caratterizzata dall'accelerata evoluzione delle tecnologie digitali, l'intersezione tra diritto e innovazione riveste un'importanza cruciale. L'intelligenza artificiale, la *governance* degli algoritmi, la cybersicurezza, la protezione dei dati personali, l'identità digitale, la responsabilità delle piattaforme e la progressiva automazione delle decisioni pubbliche sono solo alcune delle coordinate che delineano il nuovo orizzonte problematico del giurista contemporaneo.

Dinanzi a tali fenomeni, infatti, il diritto non può limitarsi a rincorrere la tecnica, adattandosi con inerzia alle sue torsioni. Occorre, piuttosto, che si riappropri della sua funzione regolatoria, che non può prescindere dalla precedente comprensione delle trasformazioni sociali e istituzionali che accompagnano il mutamento tecnologico.

È su queste basi che nasce l'idea di dare vita ad una sezione scientifica nell'ambito dei *Quaderni di Agendadigitale.eu*, dedicata al "Diritto del digitale". La creazione di un nuovo spazio di approfondimento scientifico non è mai un fatto meramente editoriale. È, al contrario, un atto che si inscrive entro una precisa congiuntura culturale e istituzionale che, in questo caso, deriva dalla consapevolezza che il diritto deve adeguarsi rapidamente al mutamento profondo delle strutture stesse del reale e, con esse, dei paradigmi giuridici chiamati a comprenderlo, regolamentarlo e – non di rado – anche contrastarlo. Questa sezione nasce, dunque, con una vocazione precisa: offrire uno spazio rigoroso di analisi scientifica, capace di accogliere e valorizzare contributi di studiosi – in prevalenza accademici – impegnati ad approfondire le complessità contemporanee.

Da questo punto di vista, la sezione "Diritto del digitale" potrà rappresentare la sede per approfondire, sul piano scientifico, gli studi portati avanti dal Centro di ricerca interdipartimentale Europeo di Studi Avanzati sull'Innovazione Digitale (*Research Center IDEAS – Innovation Digital European Advanced Studies*), dell'Università degli Studi Roma Tre, fondato per promuovere iniziative di carattere multidisciplinare che riguardino l'innovazione digitale sull'ordinamento. D'altronde, il Centro è nato proprio con l'obiettivo di rappresentare un luogo stabile di dialogo e confronto tra Istituzioni, attori privati ed Università (anche di altri Paesi), per favorire la circolazione delle idee e trovare, insieme, soluzioni ai problemi che cambiamenti del nuovo millennio sta portando con sé, nella prospettiva della tutela dei diritti delle persone e delle imprese, con la necessaria attenzione ai processi normativi.

Siamo convinti che la scienza giuridica debba oggi esercitare un compito di vigilanza e di orientamento: vigilanza, per non abdicare alla propria funzione critica dinanzi alla neutralità presunta della tecnica; orientamento, per contribuire alla definizione di un equilibrio nuovo tra libertà, sicurezza, innovazione e dignità della persona umana. In questo senso, il nostro progetto editoriale non è solo una risposta alle sfide del presente, ma anche un tentativo – forse ambizioso, ma necessario – di costruire una grammatica giuridica del futuro, elaborata con consapevolezza metodologica e profondità teorica.

In un’epoca in cui l’accelerazione digitale rischia di offuscare la riflessione e marginalizzare la misura del diritto, crediamo che vi sia più che mai bisogno di luoghi di pensiero critico e la sezione “Diritto del digitale” nasce per essere uno di quei luoghi.

Proff. Carlo Colapietro e Antonio Iannuzzi

Direttori della sezione “Diritto del digitale”

Cosa cambia se il concetto di dato personale diventa relativo - di ***Giuseppe D'Acquisto***, Funzionario del Garante per la protezione dei dati personali, Titolare dell'insegnamento di intelligenza artificiale presso il Dipartimento di Giurisprudenza dell'Università LUISS Guido Carli

La proposta di modifica del GDPR introduce una nozione più relativa di dato personale e sposta il baricentro sulla concreta capacità identificativa di chi tratta i dati. Il cambiamento promette meno oneri, ma apre nodi tecnici e nuovi rischi di irresponsabilità

Se leggiamo il testo della proposta di modifica del GDPR (Omnibus) attualmente in discussione in sede di trilogo [1], non è difficile accorgersi di un cambio di paradigma per la protezione dei dati che potrebbe essere tanto profondo quanto rapido.

Le novità legislative

Tra le modifiche proposte, quella di maggiore rilievo per l'impatto che ne potrebbe derivare è la revisione della definizione di dato personale da cui discende l'applicazione stessa della disciplina. La proposta, infatti, afferma che un'informazione non assume natura personale in senso universale, ma solo in relazione alla concreta capacità identificativa di cui, per il suo tramite, dispone il soggetto che la detiene, tenuto conto dei mezzi di cui questi può ragionevolmente avvalersi. Né tale qualificazione da riconoscersi a tale soggetto muta per il solo fatto che un eventuale ulteriore detentore in una catena di passaggi successivi del dato disponga, invece, di strumenti idonei all'identificazione.

Si tratta del cosiddetto approccio "relativo" alla nozione di dato personale, che si pone in contrapposizione rispetto all'impostazione tradizionalmente seguita - di matrice sostanzialmente "assoluta" - entro cui si è sviluppata l'applicazione del diritto della protezione dei dati negli ultimi due decenni. In base a quest'ultima, la mera esistenza di un percorso plausibile di identificazione per un singolo soggetto è stata di norma ritenuta sufficiente per qualificare un'informazione come dato personale *erga omnes*, in modo pressoché indipendente dal contesto soggettivo e tecnologico di riferimento.

Questo approccio assoluto all'interpretazione del concetto di dato personale, se da un lato ha ridotto le ambiguità interpretative e massimizzato le garanzie per gli interessati, dall'altro è stato oggetto di crescenti critiche in ragione dell'ampliamento dell'ambito applicativo della disciplina, che avrebbe

determinato significativi oneri di conformità per i soggetti economici e una certa rigidità nell'uso dei dati individuata in ultima istanza come un fattore di svantaggio competitivo a livello continentale, in particolare con riguardo allo sviluppo e all'addestramento di sistemi di intelligenza artificiale.

In questo quadro, riassunto in modo il più possibile oggettivo e neutrale, si sono rafforzate le istanze di semplificazione e di ricalibrazione dell'approccio regolatorio, nella prospettiva di coniugare un elevato livello di tutela dei diritti fondamentali con l'esigenza di non ostacolare lo sviluppo tecnologico e scientifico [2]. La nozione di dato personale è naturalmente emersa come leva strategica centrale: poiché da essa dipende l'applicabilità del GDPR, una revisione dei criteri di identificabilità incide direttamente sull'ambito oggettivo della disciplina. L'obiettivo legislativo dichiarato è dunque mitigare gli oneri di compliance senza tuttavia ridurre le tutele né la fiducia nello sviluppo tecnologico.

Ciò avviene attraverso un approccio più chiaramente relativo nel quale la qualificazione di un'informazione come personale non dipende più dalla mera esistenza, anche solo potenziale o ipotetica, di mezzi di identificazione nella disponibilità di qualsiasi soggetto, ma dalla loro concreta e ragionevole disponibilità per il soggetto che tratta i dati. Se tali mezzi non sono disponibili, i dati possono essere considerati non personali per quel soggetto, con conseguente esclusione del trattamento dall'ambito di applicazione del GDPR.

È dunque opportuno avviare sin da subito una riflessione critica su questo mutamento di paradigma, al fine di orientarne gli sviluppi piuttosto che subirli. Affrontare il tema senza pregiudizi consente di coglierne tempestivamente le implicazioni sistematiche, individuando lo spazio effettivo e i limiti della tutela. Solo attraverso tale esercizio sarà possibile verificare se e in che misura il passaggio a una nozione relativa di dato personale sia effettivamente in grado di semplificare gli adempimenti mantenendo invariato il livello di garanzie e di generare fiducia nello sviluppo tecnologico.

Il problema dell'identificazione

Per qualificare un dato come personale è necessario che esso consenta l'identificazione di una persona fisica. Tale operazione non coincide necessariamente con la conoscenza o la scoperta del nome e cognome dell'interessato. Da tempo, infatti, tanto le corti europee - dalle giurisdizioni nazionali alla Corte di giustizia [3] - quanto le autorità di protezione dei dati [4] [5], sia a livello nazionale sia attraverso il Comitato europeo per la protezione dei dati (EDPB) [6] [7], concordano nel ritenere che un individuo possa considerarsi identificato quando l'uso di un'informazione consenta di indirizzare verso di lui una decisione o un'azione mirata, anche in assenza di identificatori nominativi o di altre specifiche categorie di identificatori.

Occorre dunque chiedersi se il passaggio da una nozione assoluta a una nozione relativa di dato personale agevoli o renda più complessa l'operazione di identificazione, ovvero se relativizzazione e identificazione costituiscano concetti tra loro indipendenti.

In questa prospettiva, può essere utile adottare una formalizzazione minima del problema dell'identificazione. In assenza di un riferimento formale condiviso, il dibattito rischia infatti di scivolare su un piano meramente ideologico. La formalizzazione qui proposta non pretende di essere esaustiva rispetto a tutte le possibili modalità di identificazione [8], ma mira più limitatamente a fornire un terreno di confronto tecnico oggettivo, utile per comprendere come l'identificazione si manifesti in concreto e quali rimedi possano essere ragionevolmente ipotizzati.

La prima situazione di identificazione, che potremmo definire classica, si manifesta come un'operazione di scoperta dell'identità, che si realizza attraverso la ricomposizione di informazioni precedentemente disaccoppiate. In termini formali minimi, essa può essere rappresentata come la concatenazione tra due relazioni (nome, pseudonimo) e (pseudonimo, attributo), in modo da "assorbire" il termine intermedio e ricondurre attraverso la relazione (nome, attributo) un determinato attributo a una persona identificata. Si tratta, evidentemente, di un modello stilizzato, che non esaurisce la varietà delle possibili forme di identificazione, ma che risulta utile come punto di partenza.

Tuttavia, l'identificazione non si esaurisce in questa forma diretta di ricomposizione. Un secondo caso riguarda l'ipotesi in cui un soggetto disponga di più relazioni riferite al medesimo pseudonimo, per esempio (pseudonimo, attributo_1) e (pseudonimo, attributo_2). In tale situazione è possibile effettuare un'operazione di collegamento e "assorbimento" dello pseudonimo, ottenendo una nuova relazione (attributo_1, attributo_2). Anche in assenza del nome, questa combinazione può risultare altamente rivelatrice. Si pensi al caso in cui attributo_1 consista in un insieme di caratteristiche facilmente osservabili (genere, età, altezza, localizzazione, rete di contatti), mentre attributo_2 riguardi informazioni più sensibili o non immediatamente accessibili (stato di salute, situazione giudiziaria o finanziaria, dati genetici). Se il primo insieme di attributi è sufficientemente ricco e distintivo, esso può rendere di fatto accessibile anche il secondo, esponendo l'individuo. Detto diversamente, un attributo originariamente non accessibile diviene conoscibile per effetto del collegamento con attributi più facilmente osservabili.

Infine, può configurarsi una terza forma di identificazione, di natura probabilistica. Si immagini che un soggetto disponga delle relazioni (pseudonimo_1, attributo_1) e (pseudonimo_2, attributo_2). Pur in assenza di una perfetta coincidenza tra pseudonimo_1 e pseudonimo_2, il detentore del dato potrebbe rilevare una somiglianza tra i due - non ci riferiamo a una somiglianza percepibile dallo sguardo umano, ma a una

somiglianza misurabile attraverso indicatori numerici (ad esempio, forme di *cosine similarity* tra c.d. vettori di *embedding* [9]) - e inferire, con un certo grado di probabilità, che le due coppie si riferiscano alla medesima persona. Si tratta di un'inferenza intrinsecamente azzardata, ma non per questo irrilevante: in taluni casi l'assorbimento tra pseudonimo$_1$ e pseudonimo$_2$ può rivelarsi sufficientemente verosimile da consentire l'associazione tra attributi diversi (attributo$_1$, attributo$_2$) – l'uno facilmente osservabile, l'altro inaccessibile - e, conseguentemente, la rivelazione di caratteri o stati nascosti dell'interessato.

Queste ulteriori forme di identificazione mostrano come la riferibilità di un'informazione a una persona non dipenda esclusivamente dalla disponibilità della relazione (nome, pseudonimo), ma possa emergere anche attraverso operazioni di collegamento tra dati e inferenza, senza che sia necessario conoscere il nome (o simili identificatori anagrafici), le quali, in ragione del contesto, possono determinare una sovraesposizione dell'individuo rispetto alle sue aspettative.

Sulla disponibilità di mezzi per l'identificazione

Questa formalizzazione minima del problema dell'identificazione ci consente di affrontare un ulteriore aspetto concettuale decisivo: quello dei "mezzi" di identificazione. Si tratta di una nozione tanto centrale quanto frequentemente fraintesa.

È bene aver presente che non esiste alcun "mezzo" in grado di aumentare, di per sé, l'informazione contenuta in un dato. Rappresentare il "mezzo" come un processo input-output (come un algoritmo di intelligenza artificiale, un programma, un software) che, a partire da un dato in ingresso, generi nuova informazione rispetto a quella originariamente contenuta è un'ipotesi fuorviante e, sotto il profilo teorico, insostenibile. L'identificazione - che spesso intendiamo come una forma di accrescimento della conoscenza relativa a una persona - può avvenire esclusivamente attraverso la combinazione di informazioni parziali riferibili a quella persona, dalla cui integrazione emerge un contenuto informativo più dettagliato rispetto a quello posseduto da ciascun elemento considerato isolatamente.

In questa prospettiva, deve considerarsi "mezzo" di identificazione non uno strumento generico o indeterminato, bensì l'ulteriore porzione di informazione utilizzata per arricchire quella originaria. Riprendendo l'esempio già introdotto, nella struttura (nome, pseudonimo) e (pseudonimo, attributo), la prima coppia costituisce mezzo di identificazione rispetto alla seconda (e viceversa): è infatti attraverso l'assorbimento dello pseudonimo che diviene possibile collegare l'attributo a un soggetto nominativamente determinato, realizzando così un incremento informativo.

Questa lettura consente di ancorare il concetto di "mezzi ragionevolmente utilizzabili" a una dimensione oggettiva e verificabile, coerente con i principi della teoria dell'informazione. Al contrario, interpretazioni più vaghe o indeterminate della nozione di mezzo rischiano di spostare il dibattito su un piano ancora una volta meramente ideologico, privo di adeguati riferimenti analitici e quindi poco utile ai fini di una corretta delimitazione dell'ambito applicativo della disciplina.

Dunque, se un soggetto dispone esclusivamente della relazione (pseudonimo, attributo), ma non ha accesso - né può ragionevolmente ottenerlo - alla relazione (nome, pseudonimo), possiamo anche accettare che egli non sia in grado di ricondurre l'attributo a una persona fisica identificata o identificabile e che l'informazione di cui egli dispone sia qualificata come dato relativamente non personale. Affinché questa conclusione sia sostenibile, è tuttavia necessario assumere che l'accesso alla coppia (nome, pseudonimo) sia irragionevole in senso forte, o perché tecnicamente impedito da misure di sicurezza estremamente efficaci, o giuridicamente precluso, perché per esempio proibito per legge [10], e che la coppia (pseudonimo, attributo) non sia di per sé identificativa anche senza accesso alla coppia (nome, pseudonimo). Si tratta, bisogna rimarcarlo, di una condizione tutt'altro che generale. Per esempio, la semplicissima relazione (id_Garante_abc, ingegnere nato a Palermo) si riferisce a un'unica persona tra gli abitanti del pianeta: chi sta scrivendo questo articolo, e si fa estrema fatica a voler considerare questa relazione come un dato relativamente non personale nelle mani di voi lettori, anche senza che abbiate accesso alla ipotetica coppia (id_Garante_abc, Giuseppe D'Acquisto).

Occorre inoltre osservare che ricondurre il problema dell'identificazione a un tema di sicurezza implica anche rimettere la tenuta della soluzione a fattori eminentemente comportamentali: la scelta e l'effettiva applicazione delle misure di sicurezza, il rispetto della confidenzialità, il grado di fiducia riposto nei soggetti coinvolti. Si tratta, come noto, di un anello intrinsecamente debole, in quanto il comportamento umano è esposto a errori, negligenze e forme di manipolazione. Ne deriva che una soluzione a una questione di protezione dei dati fondata esclusivamente su presidi di sicurezza si presenta, sotto questo profilo, strutturalmente fragile [11].

Tale criticità emerge con particolare evidenza negli altri casi di identificazione considerati. Quando l'identificazione deriva dalla combinazione – certa o probabilistica - di più attributi associati al medesimo pseudonimo o a pseudonimi simili, le misure di sicurezza poste a protezione della relazione (nome, pseudonimo) risultano insufficienti - se non del tutto irrilevanti - ai fini della prevenzione del rischio di identificazione. Questa, infatti, non passa più attraverso la ricostruzione diretta dell'identità nominativa, ma attraverso operazioni di correlazione che prescindono da essa.

In queste ipotesi, la soluzione non può essere affidata a strumenti meramente organizzativi o comportamentali, ma richiede più propriamente interventi di data protection engineering, ossia soluzioni tecniche integrate nel trattamento stesso, che non intervengono sulla coppia (nome, pseudonimo) ma agiscono direttamente sulla struttura dell'informazione (pseudonimo, attributo), indebolendone significativamente la possibilità di combinazioni e assorbimenti indipendentemente da fattori umani, incidendo in tal modo più efficacemente e oggettivamente sul rischio di identificazione.

Punti di irresponsabilità e conclusioni

La relativizzazione della nozione di dato personale introduce, sul piano sistemico, due potenziali "punti di irresponsabilità": uno in capo al titolare del trattamento, l'altro in capo ai soggetti che, a vario titolo, ricevono e riutilizzano i dati.

Dal lato del titolare, il rischio è quello di un affidamento eccessivo sulla sicurezza della relazione (nome, pseudonimo) quale principale - se non esclusivo - strumento di protezione, che potrebbe generare situazioni di incauto oversharing di dati relativamente non personali. La convinzione che una custodia "impenetrabile" di tale coppia, unita alla sottovalutazione delle dinamiche di collegamento tra attributi, consenta di distribuire liberamente e lecitamente una molteplicità di relazioni (pseudonimo, attributo) riferibili in modo certo o verosimile alla medesima persona, costituisce un tipico caso in cui una sicurezza malintesa genera un paradosso da eccesso di controllo. Considerato l'atteso incremento del volume di dati in circolazione, aumenterà verosimilmente anche la quantità di ricombinazioni informative idonee a rivelare - in modo certo o inferenziale - caratteristiche dei singoli interessati originariamente non accessibili o non destinate alla conoscibilità.

Dal lato dei destinatari, il problema si manifesta in forma speculare. La constatazione che un'informazione qualificata come relativamente non personale non muta la propria natura giuridica per il solo fatto che un soggetto successivo nella catena disponga di strumenti idonei all'identificazione potrebbe innescare una naturale attitudine a una condivisione incontrollata; ciò potrebbe dare luogo, come visto, a fenomeni di identificazione "per assorbimento", che non mirano necessariamente al disvelamento di una identità, ma all'accesso a stati o caratteristiche non immediatamente osservabili attraverso la loro combinazione – certa o inferenziale - con attributi più facilmente disponibili.

Non è possibile prevedere se e in quale misura tali punti di irresponsabilità sistemici saranno sfruttati in modo opportunistico o deliberato. Ciò che rileva è che la relativizzazione del concetto di dato personale rende possibile tale sfruttamento, con scarsa possibilità di enforcement *ex-post*. Se l'obiettivo dichiarato del nuovo approccio è quello di alleggerire gli oneri di compliance senza però

ridurre il livello di tutela, occorre intervenire *ex-ante* sul dato in circolazione e "ingegnerizzarlo" affinché le relazioni (pseudonimo, attributo) non siano, per loro natura, facilmente riconducibili alla medesima persona. La tollerabilità, accettata in modo fatalista, di tali ricongiunzioni finirebbe per contraddire l'obiettivo strategico di mantenere invariato il livello di tutela.

Le soluzioni tecniche idonee a evitare un abbassamento complessivo del livello di protezione rispetto all'approccio "assoluto" esistono e consentono un moderato grado di fiducia nella possibilità di coniugare alleggerimento degli oneri e salvaguardia dei diritti, anche in un quadro di relativizzazione del concetto di dato personale. Tuttavia, ciò richiede un investimento significativo e diffuso in competenze, in particolare nel campo dell'anonimizzazione. Il rischio più grave, in questa fase di transizione, è la banalizzazione del problema: trattare la relativizzazione del concetto di dato personale come una mera semplificazione o un tema di sicurezza, senza coglierne le implicazioni strutturali.

Non è necessario reinventare strumenti già noti. Le modalità concrete di attuazione possono evolversi e specializzarsi con il progresso tecnologico e la varietà dei casi di condivisione dei dati. La responsabilità della loro efficace implementazione resta affidata, in ultima analisi, all'accountability dei titolari. Riecheggiano in tal senso, ancora con piena attualità, le indicazioni formulate nel 2014 dal Gruppo di lavoro "Articolo 29" nel parere sulle tecniche di anonimizzazione: evitare singolarità, prevenire la possibilità di collegamenti tra dati, limitare facili inferenze [12].

Bibliografia

[1] Proposta di Regolamento del Parlamento Europeo e del Consiglio che modifica i Regolamenti (UE) 2016/679, (UE) 2018/1724, (UE) 2018/1725 e (UE) 2023/2854 e le Direttive 2002/58/CE, (UE) 2022/2555 e (UE) 2022/2557 per quanto riguarda la semplificazione del quadro legislativo nel settore digitale e che abroga i Regolamenti (UE) 2018/1807, (UE) 2019/1150 e (UE) 2022/868 e la direttiva (UE) 2019/1024 (omnibus digitale), https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=CELEX:52025PC0837

[2] The Draghi report on EU competitiveness, https://commission.europa.eu/topics/competitiveness/draghi-report_en

[3] Sentenza della Corte (Seconda Sezione), 20 dicembre 2017, «Rinvio pregiudiziale – Tutela delle persone fisiche con riguardo al trattamento dei dati personali – Direttiva 95/46/CE – Articolo 2, lettera a) – Nozione di "dati personali" – Risposte scritte fornite da un candidato durante un esame professionale – Annotazioni dell'esaminatore relative a tali risposte – Articolo 12, lettere a) e b) – Portata dei diritti di accesso e di rettifica della persona interessata», causa C-434/16, par. 35, https://eur-lex.europa.eu/legal-content/IT/TXT/HTML/?uri=CELEX:62016CJ0434

[4] Garante per la protezione dei dati personali, Provvedimento del 1° giugno 2023 [doc. web n. 9913795], https://garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9913795

[5] Délibération SAN-2024-013 du 5 septembre 2024, https://www.legifrance.gouv.fr/cnil/id/CNILTEXT000050202759

[6] EDPB Guidelines 01/2025 on Pseudonymisation, https://www.edpb.europa.eu/our-work-tools/documents/public-consultations/2025/guidelines-012025-pseudonymisation_it

[7] Gruppo di lavoro Articolo 29, Parere 4/2007 sul concetto di dati personali - WP 136 - 20 giugno 2007, https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/1487717

[8] G. D'Acquisto, A. Cohen, M. Naldi and K. Nissim (2024), *From Isolation to Identification*. Privacy in Statistical Databases (PSD 2024): 3-17, https://drive.google.com/file/d/1dkpeV0y7Q9CiTuy7hKhsUDKj6RFiVGZt/view

[9] S. Lermen, D. Paleka, J.P. Swanson, M. Aerni, N. Carlini and F. Tramèr (2026), *Large-scale online deanonymization with LLMs*, https://arxiv.org/abs/2602.16800

[10] Sentenza della Corte (Seconda Sezione) del 19 ottobre 2016, Patrick Breyer contro Bundesrepublik Deutschland. Domanda di pronuncia pregiudiziale proposta dal Bundesgerichtshof. Rinvio pregiudiziale – Trattamento dei dati personali – Direttiva 95/46/CE – Articolo 2, lettera a) – Articolo 7, lettera f) – Nozione di “dati personali” – Indirizzi di protocollo Internet – Conservazione da parte di un fornitore di servizi di media online – Normativa nazionale che non consente di prendere in considerazione l’interesse legittimo perseguito dal responsabile del trattamento. Causa C-582/14. https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=CELEX:62014CJ0582

[11] R. Anderson and T. Moore (2006). *The Economics of Information Security*, Science (27 October 2006) 314: 610

[12] Gruppo di lavoro Articolo 29, Parere 5/2014 sulle tecniche di anonimizzazione, WP 216 del 10 aprile 2014, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

L'Unione Europea come 'repubblica tecnologica' – di *Andrea Venanzoni*, Vicepresidente AssoCyber

Le grandi società tecnologiche stanno assumendo un ruolo sempre più vicino a quello di infrastrutture di governo. Dalla lezione americana emerge una domanda cruciale per Europa e Italia, che devono rafforzare sovranità digitale, coordinamento istituzionale e capacità di innovazione autonoma

La 'repubblica tecnologica' americana

Le grandi società tecnologiche sono attori geopolitici ormai strategicamente irrinunciabili per l'interesse nazionale stesso, come riconosce un recente documento del Parlamento Europeo[3].

Paradosso significativo che detto riconoscimento arrivi dal cuore di un sistema istituzionale ed economico privo di quei grandi soggetti, riconducibili ormai quasi esclusivamente agli USA o alla Cina.

Di recente la pubblicazione dell'AI Index Report di Stanford HAI, come ha ricordato Stefano da Empoli[4], oltre a questa dicotomia sempre più cristallizzata e consolidata, con una Europa in grande affanno, ha illustrato nella maniera più potente possibile la ritrovata centralità del ruolo pubblico.

Ed è proprio questo il cuore della lezione americana.

Infatti, quei soggetti societari statunitensi non sono più soltanto strumenti serventi, in chiave mercatoria o di *soft power,* le esigenze dei loro governi, ma si stanno trasformando in qualcosa di decisamente diverso. È il paradigma, portando alle estreme conseguenze la lezione di Max Weber, del 'capitalismo politico', per come investigato da Alessandro Aresu[5].

Di recente, interrogandosi sul valore in termini di diritto pubblico di questo processo, Luca R. Perfetti ha sottolineato la nascita di uno Stato parallelo, uno Stato interno allo Stato, perimetrato dalla logica privata del singolo attore societario[6].

Il caso più noto e forte, in questa prospettiva, è quello di Palantir Technologies, società di analisi dei dati fondata da Peter Thiel nel 2003 e diretta da Alex Karp. Trasformatasi nel corso degli anni in un autentico

[3] S. Clapp, *Defence and artificial intelligence*, in Servizio studi del Parlamento europeo, aprile 2025.

[4] S. da Empoli, *L'Europa è schiacciata tra USA e Cina: come uscire dallo scacco*, in Agenda Digitale, 17 aprile 2026.
[5] Così Aresu definisce il capitalismo politico, *'intreccio strumentale e reale tra obiettivi economici ed esigenze di sicurezza nazionale',* A. Aresu, *Capire il capitalismo politico*, in Aspenia, 20, 2020, pp. 119.
Si veda anche A. Aresu, *Le potenze del capitalismo politico – Stati Uniti e Cina*, Roma, La Nave di Teseo, 2020.
[6] L. R. Perfetti, *Lo Stato parallelo: Brevi considerazioni su P. Thiel, The straussian moment, trad. it., Macerata, Liberilibri, 2025*, in Persona e Amministrazione, 2, 2026, pp. 743 e ss.

gigante, è una società improntata, sin dalla sua fondazione, a una precisa visione del mondo ben esposta all'attenzione pubblica dal volume di Karp, '*La repubblica tecnologica*[7]'.

Uno dei volumi più importanti degli ultimi anni e che il Wall Street Journal ha descritto come un ambizioso trattato di filosofia politica[8].

È finito il tempo della Silicon Valley esclusivamente consumeristica, annota Karp, i cui prodotti e servizi esulavano, per principio, da qualunque idea di collaborazione organica con il governo americano[9]: ormai è tempo di creare un dispositivo organico che trascenda i limitati confini del partenariato pubblico-privato per sublimarsi in una forma di governo nuova, la 'repubblica tecnologica' appunto.

In fondo, la Valley stessa, storicamente, si situa lungo la dorsale dell'interesse militare ed è stata cesellata dai finanziamenti del Dipartimento della Difesa[10]: il fatto che col passare degli anni abbia cercato di scrollarsi di dosso quella origine ingombrante è purissima *fictio*.

La vicenda di Palantir e di come il governo USA se ne sta servendo presenta una serie di profili di grande interesse anche in prospettiva europea, soprattutto nella angolazione di analisi di quanto l'uso della IA nei cicli decisionali politico-amministrativi vada rimodellando la fisionomia della forma di governo stessa.

Palantir, è noto, pur nata come società specializzata in analisi massiva dei dati attraverso piattaforme s*oftware* basate su algoritmi e crittografia mutuati da PayPal, è ormai divenuta un sistema infrastrutturale che mediante l'integrazione organica delle sue piattaforme, come Gotham o Foundry, arriva a definire una '*Ontology*' attraverso cui passa la nuova decisione esecutiva.

In tutto questo c'è anche un aspetto di natura culturale: Palantir coltiva infatti la dimensione strategica del dislocamento di propri ingegneri presso la singola realtà che utilizza la data piattaforma, procedendo a una definizione di un modello di volta in volta unico.

L'ingegnere distaccato, il *forward deployed engineer*[11], concetto tipicamente militare, è ormai forma largamente utilizzata anche nel settore civile e amministrativo, contribuendo non solo a modellare la

[7] A. Karp, *La repubblica tecnologica*, Milano, Silvio Berlusconi Editore, 2025.
[8] J. Bew, *'The Technological Republic' Review: Power in a Silicon World*, in Wall Street Journal, 287 febbraio 2025.
[9] A. Karp, *La repubblica tecnologica*, cit., p. 21.

[10] M. O'Mara, *The Code. Silicon Valley and the Remaking of America*, New York City, Penguin, 2019, specie pp. 24 e ss.

[11] Sull'importanza di questa figura, S. Sankar, *Mobilize*, New York, Simon & Schuster, 2026.

piattaforma a seconda di contesto e utilizzatore finale ma anche irradiando una mentalità diversa nel circuito decisionale.

Il processo che ha posto Palantir al centro di governo dell'infrastruttura pubblica americana è stato accompagnato da un intricato reticolo normativo e contrattuale.

Con *executive order* del 25 marzo 2025, l'amministrazione Trump ha dato mandato alle sue strutture di centralizzare il flusso di dati al fine della piena razionalizzazione dell'utilizzo dei dati stessi.

Con contratto del maggio 2025, l'amministrazione federale ha stabilito che le strutture pubbliche utilizzino, in modo centralizzato, l'infrastruttura di Palantir per processare i dati unitariamente raccolti. Il superamento della ritenzione informativa è d'altronde riconosciuto quale specialità di Palantir[12] e si innerva in un lungo *iter* di definizione di una amministrazione algoritmica[13].

Nel Luglio 2025, il Piano federale per l'IA ha inteso superare la competenza legislativa degli Stati dell'Unione, centralizzando in capo al livello federale le politiche in tema di IA, riconosciute come sfida esistenziale.

Questa dinamica di accentramento funzionale, accompagnata da una precisa visione costituzionale del potere esecutivo, conosciuta come teoria unitaria del potere esecutivo[14], è coadiuvata proprio dall'intelligenza artificiale e scandisce le linee di evoluzione, e di emersione, di una sorta di governo *start-up*.

Si tratta di una visione tipicamente americana e che vanta, in quel Paese, antecedenti concettuali assai precisi[15] ma che presenta alcuni aspetti di interesse anche in una dimensione europea.

Sfrondata degli eccessi e ricondotta ad una latitudine valoriale europea può indicare una strada utile da seguire.

---

[12] L. Ulbricht, S. Egbert, *In Palantir we trust? Regulation of data analysis platforms in public security*, in Big Data & Society, 4 agosto 2024.

[13] AA.VV., *Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies*, in Administrative Conference of the United States, febbraio 2020.

[14] D. M. Driesen, *The Unitary Theory in Comparative Context*, in UC Law Journal, 1, 2020, particolarmente critico a pp. 29 e ss.

Molto critici anche J. Crouch, M. J. Rozell, M. A. Sollenberger, *The Unitary Executive Theory: a Danger to Constitutional Government*, Lawrence, Texas University Press, 2020, in particolare pp. 10 e ss.

C. R. Sunstein, A. Vermeule, *The Unitary Executive, Past, Present, Future*, in The Supreme Court Review, 2020, pp. 83 e ss., la vedono al contrario come scarsamente eversiva.

[15] R. Hofstaedter, *L'odio per gli intellettuali in America*, Roma, LUISS University Press, 2024, pp. 233 e ss.

L'Europa come 'repubblica tecnologica'

La lezione americana è preziosa perché non schiude un orizzonte esclusivamente limitato alla sfera di semplificazione, tanto di *stock* quanto di flusso, della produzione normativa e degli incombenti amministrativi.

Indipendenza tecnologica e sovranità digitale, anche alla luce dell'US Cloud Act, devono divenire imperativi esistenziali e non solo bandiere segnaletiche di *policy*, accompagnandosi fisiologicamente a un processo di integrazione digitale istituzionale anche per l'Europa politica.

L'integrazione digitale, particolarmente evidente lungo il crinale della *cybersecurity*[16], può essere definita quale sistema coordinato di politiche tra i vari livelli di decisione[17]: deve accompagnarsi a una *road-map* empiricamente sostenibile che dal livello europeo promani a quello nazionale, senza che quest'ultimo si trasformi in duplicazione segnaletica e, quindi, in zavorra frenante.

Perché, ed appare ovvio, indipendenza e sovranità si raggiungono e mantengono prioritariamente attraverso innovazione tecnologica propria, che consenta modellazione *by design* di servizi e prodotti conformi ai principi di diritto europeo.

Questo aspetto, sovente enunciato ma scarsamente praticato, diviene ancora più indefettibile alla luce di un dato pacifico: la divaricazione sempre più cristallina tra ciò che, per riprender il lessico di un recente, bel libro di Gianfranco Pacchioni[18], possiamo definire scienza 'chiara' e scienza 'oscura', intendendo questa ultima quale tutta quella sperimentazione e quella innovazione che avvengono lontano dallo scrutinio del pubblico dibattito.

Innovazione legata al campo della difesa ma ancor più spesso, specie in tema di IA, perché portata avanti nel ventre dei grandi soggetti del Tech.

La scienza 'oscura' pone un nuovo tema di consolidamento del potere, recentemente sottolineato da Oreste Pollicino commentando il *Project Glasswing* di Anthropic[19]; la sottrazione completa di un elemento tecnologico allo spazio pubblico, un fenomeno che conferisce al soggetto privato una rendita

[16] F. Serini, *Il sistema europeo di cooperazione informativa per il contrasto alle minacce informatiche. Verso una definizione di cybersicurezza integrata*?, in Medialaws, 3, 3023, p. 171
[17] Sia consentito il rinvio a A. Venanzoni, *L'ordine costituzionale della cybersecurity*, in Forum di Quaderni Costituzionali – Rassegna, 4, 2024, specie pp. 59 e ss.
[18] G. Pacchioni, *Scienza chiara, scienza oscura. Ricerca pura, ricerca militare, Big Tech*, Bologna, Il Mulino, 2025.
[19] O. Pollicino, *Il potere oggi è di chi regola l'accesso al mondo digitale*, in Milano Finanza, 15 aprile 2026.

di posizione *ex ante* immane e che appare, ontologicamente, sfuggire a qualunque ipotesi di regolazione possibile.

Uno scenario di questo genere evidenzia nella maniera più evidente quanto e come all'innovazione non possa che prioritariamente rispondersi con innovazione propria, un aspetto questo che d'altronde Mario Draghi aveva ben evidenziato nel suo *Rapporto sulla competitività*.

Come Palantir, anche nel caso dei *Large Language Models* siamo in una prospettiva non più puramente mercatoria o concorrenziale o piattaformizzata ma ontologica, infrastrutturale.

Se diamo per assodato che la digitalizzazione di una nazione si fonda su cinque pilastri essenziali, l'adozione delle tecnologie digitali, le infrastrutture abilitanti, l'*hardware* e il *software*, i servizi digitali, e i dati, ci rendiamo conto di come e quanto la dipendenza europea da *player* americani sia enorme.

Proprio questa necessità di accelerazione, unitamente al processo semplificatorio contenuto nel *Digital Omnibus* e al tentativo di ridurre la dipendenza in tema di *chip*, con il *Chips Act*, che punta a riportare la produzione di semiconduttori al 20% del totale mondiale entro la fine del decennio, riducendo la dipendenza da fornitori esterni, porta a nutrire perplessità sull'approccio di governance multi-partecipativa in tema di intelligenza artificiale.

Come noto, a sovraintendere alla normativa AI Act, in fase esecutiva e di vigilanza e di applicazione, e consultiva, abbiamo un sistema complesso. L'Ufficio europeo per l'IA e le autorità nazionali di vigilanza del mercato individuati sul versante attuativo della normativa *de qua*, e poi un ecosistema partecipativo di funzione consultiva: il comitato europeo per l'intelligenza artificiale, composto da rappresentanti degli Stati membri dell'UE, il gruppo di esperti scientifici, composto da esperti indipendenti nel settore dell'IA, il forum consultivo, che rappresenta una selezione diversificata di portatori di interessi, sia commerciali che non commerciali.

In considerazione del fatto che l'Unione Europea è un organismo sovra-nazionale non di matrice federale e che pertanto, fisiologicamente, la decisione politica è ripartita in modo pulviscolare tra livello euro-unitario e livello nazionale, risultandone fisiologicamente rallentata, questo sistema deve essere rivisto, sostituendo gli organi con consultazioni aperte.

Del pari è necessario che a livello europeo vi sia un organismo unico chiamato ad occuparsi di IA e che ad esso siano devolute tutte le competenze che, ad oggi, *ratione materiae*, pur incidenti nel campo IA, sono diffuse in una più estesa costellazione istituzionale: si pensi ad ENISA per quanto concerne il rapporto tra IA e *cybersecurity*.

Un altro punto di partenza irrinunciabile deve essere il ricorso massivo all'*open source*[20].

L'*open source* è ad oggi l'unico ambito che consenta, realisticamente, di mitigare la dipendenza strategica e digitale da *vendor* americani.

Non per caso, sia in ambito *cloud, edge* e IoT, l'Unione ha iniziato a porre forte enfasi su soluzioni *open source*, nonostante oscillazioni che, a *Cyber Resilience Act* vigente, hanno portato a parlare di una potenziale fine dell'*open source* stesso[21].

Gli indirizzi generali sembrano, però, fortunatamente mutare.

Si può citare, a detto proposito, la *Dichiarazione di sovranità digitale europea*, siglata nel novembre 2025 e che reca una chiara impostazione connessa all'*open source*.

Un atto politico, non normativo, ovviamente, ma che pure sembra tracciare una strada di favore riconosciuto per soluzioni aperte.

Da ultimo, anche l'elaborazione di un più recente *framework cybersecurity* appare più sensibile alla dimensione *open source*[22]. L'elaborazione di una tabella di marcia europea per l'*open source* è guidata da quattro fattori determinanti: sovranità digitale e tecnologica, sicurezza dei dati e conformità, innovazione e resilienza economica, nonché sostenibilità ambientale.

Ciascuno di questi fattori evidenzia la necessità di una strategia dedicata che sfrutti le tecnologie *open source* europee per governare le vulnerabilità, superare le lacune infrastrutturali, ottimizzare e razionalizzare la sicurezza e permettere alle organizzazioni europee di agire con piena indipendenza, sicurezza e in chiave sintonica con i valori europei all'interno di un quadro digitale il più affidabile possibile.

Definire una 'repubblica tecnologica europea' passa innanzitutto mediante la definizione di canoni di interoperabilità di servizi e prodotti digitali; una infrastruttura connettiva che parlando un unico linguaggio superi quello che è uno dei problemi maggiori dello spazio europeo, l'omogeneità.

Anche alcuni Stati membri stanno adottando sul versante interno soluzioni fortemente *open source*, per lenire la dipendenza da soggetti privati americani.

---

[20] M. Klare, U. Lechner, *Digital Sovereignty: Affordances in Open Source Projects*, in Software Business, 3, 2025, pp. 20 e ss.
[21] L. Colonna, *The end of open source? Regulating open source under the cyber resilience act and the new product liability directive*, in Computer Law & Security Review, 56, 2025
[22] J. Tridgell, *Open or closing doors? The influence of 'digital sovereignty' in the EU's Cybersecurity Strategy on cybersecurity of open-source software*, in Computer Law & Security Review, 56, 2025.

Ad esempio, in Germania sempre più amministrazioni locali stanno abbandonando *software* Microsoft, sostituiti con soluzioni *open source*, e al tempo stesso si stanno definendo soluzioni, normative e fiscali, di favore per le *start-up* tecnologiche al fine di sviluppare innovazione autoctona[23].

In Francia, si sta sviluppando una piattaforma, *Visio*, modellata in funzionalità *open source*, per sostituire le piattaforme di comunicazione americane. Pur avendo intrapreso da anni ormai un processo di digitalizzazione della sfera pubblica, meno ci si era interrogati sul quoziente di dipendenza dalle realtà extra-europee[24].

Un altro aspetto essenziale è potenziare e attrarre investimenti mirati in area europea in tema *open source*, anche mediante connessioni con realtà societarie e fondazioni europee per l'implementazione concreta di *policies* in tema.

Aspetto questo che costituisce parte saliente di una ampia e condivisibile riflessione di Luciano Floridi che vede nell'*open source* applicato a modelli fondazionali IA una via privilegiata per riattivare e rivitalizzare il complessivo progetto europeo[25].

...e l'Italia come 'repubblica tecnologica'

Uno dei principali problemi italiani è l'eccessiva frammentazione del ciclo decisionale in tema di innovazione digitale.

Ad una produzione normativa sovente episodica, quando non sloganistica, manca una simmetrica impalcatura istituzionale che vigili e attui concretamente gli indirizzi, procedendo quando necessario a una rettificazione empirica delle norme, sul campo.

Si tratta, per vero, di fenomeno risalente e che col passare del tempo ha determinato una stratificazione geologica di norme, con la risultante di disorientare molto spesso persino le stesse amministrazioni[26].

L'abbondanza di norme, riconducibili funzionalmente a istituzioni differenziate, con non banali necessità di coordinamento e di collaborazione, conduce a ciò che è stata definita 'poliarchia

[23] J. Bennink, *Germany's digital minister wants a European Palantir*, in Politico, 14 aprile 2026.
[24] K. Yeung, '*The New Public Analytics as an Emerging Paradigm in Public Sector Administration*' (2022) 27(2) Tilburg Law Review, specialmente p. 28.

[25] L. Floridi, *La differenza fondamentale*, Milano, Mondadori, 2025, pp. 221 e ss.
[26] S. Ungaro, *Non è il Foia promesso, "PA disorientate da eccesso di norme e assenza di guida",* in Agenda Digitale, 2 febbraio 2017.

regolatoria[27]'. Quanto poco questo quadro possa coniugarsi con un ecosistema fattivamente devoluto alla reale innovazione appare evidente.

In quanto alla natura sloganistica di alcune previsioni normative, basta un semplice esempio empirico: nonostante riuso e utilizzo di soluzioni *open source* siano previsti e definiti dal d. lgs. 7 marzo 2005, n. 82 (codice dell'amministrazione digitale)[28] come obbligatori, prima di procedere all'acquisto di s*oftware* proprietari, la prassi ci dice ben altro.

Nonostante l'enfasi, anche in questo caso, la vita interna delle amministrazioni e la natura delle loro decisioni legate all'approvvigionamento di dotazioni strumentali concrete segnalano un quadro molto diverso[29]. Quella che a norma vigente dovrebbe essere la deroga, l'eccezione, diviene in via di prassi lo standard. Ma si tratta di standard e di decisioni che hanno aumentato il livello quali-quantitativo di dipendenza tecnologica da soggetti privati extra-europei.

In questo specifico caso, non c'è alcun dubbio che si dovrebbe intervenire normativamente assegnando un arco temporale alle amministrazioni entro il quale, obbligatoriamente, passare dal *software* proprietario a quello aperto, eliminando in radice la clausola derogatoria.

Non c'è alcun dubbio, poi, che questa frastagliata e balcanizzata modalità di gestione sia figlia di una mancanza di strategia. Un altro aspetto saliente che finisce per rimandare un quadro assai nitido di una carenza di strategia è proprio quello della *governance* istituzionale dell'IA.

Se qualcosa l'esempio americano ci insegna, persino in uno Stato a vocazione fortemente federalista e in cui il X emendamento scolpisce un favore per le comunità territoriali particolarmente cogente, è proprio la necessità di eliminare i passaggi intermedi, i *layer* decisionali, e soprattutto le contraddizioni gestionali che inevitabilmente vengono a crearsi in una stordente polifonia decisionale.

L'Italia, è noto, si è dotata, prima tra tutti i Paesi europei, di una legge sull'intelligenza artificiale, la quale attua, ma non solo, la normativa contenuta nell'IA Act, pur in alcuni casi discostandosene.

Nonostante l'approccio istituzionale scelto sia stato descritto come governo-centrico, la risultante concreta continua, secondo una tradizione italiana schiettamente patologica, a vedere una galassia di soggetti istituzionali, direttamente o indirettamente, chiamati in causa.

Alla luce della legge 23 settembre 2025, n. 132, segnatamente articolo 20, il quadro che ne risulta è quello di un vertice politico, la Presidenza del Consiglio, e due organismi insigniti direttamente della

[27] S. Compagnucci, *Energia, salute e TLC: quando troppe regole frenano l'Italia*, in Agenda Digitale, 17 aprile 2026.
[28] In particolare, gli articoli 68 e 69 del d. lgs. n. 82/2005.
[29] I. Vignoli, *Open source: cresce l'uso nei governi, ma manca visione strategica*, in Agenda Digitale, 23 giugno 2025.

competenza di *governance*, quali autorità nazionali di IA: Agenzia Nazionale per la Cybersicurezza (ACN), da un lato, e Agenzia per l'Italia Digitale (AgID), dall'altro.

Ma, appare evidente, il quadro non si può esaurire qui.

Troppi profili esulerebbero dalla ipotetica *actio finium regundorum*, quando si decidesse di scandagliare i profili realmente sostanziali chiamati in causa dall'IA.

Ed è evidente come della partita debbano essere anche il GPDP, l'AGCOM e l'AGCM, ciascuna per il segmento di sua specifica competenza.

Il quadro perfetto, verrebbe da dire, per produrre una seriale sequenza di conflitti di competenze tra agenzie ed enti di governo[30], portando ad una potenziale inerzia.

Ma la complicazione dall'aura labirintica non si arresta qui. Perché l'intero ecosistema digitale, e quindi anche dell'IA, è segmentato in maniera vertiginosa. Competenze in tema ricadono in capo al Dipartimento per la trasformazione digitale o addirittura ai singoli Ministeri, ciascuno con una strategia funzionale localizzata nell'alveo delle proprie attribuzioni.

Come se non bastasse, questa postura diffusa porta a una frantumazione stellare con la istituzione di una serie di tavoli, comitati, commissioni, dalla natura consultiva ma la cui unica risultante è l'aumento esponenziale del rumore di fondo.

Arrivati a questo punto appare indispensabile adottare una strategia gordiana; creazione di un unico organismo di rango ministeriale che assorba ogni aspetto non ricadente nella fisionomia dei compiti di una autorità indipendente.

Tutte le funzioni residue di governo e di implementazione dell'IA devono essere centralizzate in un Ministero dell'IA, il quale, simmetricamente, porti alla immediata chiusura di qualunque struttura duplicata e alla cessazione di qualunque piano particolare, strategia ministeriale, dossier, libri verdi, bianchi.

Del pari, il Ministero dell'IA si farà latore di una reale strategia di formazione culturale, di istruzione e di disseminazione per arrivare all'unico vero obiettivo esistenziale; raccordare mondo dell'istruzione, quello accademico e mercato del lavoro per comprendere, empiricamente, anche quali saranno i settori più incisi e colpiti, e in che modo, dalla diffusione dell'IA.

[30] C. Morelli, *Pro e contro della legge sull'intelligenza artificiale, l'analisi*, in Wired, 24 settembre 2025.

Il raccordo deve portare al superamento di una visione puramente cartolare, libresca e museale della connessione tra IA e lavoro, elaborando una dimensione che consenta, anche in Italia, e in Europa, di dotarsi di *forward deployed engineer* che lavorino nelle amministrazioni e che monitorino l'attuazione empirica della normativa e delle politiche pubbliche, customizzino i modelli tecnologici e i servizi digitali.

Non ultimo, la definizione di una politica di incentivazione di finanziamenti e di snellimento procedurale e burocratico al fine di ingenerare un ecosistema idoneo allo sviluppo di *start-up* innovative.

Perché, appare chiaro, senza capacità di attrarre capitali continueremo ad esportare, e perdere, i nostri talenti.

Sistema di IA nell'AI Act: definizione giuridica e criteri applicativi - di ***Anna Cataleta***, Università degli Studi Suor Orsola Benincasa - Napoli

La definizione di sistema di IA nell'AI Act delimita l'ambito applicativo del Regolamento UE 2024/1689. Il testo ricostruisce l'evoluzione della nozione, il ruolo dell'OCSE e della Commissione europea, i sette elementi costitutivi e le categorie escluse dalle Linee Guida.

L'inquadramento giuridico dell'intelligenza artificiale (di seguito anche "**IA**") richiede di precisare che il Regolamento (UE) 2024/1689 del Parlamento europeo e del Consiglio che stabilisce regole armonizzate sull'intelligenza artificiale (di seguito anche "**AI Act**" o "**Regolamento**") non estende la propria disciplina a qualsiasi soluzione tecnologica riconducibile al concetto di IA, limitandosi piuttosto a regolare esclusivamente le soluzioni di IA che rientrano nel suo perimetro applicativo e ne soddisfano i requisiti definitori.

Come già emerso in altri ambiti normativi di natura interdisciplinare, l'adozione di un apparato terminologico coerente e condiviso si rivela imprescindibile per descrivere adeguatamente quali sono le soluzioni di IA, le tecniche impiegate e i requisiti giuridici e tecnici applicabili. Tale complessità è stata evidenziata in un recente contributo, ancora in fase di revisione (Sun et al. 2026). Tale vocabolario comune dovrebbe costituire un linguaggio condiviso tra gli esperti delle diverse discipline, facilitando la comunicazione e lo scambio di conoscenze al di là dei confini settoriali, indipendentemente dal contesto specifico in cui le soluzioni di IA vengono sviluppate o utilizzate. Come rilevato nel Considerando 12 dell'AI Act[31], l'assenza di un vocabolario affidabile e univocamente utilizzabile dalle diverse comunità professionali genera rischi per la certezza del diritto, con potenziali ricadute negative sulla ricerca scientifica e sullo sviluppo tecnologico (Rossi et al. 2025).

Ai fini di verificare l'applicabilità dell'AI Act ad una determinata soluzione di IA, è necessario, in primo luogo, comprendere se queste ultime corrispondano alle definizioni normative di "**sistema di intelligenza artificiale**" o di "**modello di intelligenza artificiale per finalità generali**" previste dall'AI Act.

[31] Considerando 12 "*La nozione di «sistema di IA» di cui al presente regolamento dovrebbe essere definita in maniera chiara e dovrebbe essere strettamente allineata al lavoro delle organizzazioni internazionali che si occupano di IA al fine di garantire la certezza del diritto, agevolare la convergenza internazionale e un'ampia accettazione, prevedendo nel contempo la flessibilità necessaria per agevolare i rapidi sviluppi tecnologici in questo ambito. […]*"

Il presente lavoro si concentra sulla prima delle due categorie definitorie: sistema di intelligenza artificiale. Il Regolamento si applica unicamente ai sistemi che soddisfano i requisiti stabiliti dall'articolo 3, punto 1, e rientrino nella definizione di "sistema di IA", ossia "*un sistema automatizzato progettato per funzionare con livelli di autonomia variabili e che può presentare adattabilità dopo la diffusione e che, per obiettivi espliciti o impliciti, deduce dall'input che riceve come generare output quali previsioni, contenuti, raccomandazioni o decisioni che possono influenzare ambienti fisici o virtuali*".

La corretta interpretazione della definizione di "sistema di IA" assume quindi carattere fondamentale per la delimitazione dell'ambito di applicazione della disciplina.

Di seguito sarà illustrato più nel dettaglio l'iter che ha portato alla definizione di sistema di IA presente nell'AI Act e che cosa, in concreto, occorre considerare per comprendere se una soluzione di IA rientra nell'ambito di applicazione del Regolamento.

L'iter che ha portato alla definizione attuale di sistema di IA presente nell'AI Act

Nella storia dell'intelligenza artificiale sono state proposte diverse definizioni che tendevano a fare riferimento allo sviluppo di sistemi capaci di imitare o riprodurre il pensiero e l'azione umani, di agire razionalmente e di interagire con l'ambiente circostante. L'IA mira, infatti, a sintetizzare o riprodurre i processi cognitivi umani, quali la percezione, la creatività, la comprensione, il linguaggio o l'apprendimento. Le numerose versioni della definizione di sistema di IA proposte negli anni, fino a quella inserita nell'AI Act, riflettono le difficoltà incontrate per formulare una definizione di sistema di intelligenza artificiale (Presno Linera e Meuwese 2025).

Una delle prime definizioni di "Intelligenza artificiale" elaborata a livello delle Istituzioni dell'UE si trova nella Comunicazione della Commissione europea "*L'intelligenza artificiale per l'Europa*" del 2018 (Mantelero, Resta e Riccio 2025); secondo la Commissione, l'IA *"indica sistemi che mostrano un comportamento intelligente analizzando il proprio ambiente e compiendo azioni, con un certo grado di autonomia, per raggiungere specifici obiettivi. I sistemi basati sull'IA possono consistere solo in software che agiscono nel mondo virtuale (per esempio assistenti vocali, software per l'analisi delle immagini, motori di ricerca, sistemi di riconoscimento vocale e facciale); oppure incorporare l'IA in dispositivi hardware (per esempio in robot avanzati, auto a guida autonoma, droni o applicazioni dell'Internet delle cose)"* (Commissione europea 2018).

Successivamente, nel 2019, anche il Gruppo di esperti nominato dalla Commissione europea (*High Level Expert Group* o *HLEG*) ha inserito nel glossario degli "*Orientamenti etici per un'IA affidabile*" una definizione di sistemi di IA, ossia "*Sistemi software (ed eventualmente hardware) progettati dall'uomo che, dato un obiettivo complesso, agiscono nella dimensione fisica o digitale percependo il proprio ambiente attraverso l'acquisizione di dati, interpretando i dati strutturati o non strutturati raccolti, ragionando sulla conoscenza o elaborando le informazioni derivate da questi dati e decidendo le migliori azioni da intraprendere per raggiungere l'obiettivo dato.* " (Commissione europea 2019).

Nel dibattito sulla definizione di "sistema di intelligenza artificiale", oltre alla Commissione europea e all' HLEG, ha svolto un ruolo di primo piano anche l'Organizzazione per la cooperazione e lo sviluppo economico (OCSE o OECD). Infatti, nel mese di novembre 2018, il gruppo di esperti sull'intelligenza artificiale dell'OCSE ha istituito un sottogruppo incaricato di elaborare una descrizione di un sistema di intelligenza artificiale che fosse comprensibile, tecnicamente accurata, neutrale rispetto alla tecnologia e applicabile sia nel breve sia nel lungo periodo. Inoltre, l'intento dichiarato è stato quello di creare una definizione sufficientemente ampia da ricomprendere molte delle definizioni di intelligenza artificiale comunemente utilizzate dalle comunità scientifiche, imprenditoriali e dei decisori pubblici. Inoltre, ha contribuito all'elaborazione della Raccomandazione del Consiglio dell'OCSE sull'intelligenza artificiale (OECD 2019). Dal lavoro emerge una possibile risposta al quesito su che cosa possa essere considerato un sistema di intelligenza artificiale. Secondo l'OCSE, infatti, un sistema di IA "*Sistemi basati su macchine che, per un determinato insieme di obiettivi definiti dall'uomo, sono in grado di formulare previsioni, raccomandazioni o decisioni, influenzando ambienti reali o virtuali, e che sono progettati per operare con diversi livelli di autonomia*".

Nel mese di novembre del 2023, l'OECD ha aggiornato la definizione come segue: "*Un sistema basato su macchine che, per obiettivi espliciti o impliciti, deduce, a partire dagli input ricevuti, come generare output quali previsioni, contenuti, raccomandazioni o decisioni che possono influenzare ambienti fisici o virtuali. I diversi sistemi di intelligenza artificiale si distinguono per i rispettivi livelli di autonomia e per la capacità di adattamento successiva alla loro messa in esercizio"* (OECD 2023).

Sia la prima versione della definizione che la seconda hanno influenzato le diverse fasi della redazione della definizione di sistema di IA all'interno dell'AI Act.

Tornando ora al piano europeo, dopo aver preso in considerazione i diversi punti di vista, nel 2020, la Commissione europea nel Libro bianco sull'Intelligenza artificiale (Commissione europea 2020) ha affermato che "*Qualunque nuovo strumento giuridico dovrà comprendere una definizione di IA* ***abbastanza flessibile da accogliere il progresso tecnico****, ma anche* ***sufficientemente precisa da garantire la necessaria certezza del diritto****".*

Fatte queste premesse e osservata la complessità di trovare una definizione comune condivisa, la Commissione europea, nella proposta di Regolamento sull'intelligenza artificiale (Commissione europea 2021) è approdata alla seguente proposta: un sistema di IA è "*un software sviluppato con una o più delle tecniche e degli approcci elencati nell'allegato I, che può, per una determinata serie di obiettivi definiti dall'uomo, generare output quali contenuti, previsioni, raccomandazioni o decisioni che influenzano gli ambienti con cui interagiscono*".
Leggendo il testo è evidente che tale definizione è stata poi ulteriormente modificata nel corso del trilogo (durante il quale sono stati presi in considerazione sia la nuova definizione proposta dall'OCSE nel 2023, sia gli sviluppi tecnologici) e si è approdati alla definizione che oggi è prevista dall'art. 3 dell'AI Act[32].

È da notare che, come riportato sopra, l'intenzione delle istituzioni dell'UE è stata quella di proporre una definizione con un ambito di applicazione sufficientemente ampio in modo da consentire un adattamento ai futuri sviluppi dell'intelligenza artificiale (Fernández-Llorca et al. 2025).
Tuttavia, questa ampiezza definitoria rischia di non favorire la certezza del diritto in quanto, in alcuni casi, potrebbe risultare complesso determinare i confini della stessa e potrebbe essere difficoltoso comprenderne a pieno i limiti senza degli esempi pratici su quali soluzioni di IA vi rientrino o meno (Trincado Castán 2024).
Per tentare di sciogliere questi dubbi è intervenuta la Commissione europea.

Le linee guida della Commissione europea

[32] Regolamento (UE) 2024/1689, art. 3, par. 1, «sistema di IA»: un sistema automatizzato progettato per funzionare con livelli di autonomia variabili e che può presentare adattabilità dopo la diffusione e che, per obiettivi espliciti o impliciti, deduce dall'input che riceve come generare output quali previsioni, contenuti, raccomandazioni o decisioni che possono influenzare ambienti fisici o virtuali.

Come visto nelle pagine precedenti, una domanda centrale per comprendere da un punto di vista pratico quando una soluzione che adotta l'IA rientri nell'ambito di applicazione dell'AI Act (e, dunque, inneschi la necessità di rispettare gli obblighi dalla normativa sull'IA) è: "quando un sistema software costituisce un "sistema di IA" ai sensi del Regolamento"?

La risposta a questa domanda è preliminare alle altre valutazioni richieste dalla normativa.

Occorre, tuttavia considerare che, come chiarisce la stessa Commissione, la stragrande maggioranza dei sistemi, pur potendo tecnicamente qualificarsi come sistemi di IA, non sarà soggetta ad alcun requisito normativo specifico previsto dall'AI Act, in quanto classificabili come sistemi a "rischio limitato" (Commissione europea 2025). Alla luce di quanto detto, identificare correttamente i confini della definizione è dunque un prerequisito indispensabile per ogni operatore economico.

In questo contesto, posta la delicatezza di comprendere appieno i limiti della definizione di "sistema di IA" anche da un punto di vista operativo, l'articolo 96, paragrafo 1, lettera f), dell'AI Act stesso ha conferito alla Commissione europea il mandato di elaborare Linee guida interpretative sull'applicazione della definizione. La Commissione ha assolto tale compito pubblicando, a febbraio 2025, le linee guida sull'applicazione della definizione di sistema di IA (Commissione europea 2025) (di seguito, "Linee Guida"), un documento non vincolante destinato ad assistere fornitori, deployer e altri soggetti interessati nel determinare se un sistema costituisca o meno un sistema di IA ai sensi del Regolamento.

Va precisato che le Linee Guida non sono vincolanti ma costituiscono un supporto interpretativo fornito dalla Commissione; infatti, solo la Corte di giustizia dell'Unione europea (CGUE) può fornire interpretazioni autorevoli del regolamento sull'IA"[33]. Tuttavia, proprio perché mirano ad assicurare un'applicazione uniforme del Regolamento su tutto il territorio dell'Unione, orientando le autorità di vigilanza del mercato, i giudici nazionali e gli operatori economici, le Linee Guida assumono un rilievo pratico di primo piano.

È importante evidenziare che le Linee Guida adottano un approccio analitico che rifiuta la meccanicità applicativa; infatti, come precisato al paragrafo 6 delle stesse, *"[l]a definizione di sistema di IA non dovrebbe essere applicata meccanicamente; ciascun sistema deve essere valutato sulla base delle sue caratteristiche specifiche*". Questo principio metodologico è fondamentale per i professionisti e le imprese in quanto non esistono scorciatoie definitorie né elenchi esaustivi di sistemi inclusi o esclusi. **Ogni valutazione richiede un'analisi caso per caso**.

[33] Ibid., par. 7.

I sette elementi della definizione di sistema di IA

Le Linee Guida identificano sette elementi costitutivi della definizione di Sistema di IA, che devono essere analizzati congiuntamente secondo una prospettiva basata sul ciclo di vita del sistema.

Primo elemento: sistema automatizzato

Il termine "automatizzato" si riferisce al fatto che il sistema opera mediante macchine, intese come l'insieme di componenti hardware (processori, memorie, unità di rete) e software (codice, programmi, sistemi operativi) che ne consentono il funzionamento. La definizione è intenzionalmente ampia e, come precisato al paragrafo 13 delle Linee Guida stesse, essa ricomprende anche i sistemi di calcolo quantistico, i sistemi biologici od organici dotati di capacità computazionali.

Dal punto di vista operativo, questo elemento non solleva particolari problemi interpretativi per le imprese; infatti, qualsiasi sistema software che elabora dati in modo automatizzato soddisfa questo requisito. Il rilievo pratico è piuttosto nell'interazione con gli altri elementi della definizione.

Secondo elemento: **autonomia**

Il requisito di autonomia determina la linea di demarcazione tra i sistemi di IA e i sistemi automatizzati tradizionali. Le Linee guida richiamano il considerando 12 dell'AI Act chiarisce che "*con "livelli di autonomia variabili" si intende che il sistema deve essere progettato per operare con un certo grado di autonomia di azione rispetto al coinvolgimento umano*".

Le Linee Guida precisano che sono esclusi dalla definizione i sistemi "*progettati per funzionare esclusivamente con il pieno coinvolgimento e intervento umano manuale*". Il punto discriminante è dunque la capacità del sistema di generare output senza che questo sia controllato manualmente o specificato in modo esplicito ed esatto da un essere umano. Esemplificativamente, un sistema che richiede input manuali ma è in grado di produrre autonomamente un output (come una raccomandazione o una previsione) soddisfa il requisito dell'autonomia.

Per le imprese, questo elemento ha implicazioni importanti nella progettazione dei flussi di lavoro: l'introduzione di checkpoint di supervisione umana sui singoli output del sistema non elimina necessariamente il carattere autonomo del sistema, se il sistema è strutturalmente capace di produrre output in modo indipendente. Il grado di autonomia in alcuni casi assume particolare rilevanza; ad esempio, "*nei settori ad alto rischio individuati nell'allegato I e nell'allegato III dell'AI Act, possono, a determinate condizioni, comportare rischi potenziali e considerazioni in materia di sorveglianza umana aggiuntivi*".

Terzo elemento: adattabilità dopo la diffusione

L'adattabilità si riferisce alla capacità di autoapprendimento del sistema, che consente di modificare il proprio comportamento durante l'uso, come ribadito anche nel considerando 12 dell'AI Act. Un sistema adattivo può produrre risultati diversi per gli stessi input dopo aver "appreso" dall'esperienza accumulata nel corso dell'utilizzo.

Tuttavia, le Linee Guida chiariscono espressamente che l'adattabilità è un elemento facoltativo; infatti, l'utilizzo del termine "può" nella definizione normativa indica che un sistema può costituire un sistema di IA anche in assenza di capacità di autoapprendimento post-deployment. Il chiarimento può avere rilevanza pratica nella misura in cui determina che sistemi con architetture fisse, non aggiornabili dopo il rilascio, potrebbero comunque rientrare nella definizione di sistema di IA nel caso in cui soddisfino gli altri requisiti individuati nella definizione e non rientrino tra i sistemi che la Commissione esclude espressamente dall'ambito di applicazione della definizione di sistema di IA.

Quarto elemento: obiettivi espliciti o impliciti

I sistemi di IA operano perseguendo uno o più obiettivi, che possono essere espliciti, quindi chiaramente dichiarati e codificati dallo sviluppatore, come l'ottimizzazione di una funzione di costo, o impliciti, ovvero desumibili dal comportamento del sistema o derivanti dai dati di addestramento o dall'interazione del sistema di IA con il proprio ambiente.

Le Linee Guida, richiamando anche in questo caso il considerando 12 dell'AI Act, distinguono opportunamente tra gli obiettivi interni del sistema e la sua finalità prevista. Gli obiettivi riguardano i compiti che il sistema deve svolgere, hanno una rilevanza interna al sistema e fanno riferimento agli obiettivi dei compiti da svolgere e ai loro risultati ("*ad esempio, un sistema di assistente di IA virtuale aziendale può avere come obiettivi di rispondere alle domande degli utenti su una serie di documenti con un'elevata accuratezza e un basso tasso di errori*"). La finalità prevista ha invece valenza esterna e si riferisce al contesto applicativo in cui il sistema è progettato per operare e "*all'"uso di un sistema di IA previsto dal fornitore*" ("*ad esempio, nel caso di un sistema di assistente di IA virtuale aziendale, la finalità prevista potrebbe essere quella di assistere un determinato dipartimento di un'impresa nello svolgimento di determinati compiti*").

Questa distinzione è molto importante poiché la finalità prevista può influenzare la classificazione del livello di rischio del sistema e la determinazione degli obblighi di conformità.

Quinto elemento: capacità inferenziale

Il considerando 12 dell'AI Act, cui fanno riferimento anche le Linee Guida, identifica la capacità inferenziale come la caratteristica fondamentale che distingue i sistemi di IA dai sistemi software tradizionali. In particolare, questa capacità "*si riferisce al processo di ottenimento degli output, quali previsioni, contenuti, raccomandazioni o decisioni, che possono influenzare gli ambienti fisici e virtuali e alla capacità dei sistemi di IA di ricavare modelli o algoritmi, o entrambi, da input o dati*".

Secondo la Commissione, la formulazione normativa "deduce come" è deliberatamente ampia ed abbraccia tanto la fase di utilizzo del sistema (in cui questo genera output dagli input) quanto la fase di costruzione (in cui si sviluppano modelli o algoritmi attraverso tecniche di IA). È interessante notare altresì che le Linee Guida precisano che la nozione di inferenza è compatibile con la definizione ISO/IEC 22989, che include nei presupposti del ragionamento inferenziale non solo fatti e regole, ma anche modelli, caratteristiche e dati grezzi.

La capacità inferenziale, pertanto, costituisce un elemento fondamentale nel determinare la linea di confine tra i sistemi che rientrano nell'ambito di applicazione dell'AI Act e quelli che ne rimangono esclusi.

Sesto elemento: tipologie di output (che possono influenzare gli ambienti fisici o virtuali)

Il sesto elemento cataloga le tipologie di output che un sistema di IA può generare: previsioni, contenuti, raccomandazioni e decisioni. Le diverse tipologie di output evidenziano un diverso grado di coinvolgimento umano.

Le **previsioni**, che "*costituiscono uno degli output più comuni prodotti dal sistema di IA e richiedono il minore coinvolgimento umano*", sono stime di valori sconosciuti a partire da input noti. I sistemi di IA si distinguono dai sistemi predittivi tradizionali per la capacità di rilevare pattern complessi in ambienti dinamici; ad esempio, si pensi ai sistemi di guida autonoma o ai sistemi di previsione del consumo energetico che analizzano dati eterogenei (contatori intelligenti, previsioni meteorologiche, comportamenti dei consumatori).

I **contenuti** si riferiscono alla generazione di nuovo materiale (testi, immagini, video, audio) da parte del sistema. Questa categoria assume particolare rilievo con la diffusione dei sistemi di IA generativa, ai quali l'AI Act dedica specifiche disposizioni in merito ai requisiti di trasparenza.

Le **raccomandazioni** sono "*suggerimenti per azioni, prodotti o servizi rivolti agli utenti sulla base delle loro preferenze, dei loro comportamenti o di altri input di dati*".

Anche in questo caso, come per le previsioni, le raccomandazioni possono essere generate sia da sistemi di IA sia da sistemi che non usano l'IA. Anche in questo caso le differenze tra la qualità delle raccomandazioni varia molto; infatti "*sistemi di raccomandazione basati sull'IA, ad esempio, possono far leva sui dati su larga scala, adattarsi al comportamento degli utenti in tempo reale, fornire raccomandazioni altamente personalizzate ed espandersi in modo efficiente con l'aumento del set di dati, funzionalità raramente possedute dai sistemi non basati sull'IA*".

Le Linee Guida operano poi una distinzione importante: quando le raccomandazioni sono automaticamente applicate senza valutazione umana, esse si trasformano in **decisioni**.

Le decisioni, secondo quanto affermato nelle Linee Guida "*si riferiscono a conclusioni o scelte effettuate da un sistema. Un sistema di IA che produce una decisione automatizza processi tradizionalmente gestiti dalla valutazione umana*". In particolare, un determinato risultato è prodotto nell'ambiente senza alcun intervento umano attraverso un processo completamente automatizzato.

Tale distinzione è rilevante per determinare se si applichino obblighi aggiuntivi, ad esempio in materia di trasparenza verso gli individui soggetti a decisioni automatizzate.

Settimo elemento: interazione con l'ambiente

L'ultimo elemento evidenzia che i sistemi di IA non sono sistemi passivi, ma hanno un impatto attivo sull'ambiente in cui operano. Il riferimento ad "ambienti fisici o virtuali", richiamato nel punto precedente, abbraccia sia le interazioni con oggetti tangibili (si pensi a un braccio robotico controllato da un sistema di IA) sia quelle con spazi digitali, flussi di dati ed ecosistemi software.

Sistemi che non rientrano nell'ambito di applicazione della definizione di sistema di IA

Dopo aver analizzato gli elementi costitutivi della definizione di sistema di IA, le Linee Guida si occupano altresì di delimitarne negativamente il perimetro, individuando le categorie di sistemi che, pur presentando alcune caratteristiche superficialmente analoghe, non possono essere ricondotte alla nozione di sistema di IA ai sensi dell'AI Act.

La Commissione muove le proprie considerazioni dall'assunto del considerando 12 dell'AI Act secondo cui "*la definizione di sistema di IA dovrebbe distinguere i sistemi di IA dai tradizionali sistemi software o dagli approcci di programmazione più semplici, e non dovrebbe riguardare i sistemi basati sulle regole definite unicamente da persone fisiche per eseguire operazioni in modo automatico*". A partire da questo criterio orientativo, le Linee Guida individuano quattro categorie di sistemi esclusi. Ovviamente,

dal punto di vista del rispetto della normativa, è opportuno svolgere una valutazione caso per caso sulla singola soluzione che si intende sviluppare o adottare.

Sistemi per migliorare l'ottimizzazione matematica: secondo la Commissione, rientrano in questa categoria i sistemi utilizzati per migliorare l'ottimizzazione matematica o per accelerare e approssimare metodi di ottimizzazione tradizionali e consolidati, quali i metodi di regressione lineare o logistica. Tali sistemi, pur possedendo una capacità inferenziale, non trascendono l'elaborazione di base dei dati. Tra gli esempi presenti nelle Linee Guida si trovano i modelli basati sull'apprendimento automatico che approssimano funzioni o parametri nei problemi di ottimizzazione mantenendo nel contempo le prestazioni.

Un altro indicatore utile in sede interpretativa è la consolidata diffusione di tali sistemi nel tempo; pertanto, il fatto che un sistema sia stato impiegato stabilmente nel settore per molti anni può costituire un indizio del fatto che esso non rientri nella definizione di sistema di IA. Le Linee Guida richiamano, a titolo esemplificativo, "*sistemi basati sulla fisica possono utilizzare tecniche di apprendimento automatico per migliorare le prestazioni computazionali, accelerando le simulazioni tradizionali basate sulla fisica o stimando parametri, che vengono poi immessi nei modelli fisici consolidati. Tali sistemi non rientrerebbero nell'ambito di applicazione della definizione di sistema di IA*".

Rispetto a questa esclusione, alcuni autori affermano che la stessa si concentra sull'applicazione pratica del sistema algoritmico anziché sulle proprietà del sistema stesso e ciò non sarebbe coerente con la definizione contenuta nell'AI Act; infatti, nella definizione di sistema di IA dell'AI Act, l'applicazione pratica non è un elemento che determina se una soluzione sia o meno un sistema di IA (Algorithm Audit 2025).**Elaborazione di base dei dati:** la seconda categoria riguarda i sistemi di elaborazione di base dei dati. Questi ultimi seguono istruzioni o operazioni predefinite ed esplicite, e sono sviluppati per eseguire compiti sulla base di input manuali o norme, senza alcuna forma di apprendimento, ragionamento o modellizzazione in qualsiasi fase del ciclo di vita; infatti "*operano sulla base di regole fisse programmate dagli esseri umani, senza utilizzare tecniche di IA, come l'apprendimento automatico o l'inferenza basata sulla logica, per generare output*".

Le Linee Guida includono in questa categoria i sistemi di gestione delle banche dati utilizzati per selezionare o filtrare dati sulla base di criteri specifici, i fogli elettronici privi di funzionalità basate sull'IA, nonché i software che calcolano medie statistiche a partire da indagini. Analogamente, i sistemi destinati esclusivamente all'analisi descrittiva, al test di ipotesi e alla visualizzazione dei dati non rientrano nella definizione, secondo la Commissione.

Sistemi basati sull'euristica classica: la terza categoria riguarda i sistemi basati sull'euristica classica. La Commissione la definisce come "*una tecnica di risoluzione dei problemi incentrata su metodi basati sull'esperienza per trovare soluzioni approssimative in modo efficiente*". A differenza dei moderni sistemi di apprendimento automatico, i sistemi di euristica classica applicano algoritmi e regole predefiniti per ricavare soluzioni, senza adeguare i propri modelli sulla base delle relazioni input-output osservate.

L'esempio paradigmatico richiamato dalle Linee Guida è il programma di scacchi che utilizza un algoritmo minimax con funzioni di valutazione euristica che, pur essendo in grado di esaminare posizioni complesse sulla scacchiera, opera senza apprendimento preliminare dai dati.

Sistemi di previsione semplici: la quarta ed ultima categoria esclude i sistemi automatizzati le cui prestazioni possono essere conseguite mediante una regola di apprendimento statistico di base. Secondo la Commissione, tali sistemi, "*sebbene tecnicamente classificabili come basati su approcci di apprendimento automatico, non rientrano nella definizione di sistema di IA a causa delle loro prestazioni*", che possono essere limitate rispetto ai sistemi di IA più avanzati.

Le Linee Guida richiamano, a titolo di esempio, i sistemi che prevedono un indice azionario utilizzando sempre il prezzo medio storico, ovvero quelli che stimano la temperatura del giorno successivo sulla base della media della settimana precedente.

Rispetto a questo punto, alcuni autori hanno considerato che nella definizione di sistema di IA di cui all'articolo 3 dell'AI Act e nel considerando 12, la complessità di un sistema non è menzionata come fattore determinante per la qualificazione come sistema di IA; pertanto, questa eccezione può creare confusione in quanto non è chiaro quando una regola di apprendimento statistico possa essere considerata "di base" (Algorithm Audit 2025).

Alla luce di quanto visto sopra, le linee guida costituiscono comunque uno strumento importante per le organizzazioni, seppur con dei limiti e con le criticità evidenziate.**Conclusioni**

L'analisi svolta consente di trarre alcune considerazioni conclusive sulla definizione di sistema di IA nell'AI Act e sulle sue implicazioni pratiche.

La definizione di sistema di IA di cui all'articolo 3, punto 1, dell'AI Act rappresenta il risultato di un lungo iter evolutivo, che ha attraversato un complesso dibattito istituzionale, tecnico e accademico. Come emerso dalla ricostruzione dell'iter definitorio, le versioni succedutesi nel tempo riflettono la difficoltà strutturale di fissare in termini giuridici un fenomeno tecnologico in rapida evoluzione. La scelta di una definizione funzionale rappresenta un compromesso deliberato tra l'esigenza di certezza del diritto e la

necessità di non cristallizzare la normativa attorno a tecnologie che potrebbero risultare già superate al momento della sua applicazione.

Questa tensione non è stata interamente risolta; infatti, l'ampiezza definitoria rischia di generare un'incertezza applicativa che le stesse Linee Guida della Commissione non sembrano in grado di dissipare completamente.

In ogni caso, allo stato attuale, è necessario svolgere una valutazione caso per caso delle proprie soluzioni partendo dai criteri indicati dalla Commissione nelle Linee Guida. Occorre altresì tenere in considerazione che le indicazioni della Commissione non sono vincolanti e possono costituire una guida autorevole e per avere maggiori certezze interpretative occorrerà attendere i provvedimenti delle autorità competenti e le sentenze della Corte di giustizia dell'Unione europea.

Bibliografia

Algorithm Audit. *Implementation of the AI Act: Definition of an AI System*. White paper. febbraio 2025. https://algorithmaudit.eu.

Commissione europea. *L'intelligenza artificiale per l'Europa*. COM(2018) 237 final. Bruxelles: Commissione europea, 25 aprile 2018.

Commissione europea. *Piano coordinato sull'intelligenza artificiale*. COM(2018) 795 final. Bruxelles: Commissione europea, 7 dicembre 2018.

Commissione europea. *Libro bianco sull'intelligenza artificiale – Un approccio europeo all'eccellenza e alla fiducia*. COM(2020) 65 final. Bruxelles: Commissione europea, 19 febbraio 2020.

Commissione europea. *Proposta di regolamento del Parlamento europeo e del Consiglio che stabilisce regole armonizzate sull'intelligenza artificiale (legge sull'intelligenza artificiale) e modifica alcuni atti legislativi dell'Unione*. COM/2021/206 final. Bruxelles: Commissione europea, 21 aprile 2021.

Commissione europea. *Orientamenti sulla definizione di sistema di intelligenza artificiale stabilita dal Regolamento (UE) 2024/1689*. Bruxelles: Commissione europea, febbraio 2025. https://ec.europa.eu/newsroom/dae/redirection/document/112455.

Commissione europea, Gruppo indipendente di esperti sull'intelligenza artificiale. *Una definizione di IA: principali capacità e discipline scientifiche*. Bruxelles: Commissione europea, 8 aprile 2019.

Commissione europea, Gruppo indipendente di esperti sull'intelligenza artificiale. *Orientamenti etici per un'IA affidabile*. Bruxelles: Commissione europea, aprile 2019.

Fernández-Llorca, David, Emilia Gómez, Ignacio Sánchez e Gabriele Mazzini. "An Interdisciplinary Account of the Terminological Choices by EU Policymakers Ahead of the Final Agreement on the

AI Act: AI System, General Purpose AI System, Foundation Model, and Generative AI." *Artificial Intelligence and Law* 33, n. 4 (2025): 875–888. https://doi.org/10.1007/s10506-024-09412-y.

Mantelero, Alessandro, Giorgio Resta e Giovanni Maria Riccio, a cura di. *Intelligenza artificiale. Commentario. Regolamento (UE) 2024/1689*. Milano: Ipsoa, 2025.

OECD. *Recommendation of the Council on Artificial Intelligence*. OECD/LEGAL/0449. Paris: OECD, 22 maggio 2019. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

OECD. "Updates to the OECD's Definition of an AI System Explained." *OECD.AI*, 29 novembre 2023. https://oecd.ai/en/wonk/ai-system-definition-update.

Parlamento europeo e Consiglio dell'Unione europea. Regolamento (UE) 2024/1689 del 13 giugno 2024 che stabilisce regole armonizzate sull'intelligenza artificiale e modifica i regolamenti (CE) n. 300/2008, (UE) n. 167/2013, (UE) n. 168/2013, (UE) 2018/858, (UE) 2018/1139 e (UE) 2019/2144 e le direttive 2014/90/UE, (UE) 2016/797 e (UE) 2020/1828 (regolamento sull'intelligenza artificiale). *Gazzetta Ufficiale dell'Unione europea*, L 1689, 12 luglio 2024.

Presno Linera, Miguel Ángel e Anne Meuwese. "Regulating AI from Europe: A Joint Analysis of the AI Act and the Framework Convention on AI." *Theory and Practice of Legislation* 13, n. 3 (2025): 292–311. https://doi.org/10.1080/20508840.2025.2492524.

Rossi, Arianna, Francesca Gennari, Ilaria Fagioli, Alessandro Mazzarini, Fabrizio Moncelli, Denise Amram, Simona Crea e Andrea Parziale. "The AI System Definition under the AI Act, a New Nomen Rosae?" *CEUR Workshop Proceedings* 4074 (2025). https://ceur-ws.org/Vol-4074/short6-16.pdf.

Sun, Yuanyuan, Timothy Parker, Lara Gierschmann, Sana Shams, Teo Canmetin, Mathieu Duteil, Rokas Gipiškis e Ze Shen Chin. "Defining AI Models and AI Systems: A Framework to Resolve the Boundary Problem." *arXiv*:2603.10023 [cs.CY], 22 febbraio 2026. https://arxiv.org/abs/2603.10023.

Trincado Castán, Carlos. "The Legal Concept of Artificial Intelligence: The Debate Surrounding the Definition of AI System in the AI Act." *BioLaw Journal – Rivista di BioDiritto*, n. 1 (2024). https://doi.org/10.15168/2284-4503-3000

Dispositivi elettromedicali e tutela della persona – di ***Lorenzo Maria Lucarelli Tonini***, Dottore di Ricerca in Discipline Giuridiche, presso l'Università degli Studi "Roma Tre"

La sanità digitale sta ridefinendo il rapporto tra tutela della salute, innovazione tecnologica e protezione della persona. In questo quadro, i dispositivi elettromedicali mostrano come cybersicurezza, dati personali e sicurezza clinica siano ormai profili strettamente intrecciati

La tutela della salute nella sanità digitale

Il diritto alla salute costituisce un diritto che ha direttamente subito nel tempo un forte sviluppo, dovuto al mutare di molteplici elementi, così comportando un continuo adeguamento del proprio contenuto nel rispetto di quanto delineato dall'art. 32 della Costituzione italiana[34].

In tale contesto, ad oggi rileva sempre di più l'uso delle nuove tecnologie[35], utilizzate per trattare informazioni personali dei pazienti sia al fine di garantire loro l'accesso ai servizi sanitari, quindi per l'erogazione delle cure, sia per finalità di ricerca o di governo del sistema sanitario stesso.

[34] Ampiamente sul tema R. BALDUZZI, *Salute* (diritto alla), in S. CASSESE (diretto da), *Dizionario di diritto pubblico*, Giuffré, Milano, 2006; M. CARTABIA, *La giurisprudenza costituzionale relativa all'art. 32, secondo comma, della Costituzione italiana*, in *Quad. cost*., 2012, 455 ss.; F. MINNI, A. MORRONE, *Il diritto alla salute nella giurisprudenza della Corte Costituzionale italiana*, in *Rivista AIC*, 2/2013, 1 ss.; D. MORANA, *La salute come diritto costituzionale. Lezioni*, Giappichelli, Torino, 2015, 76 ss.; S. BARBARESCHI, *Tecniche di individuazione del nucleo essenziale del diritto alla salute nella giurisprudenza costituzionale*, in C. COLAPIETRO, M. ATRIPALDI, G. FARES E A. IANNUZZI (a cura di), *I modelli di welfare sanitario tra qualità e sostenibilità. Esperienze a confronto*, Esi, Napoli, 2018, 335 ss.

[35] C. IRTI, *Riflessioni sull'impatto delle nuove tecnologie nell'ambito sanitario, tra regole tecniche e regole giuridiche*, in *EJPLT*, 1/2023, 34 ss.; C. DI COSTANZO, *L'impiego delle nuove tecnologie nel settore della salute: problematiche e prospettive di diritto costituzionale*, in *Consulta Online*, 1/2023, 214 ss.; A. PIETROLETTI, A. NICOTRA, *Tutela della salute, sistemi digitali e privacy*, in *Inf. e dir*., 1/2022, 283 ss.; A. THIENE, S. CORSO (a cura di), *La protezione dei dati sanitari. Privacy e innovazione tecnologica tra salute pubblica e diritto alla riservatezza*, Jovene, 2023; F. MASCHIO, *Innovazione digitale dei sistemi sanitari, eHealth e questioni di data governance, nell'agenda digitale sanitaria nazionale ed europea*, in *Comparazione e diritto civile*, 2/2021, 559 ss.; R. MICCÙ, *Questioni attuali intorno alla digitalizzazione dei servizi sanitari nella prospettiva multilivello*, in *federalismi.it*, 5/2021; F. COVINO, *Uso della tecnologia e protezione dei dati personali sulla salute tra pandemia e normalità*, in *federalismi.it*, 5/2021, 42 ss.; F. APERIO BELLA, *L'accesso alle tecnologie innovative nel settore salute tra universalità e limiti organizzativi (con una postilla sull'emergenza sanitaria)*, in *Persona e Amministrazione*, 1/ 2020, 222 ss.; C. FARALLI, R. BRIGHI, M. MARTONI (a cura di), *Strumenti, diritti, regole e nuove relazioni di cura: il paziente europeo protagonista nell'eHealth*, Giappichelli, Torino, 2015; D.A. LIMONE (a cura di), *La sanità digitale*, in *Rivista elettronica di diritto, economia, management*, 3/ 2014.

Ad essere direttamente coinvolti sono sia il settore pubblico (il Servizio Sanitario Nazionale – SSN), che quello privato[36], spesso fornitore di servizi - tecnologici - necessari per garantire l'esercizio del servizio pubblico stesso.
Anche il legislatore ha ormai avuto modo di riconoscere formalmente l'importanza che le tecnologie e i dati hanno assunto nel settore sanitario, a fronte di quel processo di «giuridificazione» della persona in corso[37] a più livelli.

La Commissione europea, ad esempio, definisce con i termini «sanità e assistenza digitali» gli «strumenti e servizi che sfruttano le tecnologie dell'informazione e della comunicazione per migliorare la prevenzione, la diagnosi e il trattamento delle patologie, il monitoraggio e la gestione della salute e degli aspetti dello stile di vita che influiscono sulla salute». Rispetto alla molteplicità di tali tecnologie, centrali sono i c.d. "medical devices", ovvero i dispositivi elettromedicali attraverso i quali, in virtù delle tecnologie implementate e dei dati che vengono trattati, è ormai possibile raggiungere importanti risultati diagnostici, interventistici e di monitoraggio.

Tuttavia, a fronte di uno sviluppo tecnologico assai veloce, l'avvento della "sanità digitale" è piuttosto recente, in quanto è a partire dall'emanazione del primo "Piano d'Azione per la Sanità Elettronica"[38], adottato dalla Commissione europea nel 2004, che è stato riconosciuto rilievo giuridico a tale fenomeno, ove «il ricorso agli strumenti della sanità elettronica [...], consente di ottenere in tempo utile dati completi ed integrati». Successivamente, con una Comunicazione del 2012[39], la Commissione ha fissato un termine temporale, precisamente nel 2020, per raggiungere una maggiore interoperabilità dei servizi di sanità elettronica e sostenere la ricerca, lo sviluppo e l'innovazione, incentivandone la diffusione.
Ancora, con una Comunicazione del 2018[40], è stata affermata la necessità per i cittadini di accedere in maniera sicura ai propri dati, in ogni momento e ovunque si trovino all'interno dell'Ue, potendoli condividere attraverso strutture comuni e formati interoperabili, anche al fine di garantirne una migliore utilizzazione nello sviluppo di una medicina sempre più personalizzata e di precisione[41].

---

[36] C. COLAPIETRO, M. ATRIPALDI, G. FARES E A. IANNUZZI (a cura di), *I modelli di welfare sanitario tra qualità e sostenibilità. Esperienze a confronto*, Esi, Napoli, 2018.
[37] S. RODOTÀ, *Il corpo "giuridificato"*, in S. RODOTÀ, P. ZATTI, *Trattato di biodiritto*, Giuffrè, Milano, 2011, 56 ss. Ne evidenzia l'approccio G. AZZARITI, *Stefano Rodotà e l'uso costituzionalmente orientato delle nuove tecnologie*, in *Costituzionalismo.it*, 2, 2022.
[38] COM(2004)356 del 30 aprile 2004.
[39] COM/2012/0736 del 6 dicembre 2012.
[40] COM(2018)233 del 24 aprile 2018.
[41] Sia consentito rinviare in materia a L.M. LUCARELLI TONINI, *Il diritto di accesso alle cure: nuovi orizzonti nella personalizzazione del trattamento sanitario*, in *Teoria e prassi del diritto*, 2/2023.

Con una Raccomandazione del 2019[42] la Commissione ha poi sottolineato che «l'obiettivo principale della rete di assistenza sanitaria online è sostenere lo sviluppo di sistemi, servizi e applicazioni interoperabili di eHealth sostenibili» agevolando lo scambio di informazioni e la cooperazione tra Stati membri.

In siffatto contesto, si pone il più recente programma *EU4Health 2021-2027 – Una visione per un'Unione europea più sana*[43] che, adottato sulla base dell'emergenza pandemica, ha apportato un ingente sostegno finanziario nel settore sanitario, con un'attenzione specifica ai progetti di digitalizzazione.

Ulteriore efficacia propulsiva è poi derivata dalla Comunicazione "*L'Unione europea della salute*"[44] che, nel focalizzare l'attenzione sull'importanza rivestita dal processo di digitalizzazione della sanità, ne individua la pietra miliare nell'infrastruttura dello "*Spazio europeo dei dati sanitari*" così come previsto nel Reg. UE 2025/327[45].

Infine, l'Unione ha avuto modo di adottare un recentissimo Piano d'azione[46] con il quale è stata richiamata l'attenzione alla necessità di rafforzare la cybersicurezza delle strutture sanitarie e dei prestatori di servizi coinvolti.

Quanto alla normativa nazionale, è solo di recente, sulla spinta delle esigenze insorte con la pandemia da Covid-19, nonché a seguito degli stimoli provenienti a livello sovranazionale e realizzati mediante il "Piano Nazionale di Ripresa e Resilienza" (PNRR - Missione6 "Salute") che il processo di digitalizzazione sanitaria è stato oggetto di effettivi interventi di razionalizzazione, sistematizzazione e impulso.

Invero, già il "Codice dell'Amministrazione Digitale" (CAD)[47], all'art. 2, aveva posto generali obiettivi di digitalizzazione riguardanti, trasversalmente, tutte le amministrazioni pubbliche. Interventi concreti di regolazione sistematica della sanità digitale, però, sono derivati dal "Piano triennale per l'informatica

[42] Raccomandazione (Ue) n. 2019/243/UE del 6 febbraio 2019, Considerando n. 15.
[43] Il programma è stato istituito dal Regolamento (Ue) 2021/522 del 24 marzo 2021.
[44] COM (2024) 206 del 22 maggio 2024.
[45] Reg. Ue 2025/327 dell'11 febbraio 2025.
[46] COM(2025) del 15 gennaio 2025.
[47] Art. 2 del decreto legislativo 7 marzo 2005, n. 82.
Inoltre, l'art. 14 del CAD, dando seguito all bipartizione di competenze tra Stato e Regioni sulla scorta dell'articolo 117 Cost., prevede che la disciplina relativa al coordinamento informatico dei dati dell'amministrazione statale, regionale e locale, spetti allo Stato.

nella pubblica amministrazione", che, come previsto dalla legge 28 dicembre 2015, n. 208, art. 1, co. 513, è il documento predisposto dall'Agenzia per l'Italia digitale (Agid) che contiene, «per ciascuna amministrazione o categoria di amministrazioni, l'elenco dei beni e servizi informatici e di connettività e dei relativi costi [...] da sostenere per innovazione».

È però con il programma stabilito nella "Strategia per la crescita digitale 2014-2020"[48] che si è inteso rafforzare nello specifico il processo di digitalizzazione della sanità, attraverso l'«integrazione» delle frammentate e differenti iniziative esistenti e il raggiungimento di un «contesto organico», mediante un piano di azioni pianificate. In tale ambito, rivestono, inoltre, un ruolo fondamentale di specificazione il "Patto per la Sanità Digitale" del 24 giugno 2016[49], al quale la Strategia espressamente rimanda quale strumento di accordo tra Governo, Regioni e Province autonome, nonché il "Piano Nazionale della Sanità Digitale 2022-2027".

I dispositivi elettromedicali: cybersicurezza e protezione dei dati personali

In tale contesto, il caso dei dispositivi elettromedicali rappresenta una fattispecie centrale non solo per delineare l'impatto che l'innovazione tecnologica ha sulla tutela della salute, ma anche su altre e nuove situazioni giuridiche quali la protezione dei dati personali e la cybersicurezza.

Proprio prendendo le mosse dalla situazione giuridica la cui normazione è più recente, ossia la cybersicurezza, occorre anzitutto rappresentare come questa sia ben cristallizzata, in via di principio, nella "Dichiarazione europea sui diritti e i principi digitali per il decennio digitale"[50] che, seppure animata da un mero intento programmatico, esprime l'intenzione di fornire al concetto di sicurezza dell'ambiente digitale, quindi alla cybersicurezza, autonoma dignità e riconoscimento rispetto agli altri diritti fondamentali, affermando, al punto n. 16, che «ogni persona dovrebbe avere accesso a tecnologie, prodotti e servizi digitali che siano sicuri e protetti e tutelino la vita privata».

---

[48] La "Strategia" è stata approvata dal Consiglio dei Ministri il 3 marzo 2015.

[49] Il Patto è previsto dalla "Strategia per la crescita digitale" 2014-2020 e dal "Patto per la Salute" 2014-2016.

[50] Parlamento europeo, Consiglio, Commissione europea – Dichiarazioni comuni, *Dichiarazione europea sui diritti e i principi digitali per il decennio digitale* (2023/C 23/01).
Circa la Dichiarazione v.: D.A. LIMONE, *"Dichiarazione europea sui principi e diritti digitali" - Per il decennio digitale 2020-2030*, in *Rivista elettronica di Diritto, Economia, Management*, 3/2025; F. ZORZI GIUSTINIANI, *La Dichiarazione europea sui diritti e principi digitali, il Trans-Atlantic Data Privacy Framework e l'Unione Federale dei consumatori tedeschi contro Meta Platforms Ireland*, in *Nomos*, 1/2022; M. PIANOFORTE, *Cittadinanza digitale: la proposta di una dichiarazione solenne sui diritti e i principi digitali*, in *Eurojus*, 14.02.2022; P. DE PASQUALE, *Verso una Carta dei diritti digitali (fondamentali) dell'Unione europea?*, in *Il diritto dell'Unione europea*, 17.03.2022; S. TOMMASI, *La proposta di Dichiarazione europea sui diritti e i principi digitali per il decennio digitale COM(2022) 28 final del 26 gennaio 2022*, in *Persona e mercato*, 1/2022, 168-171.

A livello nazionale, seppure privo di valore cogente, l'art. 13 della "Dichiarazione dei diritti in Internet"[51], rubricato «sicurezza in rete», identifica la sicurezza quale interesse sia pubblico che privato, sottolineando in particolare che «la sicurezza in Rete deve essere garantita come interesse pubblico, attraverso l'integrità delle infrastrutture e la loro tutela da attacchi, e come interesse delle singole persone».

Tale esigenza, oltre che da diversi atti di *soft law*, risulta oggi rafforzata anche da un insieme di provvedimenti normativi che risultano portatori di una nuova consapevolezza. Tra questi, centrale è anzitutto l'articolo 2, punto 1) del Regolamento (UE) 2019/881, c.d. "Cybersecurity Act"[52], il quale definisce la "cybersicurezza" quale «insieme delle attività necessarie per proteggere la rete e i sistemi informativi, gli utenti di tali sistemi e altre persone interessate dalle minacce informatiche», evolvendosi quindi da protezione delle infrastrutture informatiche anche a protezione delle persone e, in questo contesto, dei pazienti e del personale sanitario.

Diverse sono poi le fonti normative di rango primario, europee e nazionali, in materia, in primis l'ormai abrogata Direttiva "NIS1"[53] , oggi sostituita dalla Direttiva "NIS2"[54] e volta ad istituire un livello rafforzato di governance della cybersicurezza, così come recepita a livello nazionale con il D.Lgs. 138/2024[55].

Con il Regolamento Ue 2019/881 del 17 aprile 2019, c.d. "Cybersecurity Act", sono stati invece integrati i poteri dell'Agenzia dell'Unione europea per la Cybersicurezza (ENISA), ed è stato previsto un «sistema di certificazione della cybersicurezza per le tecnologie dell'informazione e della comunicazione».

---

[51] Camera dei Deputati – Commissione per i diritti e i doveri relativi ad Internet, *Dichiarazione dei diritti di Internet,* 14 luglio 2015.
A riguardo v.: M. BASSINI, *Né costituzione né legge. La Dichiarazione dei diritti in Internet verso una missione culturale*, in *Medialaws*, 28 luglio 2015; A. MORELLI, *I diritti e la Rete. Notazioni sulla bozza di Dichiarazione dei diritti in Internet*, in *federalismi.it*, 1/2015; L. NANNIPIERI, *Sulla "Dichiarazione dei diritti in Internet". Alcune annotazioni critiche*, in *Inf. e dir*., 2/2014, 127-138; L. ABBA, A. ALÙ, *Il valore della Carta dei diritti di Internet*, Esi, Napoli, 2020.

[52] Regolamento (UE) 2019/881 *relativo all'ENISA, l'Agenzia dell'Unione europea per la cibersicurezza, e alla certificazione della cibersicurezza per le tecnologie dell'informazione e della comunicazione, e che abroga il regolamento (UE) n. 526/2013 («regolamento sulla cibersicurezza»)*, del 17 aprile 2019.

[53] Direttiva (UE) 2016/1148 del 6 luglio 2016..

[54] Direttiva (UE) 2022/2555 del 14 dicembre 2022.
A riguardo si veda: S. UNGARO (a cura di*), NIS2 e Cybersecurity,* in *Rivista elettronica di Diritto, Economia, Management*, 1/2025; M. BUFFA, *La Direttiva NIS II cybersecurity in Europa: tra innovazione, formazione e diritto vivente*, in *Democrazie e Diritti Sociali*, 1/2023, 47 ss.

[55] Decreto legislativo 4 settembre 2024, n. 138, in tema di *recepimento della direttiva (UE) 2022/2555, relativa a misure per un livello comune elevato di cibersicurezza nell'Unione, recante modifica del regolamento (UE) n. 910/2014 e della direttiva (UE) 2018/1972 e che abroga la direttiva (UE) 2016/1148.*

Con il c.d. “Cyber Resilience Act” (Reg. Ue 2024/2847 del 23 ottobre 2024)[56], si è poi inteso sviluppare «un quadro normativo orizzontale dell’Unione che stabilisca requisiti di cybersicurezza completi per tutti i prodotti con elementi digitali» e che, tuttavia, espressamente, non trova applicazione nel campo dei dispositivi medici.Tale normativa, dalla quale non emerge ancora oggi in maniera chiara, di fronte ad un’esigenza che è invece forte, evidente socialmente percepita, una formale qualificazione della cybersicurezza alla stregua di un diritto, ma pone tuttavia sempre maggiori condizioni, anche giuridiche, in tale direzione.

Ciò è ancora più chiaro nel settore sanitario, dove le tecnologie utilizzate, in questo caso i dispositivi elettromedicali, a fronte di una evidente utilità, sono in grado di mettere in pericolo, qualora non siano rispettate le condizioni di cybersicurezza, non solo i dati dei pazienti e del personale medico, ma anche la loro stessa salute.

Sul punto, come noto, il legislatore europeo ha emanato già da qualche anno un apposito Regolamento relativo ai dispositivi medici, il Reg. UE 2017/745 (MDR)[57], allo «scopo di stabilire un quadro normativo solido, trasparente, prevedibile e sostenibile per i dispositivi medici, che garantisca un livello elevato di sicurezza e di salute sostenendo nel contempo l'innovazione», fissando standard di qualità e sicurezza.

Il Regolamento, destinato ai produttori di dispositivi medici, considera tali qualunque strumento, apparecchio, apparecchiatura, software, impianto, reagente, materiale o altro articolo, destinato dal fabbricante a essere impiegato sull'uomo, da solo o in combinazione, per una o più delle seguenti destinazioni d'uso mediche specifiche, ossia la diagnosi, la prevenzione, il monitoraggio, la previsione, la prognosi, il trattamento o attenuazione di malattie, la compensazione di una lesione o di una disabilità, lo studio, sostituzione o modifica dell'anatomia oppure di un processo o stato fisiologico o patologico, o la fornitura di informazioni attraverso l'esame in vitro di campioni provenienti dal corpo umano.

Possono quindi essere immessi sul mercato solo i dispositivi che rispettano i requisiti di sicurezza previsti dal Regolamento, ossia requisiti relativi alla sicurezza generale del dispositivo, ad esempio relativamente ai materiali utilizzati o al funzionamento, sia requisiti specifici derivanti dalla destinazione d’uso.

Principio generale previsto dal Regolamento è che tali dispositivi debbano fornire le prestazioni previste dal loro fabbricante e debbano essere progettati e fabbricati in modo tale che, in normali condizioni

[56] Regolamento (UE) 2024/2847 del 23 ottobre.
[57] Regolamento (UE) 2017/745 del 5 aprile 2017.

d'uso, siano adatti alla loro destinazione. Essi, prevede il Regolamento, devono essere prodotti in modo da essere sicuri ed efficaci, non compromettendo lo stato clinico o la sicurezza dei pazienti, né la sicurezza e la salute degli utilizzatori ed eventualmente di altre persone. Sono invece considerati accettabili gli eventuali rischi associabili al loro utilizzo, qualora siano stati considerati i benefici apportati al paziente e siano compatibili con un elevato livello di protezione della salute e della sicurezza, tenendo conto dello stato dell'arte generalmente raggiunto.
Quanto alla componente tecnologica, qui di principale interesse, il Regolamento, nell'allegato 1, punto 14.2, prevede che i dispositivi siano progettati e fabbricati in modo tale da eliminare o ridurre, per quanto possibile, i rischi associati alla possibile interazione negativa tra il software e l'ambiente tecnologico in cui opera e interagisce. Inoltre, i dispositivi contenenti sistemi elettronici programmabili, compresi i software, o i software che costituiscono dispositivi a sé stanti, devono essere progettati in modo tale da garantire la riproducibilità, l'affidabilità e le prestazioni in linea con la destinazione d'uso per essi prevista.
Per tale ultima categoria di dispositivi è previsto, ancora, che il software sia sviluppato e fabbricato conformemente allo stato dell'arte, tenendo conto dei principi del ciclo di vita dello sviluppo, della gestione del rischio, compresa la sicurezza delle informazioni, della verifica e della convalida. Inoltre, al fine di non compromettere le condizioni di sicurezza implementate, i fabbricanti devono indicare i requisiti minimi in materia di hardware, le caratteristiche delle reti informatiche e le misure di sicurezza informatica, necessari per il funzionamento del software secondo le previsioni.

Sulla base del Regolamento MDR e degli obblighi previsti in tema di progettazione e sviluppo di dispositivi medici, tra i quali rientrano anche i dispositivi elettromedicali, il tema della cybersicurezza è quindi affrontato in maniera generale, rappresentando solo uno dei diversi elementi costruttivi tesi a garantire la sicurezza del dispositivo.
È al contempo possibile evidenziare come, complice il fatto che trattasi di una normativa risalente a quasi un decennio fa, tale approccio alla sicurezza e alla cybersicurezza, limitato alla progettazione "conformemente allo stato dell'arte" rispetto alla destinazione d'uso, se da una parte può apparire assai generico e forse insufficiente, dall'altra, tramite tale clausola, consente l'adeguamento allo sviluppo tecnologico, nonché, alle normative speciali e di successiva emanazione.

È proprio in questo solco che si inseriscono oggi il Reg. Ue 679/2016 e la Direttiva NIS2, delle quali occorre ora vagliare i punti di contatto con il Regolamento MDR.

Rispetto alla normativa in tema di protezione dei dati personali, questa trova un'applicazione specifica che, tuttavia, visto il settore di utilizzo dei dispositivi medici e di quelli elettromedicali in particolare, basati sull'acquisizione ed utilizzo di dati relativi ai pazienti e agli operatori sanitari, diviene invece di carattere generale.

Anzitutto, la tutela della salute mediante dispositivi medici tecnologici, comportando un grande trattamento di dati, richiede il rispetto delle garanzie previste dal GDPR (Reg. Ue 679/2016), il quale, quanto alla categoria dei dati relativi alla salute[58] riflette le intenzioni del legislatore europeo di garantirne una protezione rafforzata e al contempo riconoscerne la legittimità del trattamento nei casi previsti nel paragrafo 2 dell'art. 9 GDPR, ad esempio quando ricorra la "finalità di cura".

Rispetto alla richiamata condizione, che legittima nel caso di specie il trattamento dei dati personali, si veda come tali dispositivi abbiano una doppia anima: infatti, se nell'ambito del normale uso per finalità assistenziali sia proprio la c.d. "finalità di cura" a legittimarne l'utilizzo, negli altri casi, ad esempio per lo svolgimento di indagini cliniche sul dispositivo stesso o per finalità di ricerca, la base giuridica è solitamente da ricondurre in un consenso specifico ed espresso.

Quanto agli specifici requisiti di sicurezza, il GDPR prevede, all'art. 32, l'adozione di garanzie di sicurezza del trattamento dei dati personali, anzitutto basate sulla valutazione dello «stato dell'arte e dei costi di attuazione, nonché della natura, dell'oggetto, del contesto e delle finalità del trattamento, come anche del rischio di varia probabilità e gravità per i diritti e le libertà delle persone fisiche», rimettendo quindi al produttore, per quanto riguarda le misure strutturali del dispositivo, e all'utilizzatore (la struttura sanitaria), principalmente per quanto riguarda quelle di propria competenza, l'onere di decidere quali misure adottare.

Tuttavia, come già evidenziato, i rischi connessi ai dispositivi medici e a quelli elettromedicali in particolare, non riguardano solo i dati personali ma anche quelli non personali, nonché la persona nella sua dimensione fisica: da un attacco informatico, infatti, possono derivare diretti riflessi sulla salute della persona, qualora, ad esempio, venga inficiato il corretto funzionamento di un dispositivo che controlla il dosaggio di un farmaco.

[58] M. IASELLI (a cura di), *La tutela dei dati personali in ambito sanitario*, Giuffrè, 2020; C. COLAPIETRO, F. LAVIOLA, *I trattamenti di dati personali in ambito sanitario*, in *Dirittifondamentali.it*, 2/2019; C. FILIPPI, S. MELCHIONNA, *I trattamenti di dati in ambito sanitario*, in G. BUSIA, L. LIGUORI, O. POLLICINO (a cura di), *Le nuove frontiere della privacy nelle tecnologie digitali. Bilanci e prospettive*, Aracne, Roma, 2016, 469 ss.; G. FINOCCHIARO, *Il trattamento dei dati sanitari: alcune riflessioni critiche a dieci anni dall'entrata in vigore del Codice in materia di protezione dei dati personali*, in *Sanità pubblica e privata*, 2/2009.

Se il Regolamento MDR ha posto specifici requisiti di costruzione del dispositivo, è con la Direttiva NIS2, così come recepita per il tramite del D.Lgs. 4 settembre 2024 n. 138 ed innovando rispetto alla NIS2, che sono state previste condizioni specifiche ma al contempo trasversali di cybersicurezza, in maniera proporzionata al rischio connesso al settore di applicazione e alla dimensione organizzativa di riferimento.

In particolare, i produttori di dispositivi medici, così come qualificati dal Regolamento MDR, sono espressamente chiamati, come si evince dall'allegato 1 della NIS2, all'applicazione della normativa, nonché quando operanti in uno dei settori ad "alta criticità" ricompresi nel citato allegato, quindi nel caso in cui la produzione si inserisca all'interno di un'emergenza di sanità pubblica.

Quanto agli obblighi in materia di cybersicurezza, questi sono stati specificamente declinati nell'art. 24 del decreto legislativo n. 138 e consistono, principalmente, nell'adozione di misure tecniche, operative e organizzative adeguate e proporzionate alla gestione dei rischi posti alla sicurezza dei sistemi informativi e di rete, nonché per prevenire o ridurre al minimo l'impatto di eventuali incidenti. Inoltre, il decreto prevede specifiche misure tra le quali rileva, con particolare riguardo al tema dei dispositivi elettromedicali, l'obbligo di garantire la sicurezza della catena di approvvigionamento: rispetto a tale dimensione, diversamente da quanto previsto dalla legge[59], l'elemento della cybersicurezza, nell'ambito delle strutture ospedaliere e a fronte di una potenziale disponibilità di un parco tecnologico sempre più all'avanguardia, non sempre viene privilegiato, complice anche un deficit di carattere economico e la contrapposta disponibilità di strumenti risalenti.

Rispetto alla sicurezza informatica dei dispositivi medici, è rilevante notare come, a fronte della citata normativa di carattere trasversale, anche le Autorità di settore si stiano adeguando per darvi concreto seguito: ad esempio l'Agenas, l'Agenzia Nazionale per i Servizi Sanitari Regionali, ha recentemente emanato, nel settembre 2025, specifiche "Linee Guida di indirizzo per il percorso evolutivo dei Sistemi Medicali per Telemonitoraggio"[60], fissando, per questa tipologia di dispositivi, requisiti chiari e dettagliati, tra i quali rientrano: la presenza della c.d. certificazione MDR, per cui tutti i dispositivi devono essere conformi al Regolamento (UE) 2017/745; requisiti di connettività, per cui i dispositivi devono poter trasmettere dati in modo sicuro; condizioni di sicurezza informatica, con indicazioni specifiche per produttori, distributori e strutture sanitarie.

[59] Si fa riferimento, in particolare, alla recente introduzione, nell'ambito del d.lgs. 36/2023, dell'art. 108 co. 4, relativo alla valutazione degli elementi di cybersicurezza nelle procedure di evidenza pubblica aventi ad oggetto beni e servizi informatici, così come ulteriormente specificato dall'ACN con le recentissime "Linee Guida per l'applicazione dei criteri di premialità di cui all'articolo 14 della Legge n. 90/2024".

[60] Agenzia Nazionale per i Servizi Sanitari Regionali - *Linee Guida di indirizzo per il percorso evolutivo dei Sistemi Medicali per Telemonitoraggio*, del 9 dicembre 2024.

Più ampiamente, con decreto del Ministero della Salute n. 25 del 14 gennaio 2025[61] e facendo seguito alle previsioni del D.Lgs. n. 138, è stato istituito presso il Ministero della Salute un apposito Tavolo di settore per l'attuazione dalla Direttiva NIS2, il quale vede, tra i diversi temi cardine, proprio quello della cybersicurezza dei dispositivi medici e che, si auspica, nell'ambito del rapporto tra le diverse fonti citate, giunga alla previsione di requisiti di cybersicurezza specifici per i dispositivi elettromedicali.

In tale direzione sembra volgere anche la recente previsione di istituzione del c.d. "Centro europeo di sostegno alla cybersicurezza negli ospedali e dei fornitori di servizi sanitari": trattasi di una articolazione prevista normativamente come interna all'ENISA, e menzionata in una apposita comunicazione della Commissione europea[62] del 15 gennaio 2025, con la quale è stato presentato il "Piano d'azione sulla cybersicurezza degli ospedali e dei fornitori di servizi sanitari".

Quali sfide per una sanità digitale sicura?

Evidenziato il quadro di riferimento in merito ai dispositivi elettromedicali, appare anzitutto evidente come la continua e traversale opera di regolazione sia caratterizzata dalla necessità di contemperamento tra il fine della tutela della persona e lo sviluppo tecnologico, quest'ultimo sia in funzione servente rispetto al primo, sia in sé considerato.
Le situazioni giuridiche coinvolte sono, inoltre, di primario interesse costituzionale, sia in termini di tutela dei diritti fondamentali della persona, sia in termini di corretta gestione, "efficace" ed "efficiente" della "cosa pubblica".
Il caso dei dispositivi medici elettromedicali evidenzia come, al fine della loro conformità sia ormai richiesta l'applicazione congiunta di un vasto novero di principi e adempimenti. In particolare, anche in ragione della novità della normativa, il settore della sanità digitale si trova oggi investito dall'esigenza di garantire nuove istanze di protezione della persona, anche attraverso garanzie di cybersicurezza.

Alla luce di quanto evidenziato sul punto, deriva anzitutto la constatazione di come la cybersicurezza, elemento centrale rispetto ai dispositivi altamente tecnologici, avendo raggiunto ormai una sostanziale

[61] Decreto del Ministero della Salute del n. 25 del 14 gennaio 2025, rubricato "Istituzione presso il Ministero della Salute del tavolo settoriale per l'attuazione della Direttiva NIS".
[62] Comunicazione COM(2025) 10 final, del 15 gennaio 2025.

autonomia rispetto alle normative entro le quali prima era precedentemente sistematizzata, ad esempio la protezione dei dati personali, non sia più un'esigenza di carattere solo settoriale ma, invece, trasversale e, come tale, questa si estrinsechi attraverso obblighi che riguardano orizzontalmente anche il settore sanitario e in particolare il settore dei dispositivi medici, i quali trovano quindi la propria fonte non solo nel Regolamento MDR, ma anche nella Direttiva NIS2, nel GDPR per quanto riguarda i profili di tutela dei dati personali, e negli atti emanati o in via di emanazione da parte delle autorità di settore.

Più ampiamente, e ciò è ancor più evidente nel settore sanitario, sembra che la cybersicurezza a cui fare riferimento sia in via di evoluzione e, da una mera caratteristica accessoria, di un'attività, di uno strumento o di un diritto, sia ormai sempre più connotata da elementi tipici dei diritti stessi in sé considerati. La cybersicurezza, in questo senso, dovrebbe essere intesa, in un'ottica evolutiva, quale "nuovo" diritto che, parallelamente alla protezione dei dati personali e con esso sovente intersecandosi, sia funzionale a garantire la libertà e la dignità delle persone, di agire nell'ambiente digitale e di vivere in una realtà materiale sempre più caratterizzata da conseguenze effettive, anche pregiudizievoli, derivanti dall'uso delle tecnologie.

Bibliografia

- APERIO BELLA F., *L'accesso alle tecnologie innovative nel settore salute tra universalità e limiti organizzativi (con una postilla sull'emergenza sanitaria)*, in *Persona e Amministrazione*, 1/ 2020;
- AZZARITI G., *Stefano Rodotà e l'uso costituzionalmente orientato delle nuove tecnologie*, in *Costituzionalismo.it*, 2, 2022;
- BALDUZZI R., *Salute* (diritto alla), in CASSESE S. (diretto da), *Dizionario di diritto pubblico*, Giuffré, Milano, 2006;
- BARBARESCHI S., *Tecniche di individuazione del nucleo essenziale del diritto alla salute nella giurisprudenza costituzionale*, in COLAPIETRO C., ATRIPALDI M., FARES G. e IANNUZZI A. (a cura di), *I modelli di welfare sanitario tra qualità e sostenibilità. Esperienze a confronto*, Esi, Napoli, 2018;
- BASSINI M., *Né costituzione né legge. La Dichiarazione dei diritti in Internet verso una missione culturale*, in *Medialaws*, 28 luglio 2015;
- BUFFA M., *La Direttiva NIS II cybersecurity in Europa: tra innovazione, formazione e diritto vivente*, in *Democrazie e Diritti Sociali*, 1/2023;

- CARTABIA M., *La giurisprudenza costituzionale relativa all'art. 32, secondo comma, della Costituzione italiana*, in *Quad. cost.*, 2012;
- COLAPIETRO C., ATRIPALDI M., FARES G. e IANNUZZI A. (a cura di), *I modelli di welfare sanitario tra qualità e sostenibilità. Esperienze a confronto*, Esi, Napoli, 2018;
- COLAPIETRO C., LAVIOLA F., *I trattamenti di dati personali in ambito sanitario*, in *Dirittifondamentali.it*, 2/2019;
- COVINO F., *Uso della tecnologia e protezione dei dati personali sulla salute tra pandemia e normalità*, in *federalismi.it*, 5/2021;
- DE PASQUALE P., *Verso una Carta dei diritti digitali (fondamentali) dell'Unione europea?*, in *Il diritto dell'Unione europea*, 17.03.2022;
- DI COSTANZO C., *L'impiego delle nuove tecnologie nel settore della salute: problematiche e prospettive di diritto costituzionale*, in *Consulta Online*, 1/2023;
- FARALLI C., BRIGHI R., MARTONI M. (a cura di), *Strumenti, diritti, regole e nuove relazioni di cura: il paziente europeo protagonista nell'eHealth*, Giappichelli, Torino, 2015;
- FILIPPI C., MELCHIONNA S., *I trattamenti di dati in ambito sanitario*, in BUSIA G., LIGUORI L., POLLICINO O. (a cura di), *Le nuove frontiere della privacy nelle tecnologie digitali. Bilanci e prospettive*, Aracne, Roma, 2016;
- FINOCCHIARO G., *Il trattamento dei dati sanitari: alcune riflessioni critiche a dieci anni dall'entrata in vigore del Codice in materia di protezione dei dati personali*, in *Sanità pubblica e privata*, 2/2009;
- IASELLI M. (a cura di), *La tutela dei dati personali in ambito sanitario*, Giuffrè, 2020;
- IRTI C., *Riflessioni sull'impatto delle nuove tecnologie nell'ambito sanitario, tra regole tecniche e regole giuridiche*, in *EJPLT*, 1/2023;
- LIMONE D.A. (a cura di), *La sanità digitale*, in *Rivista elettronica di diritto, economia, management*, 3/2014;
- LIMONE D.A., *"Dichiarazione europea sui principi e diritti digitali" - Per il decennio digitale 2020-2030*, in *Rivista elettronica di Diritto, Economia, Management*, 3/2025;
- LUCARELLI TONINI L.M., *Il diritto di accesso alle cure: nuovi orizzonti nella personalizzazione del trattamento sanitario*, in *Teoria e prassi del diritto*, 2/2023;
- MASCHIO F., *Innovazione digitale dei sistemi sanitari, eHealth e questioni di data governance, nell'agenda digitale sanitaria nazionale ed europea*, in *Comparazione e diritto civile*, 2/2021;
- MICCÙ R., *Questioni attuali intorno alla digitalizzazione dei servizi sanitari nella prospettiva multilivello*, in *federalismi.it*, 5/2021;

- MINNI F., MORRONE A., *Il diritto alla salute nella giurisprudenza della Corte Costituzionale italiana*, in *Rivista AIC*, 2/2013;
- MORANA D., *La salute come diritto costituzionale. Lezioni*, Giappichelli, Torino, 2015;
- MORELLI A., *I diritti e la Rete. Notazioni sulla bozza di Dichiarazione dei diritti in Internet*, in *federalismi.it*, 1/2015;
- NANNIPIERI L*., Sulla "Dichiarazione dei diritti in Internet". Alcune annotazioni critiche*, in *Inf. e dir.*, 2/2014;
- ABBA L., ALÙ A., *Il valore della Carta dei diritti di Internet*, Esi, Napoli, 2020;
- PIANOFORTE M., *Cittadinanza digitale: la proposta di una dichiarazione solenne sui diritti e i principi digitali*, in *Eurojus*, 14.02.2022;
- PIETROLETTI A., NICOTRA A., *Tutela della salute, sistemi digitali e privacy*, in *Inf. e dir.*, 1/2022;
- RODOTÀ S., *Il corpo "giuridificato"*, in RODOTÀ S., ZATTI P., *Trattato di biodiritto*, Giuffrè, Milano, 2011;
- THIENE A., CORSO S. (a cura di), *La protezione dei dati sanitari. Privacy e innovazione tecnologica tra salute pubblica e diritto alla riservatezza*, Jovene, 2023;
- TOMMASI S., *La proposta di Dichiarazione europea sui diritti e i principi digitali per il decennio digitale COM(2022) 28 final del 26 gennaio 2022*, in *Persona e mercato*, 1/2022;
- UNGARO S. (a cura di*), NIS2 e Cybersecurity,* in *Rivista elettronica di Diritto, Economia, Management*, 1/2025;
- ZORZI GIUSTINIANI F*., La Dichiarazione europea sui diritti e principi digitali, il Trans-Atlantic Data Privacy Framework e l'Unione Federale dei consumatori tedeschi contro Meta Platforms Ireland*, in *Nomos*, 1/2022.

Il caso Anthropic vs OpenAI negli USA: sicurezza nazionale, concorrenza e diritti nell'era dell'AI – di ***Samuele Minuti***, Dottorando di Ricerca in Law and Cognitive Neuroscience presso l'Università degli Studi "Niccolò Cusano"

Dal bando federale contro Anthropic al nuovo asse tra Pentagono e OpenAI: come la sicurezza nazionale sta ridisegnando mercato, governance e diritti nell'intelligenza artificiale, e quali lezioni per l'Europa e per l'Italia

Il 27 febbraio 2026 il presidente Donald Trump ha annunciato di aver ordinato a "ogni agenzia federale" di cessare immediatamente l'uso delle tecnologie di Anthropic, con un periodo di sei mesi per completare il phase-out presso il Dipartimento della Difesa e le amministrazioni che utilizzano i suoi sistemi nei processi critici[63]. L'annuncio è arrivato all'indomani di un ultimatum del Pentagono: o Anthropic avrebbe accettato di rimuovere le limitazioni sull'uso militare dei propri modelli di AI, consentendone un impiego sostanzialmente illimitato, oppure sarebbe stata considerata un rischio per la sicurezza della supply chain.[64]

Qualche ora dopo, il segretario alla Difesa Pete Hegseth ha effettivamente designato Anthropic come "supply-chain risk", una qualifica solitamente riservata a società di Paesi rivali, che comporta il divieto per i contractor del settore difesa di utilizzare i suoi prodotti nei progetti collegati al Pentagono. In un post su Truth Social, Trump ha definito Anthropic una "AI fuori controllo" gestita da "left-wing nut jobs" e ha promesso che il governo "non avrà più bisogno" dei suoi servizi, presentando la decisione come risposta alla volontà dell'azienda di imporre vincoli etici ai militari.

Come affermato dal matematico Clive Humby, l'IA generativa[65] nasce e si sviluppa in un contesto in cui i dati personali sono diventati "il nuovo petrolio" dell'economia digitale e l'individuo rischia di vivere in

[63] Vedi https://www.abc.net.au/news/2026-02-28/us-president-orders-six-month-phaseout-of-anthropic-technology/106400088.
[64] Vedi https://www.bbc.com/news/articles/cn48jj3y8ezo.
[65] M. FLORA, L. MAGGI, *Intelligenza artificiale generativa. Opportunità e sfide legali*, Pacini Editore, Pisa, 2024. Vedi anche DI BELLO B., *Ai pro. Progettare, creare e comunicare con l'intelligenza artificiale generativa*, Hoepli,

un “castello di vetro”, per usare l’espressione di Stefano Rodotà: esposto, profilato, osservato in ogni sua azione online e offline.

In questo quadro, decisioni come quella contro Anthropic non riguardano solo un fornitore di AI, ma toccano l’equilibrio tra potere pubblico, diritti fondamentali e mercato dei dati; dunque, parlano direttamente anche al modello europeo di protezione dei dati e di regolazione dell’AI.

Dal modello Biden alla svolta Trump: due stagioni dell’AI policy USA

Per una corretta collocazione sistematica del caso, occorre richiamare l’Executive Order 14110, firmato da Joe Biden il 30 ottobre 2023, volto a delineare una cornice organica per uno sviluppo dell’intelligenza artificiale qualificato come “safe, secure and trustworthy”. L’ordine prevedeva, inter alia, obblighi di condivisione con il governo dei risultati dei test di sicurezza relativi ai modelli di frontiera, l’elaborazione di linee guida da parte del NIST in materia di red teaming[66], nonché misure dirette alla mitigazione dei rischi sistemici e delle vulnerabilità. A ciò si aggiungevano interventi concernenti la tutela della privacy, la non discriminazione, la sicurezza informatica, i diritti dei lavoratori e la concorrenza.

In particolare, l’Executive Order 14110[67] ha assolto a tre funzioni principali: in primo luogo, ha previsto l’istituzione della figura del Chief AI Officer presso le principali agenzie federali, con compiti di governance interna; in secondo luogo, ha rafforzato il ruolo delle autorità antitrust — quali la Federal Trade Commission e il Dipartimento di Giustizia (Antitrust Division) — nella vigilanza sui fenomeni di concentrazione e sulle pratiche anticoncorrenziali nel settore dell’intelligenza artificiale; in terzo luogo, ha sollecitato le agenzie federali a valutare l’impatto dei sistemi automatizzati su ambiti sensibili quali i diritti civili, l’accesso alle prestazioni sociali, la giustizia penale, l’immigrazione e l’occupazione.

Come evidenziato anche in dottrina, il Regolamento generale sulla protezione dei dati rappresenta, nel contesto europeo, un passaggio analogo in chiave “fundamental rights oriented”, segnando il

Milano, 2024 e RUFFOLO U., AMIDEI A., *Diritto dell’Intelligenza Artificiale. Proprietà industriale e intellettuale. CorpTech. Giustizia Predittiva. Transumanesimo. AI generativa. Metaverso*, LUISS University Press, Roma, 2024.

[66] Per un approfondimento si veda, tra gli altri, PREVITI L., Pubblici poteri e cybersicurezza: il lungo cammino verso un approccio collaborativo alla gestione del rischio informatico, in Federalismi.it, 25, 2022.
[67] Executive Order 14110, *Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, Joe Biden, 30 ottobre 2023, disponibile online presso il sito ufficiale della Casa Bianca.

superamento di un approccio meramente market-driven a favore del riconoscimento del diritto alla protezione dei dati personali quale diritto fondamentale, autonomo ma strettamente connesso alla tradizionale tutela della privacy. Tale impostazione ha influenzato anche il dibattito statunitense in materia di intelligenza artificiale; tuttavia, negli Stati Uniti gli strumenti normativi risultano tuttora più frammentati e maggiormente esposti alla discrezionalità dell'esecutivo, a differenza del modello europeo fondato su un regolamento direttamente applicabile.

L'insediamento di Donald Trump nel gennaio 2025 ha segnato una significativa discontinuità: mediante un nuovo Executive Order, la Casa Bianca ha revocato l'EO 14110, ritenuto idoneo a imporre "costi e burocrazia soffocanti" a un settore considerato strategico per la leadership globale degli Stati Uniti.

Il successivo ordine del dicembre 2025, intitolato "Ensuring a National Policy Framework for Artificial Intelligence"[68], ha ribadito l'obiettivo di un assetto regolatorio "pro-innovazione", impegnando il Dipartimento di Giustizia a contestare le normative statali giudicate eccessivamente onerose e affermando esplicitamente la volontà di eliminare gli ostacoli alla leadership nazionale nel settore dell'intelligenza artificiale.

In tale rinnovato contesto, più che alla costruzione di un corpus normativo organico, la governance dell'intelligenza artificiale appare affidata in larga misura alla discrezionalità dell'esecutivo, secondo una logica che combina deregolamentazione generale e interventi selettivi — talora particolarmente incisivi — nei confronti di singoli operatori. In questa prospettiva, il caso Anthropic non si configura come un episodio isolato, bensì come espressione di una più ampia strategia nella quale la sicurezza nazionale è utilizzata quale leva per orientare il mercato e per delimitare il perimetro di accettabilità delle condotte dei fornitori di intelligenza artificiale[69].

Dentro lo scontro Pentagono-Anthropic: cronologia di una rottura

Anthropic, fondata nel 2021 da ex dirigenti di OpenAI, si è rapidamente affermata come uno dei principali sviluppatori di modelli linguistici avanzati — in particolare la serie Claude — caratterizzando la propria attività per una marcata attenzione ai profili di sicurezza e alla mitigazione dei rischi estremi.

[68] Ensuring a National AI Policy Framework, *Ensuring a National Policy Framework for Artificial Intelligence*, Donald Trump, 2025, disponibile online presso fonti ufficiali del governo degli Stati Uniti.

[69] Per un approfondimento si veda, C. NOVELLI, L. FLORIDI, A. GAUR, *Two futures of AI regulation under the Trump administration*, 2025, https://digi-con.org/two-futures-of-ai-regulation-under-the-trump-administration/.

A partire dal 2024, i suoi sistemi sono stati adottati da diverse agenzie federali e dal Dipartimento della Difesa per applicazioni di analisi testuale, assistenza alla redazione e supporto operativo, nonché da primari operatori del settore tecnologico privato.

Nel 2025, Anthropic ha concluso un contratto pluriennale con il Pentagono per l'impiego dei modelli Claude in reti classificate; tuttavia, in sede di rinnovo, la società ha subordinato la prosecuzione del rapporto all'accettazione di due condizioni ritenute non negoziabili: da un lato, il divieto di utilizzo dei propri modelli per attività di sorveglianza di massa nei confronti di cittadini statunitensi e per sistemi d'arma completamente autonomi; dall'altro, l'impegno a garantire un "controllo umano significativo" nelle decisioni di targeting e di ingaggio. Il Dipartimento della Difesa ha qualificato tali clausole come eccessivamente restrittive, sostenendo che le forze armate devono poter impiegare l'intelligenza artificiale in tutte le modalità consentite dal diritto internazionale umanitario e dalla normativa interna, senza che un operatore privato introduca unilateralmente ulteriori limitazioni.

All'esito di una fase negoziale — nel corso della quale, secondo fonti giornalistiche, sembrava profilarsi una soluzione di compromesso volta a ridimensionare, ma non eliminare, i divieti — la tensione è ulteriormente cresciuta quando l'amministratore delegato Anthony Amodei ha pubblicamente ribadito l'impossibilità, "in coscienza", di acconsentire a un utilizzo illimitato dei modelli in tali contesti. A fronte di ciò, il Segretario alla Difesa Pete Hegseth ha prospettato possibili "conseguenze" in caso di mancato adeguamento delle condizioni contrattuali, mentre il Presidente Donald Trump ha progressivamente politicizzato il confronto, accusando la società di voler "ricattare" il Dipartimento della Difesa.

Da tale dinamica è scaturita una sequenza di interventi coordinati: l'adozione di un ordine presidenziale di phase-out, la designazione di Anthropic quale "supply chain risk", la comunicazione ai contraenti del settore difesa del divieto di utilizzo dei suoi prodotti e l'emanazione di direttive rivolte alle amministrazioni civili per l'interruzione dell'impiego dei relativi sistemi nei servizi interni. In tale contesto, il Dipartimento del Tesoro, la Federal Housing Finance Agency e il Dipartimento di Stato — che aveva denominato "StateChat" un chatbot interno basato sui modelli della società — hanno annunciato la migrazione verso soluzioni alternative, frequentemente individuate in ChatGPT quale opzione di riferimento.

OpenAI, la finestra di opportunità e il nuovo asse con il Pentagono

Poche ore dopo il bando nei confronti di Anthropic, OpenAI ha annunciato la conclusione di un accordo con il Dipartimento della Difesa per la fornitura dei propri modelli — inclusi GPT-4.1 e successivi —

destinati all'impiego in reti classificate, con applicazioni che spaziano dall'analisi di intelligence al supporto decisionale operativo. Tale intesa consente a OpenAI di consolidare la propria posizione quale partner privilegiato del governo federale in una fase in cui numerose amministrazioni sono chiamate a sostituire, in tempi ristretti, infrastrutture di intelligenza artificiale già integrate nei processi operativi e negli strumenti di lavoro quotidiani.

L'amministratore delegato Sam Altman ha espresso pubblicamente sostegno alle "red lines" indicate da Anthony Amodei, affermando che anche OpenAI rifiuterà contratti che prevedano attività di sorveglianza di massa o lo sviluppo di sistemi d'arma completamente autonomi, e precisando che tali limiti sono stati formalizzati in apposite clausole contrattuali con il Pentagono. La differenza rilevante risiede tuttavia nel fatto che, nel caso di OpenAI, tali clausole sono state accettate dal governo, integrate nelle policy interne e presentate come standard applicabili — almeno in linea teorica — all'insieme dei fornitori di intelligenza artificiale operanti nel settore della difesa.

Allo stesso tempo, il trasferimento forzato dei contratti pubblici da Anthropic verso altri operatori — in primis OpenAI e, in misura minore, Google — appare suscettibile di rafforzare ulteriormente assetti di tipo oligopolistico nel mercato dell'intelligenza artificiale generativa, anche alla luce del fatto che tali soggetti esercitano un controllo significativo su infrastrutture cloud, piattaforme digitali e servizi essenziali per la trasformazione tecnologica. Ne consegue che una decisione formalmente giustificata da esigenze di sicurezza nazionale è idonea a produrre effetti diretti sulla struttura concorrenziale del mercato, determinando un rafforzamento dell'operatore percepito come maggiormente allineato — o, quantomeno, più capace di negoziare — con le priorità dell'esecutivo.

Le azioni legali di Anthropic e il primo stop del giudice

Il 9 marzo 2026, Anthropic ha proposto due distinti ricorsi in sede federale: il primo dinanzi al tribunale distrettuale del Northern District of California, il secondo dinanzi alla Corte d'Appello del District of Columbia, contestando sia la designazione quale "supply chain risk" sia l'ordine presidenziale emanato da Donald Trumo volto alla cessazione dei rapporti con tutte le agenzie federali.

Nel ricorso presentato in California, la società sostiene che l'azione del Dipartimento della Difesa si ponga in violazione dell'Administrative Procedure Act (APA), in quanto la designazione sarebbe stata adottata in assenza del rispetto delle garanzie procedurali previste, quali una valutazione trasparente del rischio, un'adeguata motivazione, la notifica al soggetto interessato e la possibilità per quest'ultimo di partecipare al procedimento. Anthropic deduce, inoltre, che il bando integri una forma di ritorsione

illegittima in risposta alle proprie posizioni in materia di sorveglianza di massa e sistemi d'arma autonomi, in violazione del Primo Emendamento della Costituzione statunitense. Secondo la prospettazione attorea, l'intervento pubblico sarebbe finalizzato a compromettere in modo significativo il valore economico della società, nonché a dissuadere altri operatori dal negoziare limiti all'utilizzo delle proprie tecnologie.

Quanto alla quantificazione del pregiudizio, la società fa riferimento a "centinaia di milioni di dollari" di ricavi a rischio nel breve periodo, evidenziando altresì come la qualificazione quale rischio per la supply chain possa determinare effetti reputazionali tali da indurre anche clienti privati a interrompere i rapporti commerciali.

Un primo riscontro in sede giurisdizionale è intervenuto il 26 marzo 2026, quando la giudice federale Rita Lin ha disposto la sospensione temporanea dell'efficacia del bando, rilevando come le misure adottate dal governo appaiano orientate a "cripple Anthropic" e a disincentivare il dibattito pubblico sulle criticità sollevate dall'azienda in ordine all'impiego militare dell'intelligenza artificiale. Nell'ordinanza, il giudice qualifica tali interventi come una "classica ritorsione di Primo Emendamento", osservando che, ove si fosse trattato di un ordinario contenzioso contrattuale, il Dipartimento della Difesa si sarebbe limitato a non rinnovare il contratto, senza attivare una sequenza di decisioni eccedente quanto ragionevolmente necessario per la tutela della sicurezza nazionale.

La decisione non definisce il merito della controversia, ma evidenzia come una parte della magistratura sia orientata a sottoporre a rigoroso scrutinio l'utilizzo — potenzialmente punitivo — di qualificazioni quali "supply chain risk", in particolare quando esse incidono su operatori nazionali per ragioni che intrecciano esigenze di sicurezza e forme di dissenso di natura politico-valoriale.

Etica aziendale, ritorsione regolatoria e separazione dei poteri

Il caso solleva una questione di fondo: se un'impresa possa legittimamente rifiutarsi di fornire servizi per determinate finalità — quali la sorveglianza di massa o lo sviluppo di sistemi d'arma autonomi — senza esporsi al rischio di essere destinataria di misure statali di carattere eccezionale, eccedenti la mera perdita del contratto.

La risposta che sembra emergere, quantomeno alla luce della reazione iniziale dell'amministrazione Donald Trump, appare intrinsecamente ambivalente: positiva sul piano formale — in quanto nessun operatore economico può essere coattivamente obbligato a concludere un contratto — ma negativa

sul piano sostanziale, laddove il rifiuto venga interpretato come una minaccia alla sicurezza nazionale o come un tentativo di imporre allo Stato un autonomo sistema di vincoli etico-normativi. In tale contesto si colloca il concetto di "ritorsione regolatoria", inteso quale utilizzo di strumenti giuridici predisposti per finalità generali — quali la tutela della supply chain da interferenze esterne, i poteri speciali (golden power) o misure di natura emergenziale — al fine di colpire soggetti che assumono determinate posizioni politiche o valoriali.

Qualora l'autorità giudiziaria dovesse accertare che la ratio effettiva del bando non risiede nella gestione di un rischio tecnico, bensì nella volontà di sanzionare una posizione critica in ordine all'impiego dell'intelligenza artificiale, ciò potrebbe determinare la formazione di un precedente di rilievo, idoneo a rafforzare la tutela di quegli operatori che intendano fissare proprie "red lines" nell'utilizzo delle tecnologie sviluppate.

Sul versante europeo, come ricostruito in dottrina, la giurisprudenza della Corte di giustizia — a partire dalle pronunce "Digital Rights Ireland"[70], "Tele2 Sverige"[71] e "Ligue des droits humains"[72] — ha progressivamente delineato un limite invalicabile rispetto a pratiche di sorveglianza di massa e conservazione generalizzata dei dati giustificate da esigenze di sicurezza, imponendo un rigoroso scrutinio di proporzionalità.

---

[70] Cfr. CGUE, 8 aprile 2014, Digital Rights Ireland Ltd, C-293/12 e C-594-12, dove la Corte ha dichiarato invalida la Direttiva 2006/24/CE sulla conservazione dei dati, che obbligava gli ISP a conservare per lunghi periodi i dati di traffico e localizzazione degli utenti. Quest'ultima obbligava gli operatori di telecomunicazioni a conservare i dati relativi al traffico e alla localizzazione degli utenti per un periodo compreso tra 6 mesi e 2 anni, al fine di combattere reati gravi, come il terrorismo e la criminalità organizzata.

[71] Cfr. CGUE, 21 dicembre 2016, Tele2 Sverige AB, C-203/15, caso in cui la società in questione aveva smesso di conservare i dati di localizzazione e traffico in seguito alla decisione Digital Rights Ireland. Ciò fece si che l'autorità svedese gli ordinasse di continuare a farlo, in conformità con la legge nazionale. Questo fu il motivo dell'impugnazione da parte di Tele2, la quale sosteneva che la conservazione richiesta dalla norma nazionale violasse i diritti fondamentali sanciti dalla Carta dei diritti fondamentali dell'UE.

[72] Cfr. CGUE, Ligue des droits humains c. Conseil des ministers, C-511/18, 2020 dove la Corte di Giustizia dell'Unione Europea (CGUE) ha dichiarato illegittima la sorveglianza di massa e la conservazione indiscriminata dei dati personali da parte delle autorità pubbliche, in violazione degli artt. 7 e 8 della Carta dei Diritti Fondamentali dell'UE e del Regolamento (UE) 2016/679 (GDPR). La Corte ha affermato che il trattamento dei dati per finalità di sicurezza deve rispettare il principio di proporzionalità e garantire basi giuridiche solide, limiti temporali stringenti e controlli indipendenti. La pronuncia ha avuto un impatto sulle Smart Cities, in particolare sull'uso della videosorveglianza avanzata e del riconoscimento facciale, escludendo il monitoraggio indiscriminato della popolazione. Tale orientamento ha influenzato la successiva regolamentazione dell'Artificial Intelligence Act, stabilendo criteri più stringenti per l'uso delle tecnologie di sorveglianza nelle amministrazioni pubbliche.

In tale prospettiva, il contenzioso che coinvolge Anthropic può essere interpretato come il tentativo — nel contesto statunitense — di trasporre una logica analoga nell'ambito dell'intelligenza artificiale: anche la sicurezza nazionale non può legittimare l'adozione di misure di carattere punitivo nei confronti di soggetti che sollevino criticità in ordine ai limiti giuridicamente ammissibili della sorveglianza algoritmica.

La dimensione concorrenziale: AI, antitrust e "campioni nazionali"

Sul piano concorrenziale, l'effetto immediato del bando consiste nel riorientamento di una quota significativa della domanda pubblica di intelligenza artificiale verso un numero ristretto di fornitori alternativi, in particolare OpenAI e Google.

Tale dinamica si inserisce in un contesto nel quale il Dipartimento di Giustizia statunitense è impegnato in un articolato contenzioso antitrust nei confronti di Google e di altri grandi operatori tecnologici, nell'ambito del quale si discute anche della possibilità di introdurre limitazioni agli investimenti in imprese concorrenti nel settore dell'intelligenza artificiale — tra cui Anthropic, partecipata dalla stessa Google — al fine di ridurre i conflitti di interesse e contenere fenomeni di concentrazione del potere economico.

Per Anthropic, l'eventuale imposizione di restrizioni all'afflusso di nuovi capitali da parte di Google, unitamente alla perdita accelerata di contratti pubblici e agli effetti reputazionali derivanti dalla designazione quale "supply chain risk", rischia di tradursi in una significativa compressione della capacità competitiva nel medio periodo. Più in generale, la combinazione tra un enforcement antitrust selettivo e l'adozione di misure di sicurezza nazionale dirette nei confronti di un singolo operatore può produrre effetti paradossali, finendo per rafforzare proprio quei "campioni nazionali" che la retorica antimonopolistica dichiara di voler contenere.

In tale prospettiva, emerge una questione di fondo per i regolatori — anche europei — relativa alla reale finalità delle politiche pubbliche: se, cioè, esse siano effettivamente orientate alla promozione di un ecosistema pluralistico e competitivo di fornitori di intelligenza artificiale, ovvero piuttosto al consolidamento di un numero limitato di grandi operatori strettamente allineati alle priorità strategiche dei governi.

Copyright, web scraping e potere contrattuale

Parallelamente, Anthropic è stata coinvolta in un ulteriore fronte di rilievo, relativo al rapporto tra l'addestramento dei modelli di intelligenza artificiale e la tutela del diritto d'autore. Nel 2025, la società ha concluso un accordo transattivo con un ampio gruppo di autori ed editori che avevano promosso un'azione giudiziaria per l'utilizzo non autorizzato di opere protette nell'addestramento dei modelli Claude; l'intesa presenta un valore potenziale fino a 1,5 miliardi di dollari, articolato in pagamenti dilazionati e royalty future.

La dottrina ha dedicato particolare attenzione al fenomeno del web scraping quale tecnica di raccolta dei dati funzionale all'addestramento dei modelli generativi, evidenziando come una quota significativa del traffico Internet globale sia ormai riconducibile a bot impiegati per la raccolta massiva di informazioni destinate ad alimentare sistemi di intelligenza artificiale, inclusi i large language model. In tale contesto, vengono frequentemente richiamati i principali dataset pubblici utilizzati — quali Common Crawl, LAION e archivi open — nonché le criticità sottese alla retorica dell'"open Internet", dietro la quale possono celarsi trattamenti illeciti di dati personali, in particolare quando contenuti provenienti da social network o servizi digitali sono riutilizzati in assenza di una base giuridica adeguata.

L'esperienza applicativa conferma tali criticità: il Garante per la protezione dei dati personali ha già adottato provvedimenti d'urgenza nei confronti di sistemi quali Replika[73], ChatGPT[74] e DeepSeek[75], sollevando rilievi in materia di trasparenza, basi giuridiche del trattamento, tutela dei minori e sicurezza dei dati, oltre a emanare linee guida specifiche sul web scraping e sull'intelligenza artificiale generativa, nelle quali si raccomanda l'adozione di misure tecniche e organizzative idonee a limitare la raccolta indiscriminata di dati.

In tale scenario, accordi transattivi di dimensioni miliardarie — come quello concluso da Anthropic con autori ed editori — evidenziano come soltanto operatori dotati di ingenti risorse finanziarie siano in grado di "internalizzare" il rischio giuridico, acquistando una sostanziale stabilità

[73] Cfr. GPDP, provvedimento del 2 febbraio 2023, n. 39 [doc. web n. 9852214]
[74] Cfr. GPDP, provvedimento del 30 marzo 2023, n. 112 [doc. web n. 9870832], nello specifico OpenAI avrebbe trattato dati personali per l'addestramento di ChatGPT senza aver previamente identificato una base giuridica conforme alle disposizioni del GDPR. Oltre a ciò, gli utenti non sarebbero stati adeguatamente informati circa le modalità e le finalità del trattamento dei loro dati personali e, infine, la piattaforma avrebbe consentito l'accesso a minori di 13 anni senza implementare sistemi efficaci per verificarne l'età, esponendoli a contenuti potenzialmente inappropriati.
[75] Cfr. GPDP, provvedimento del 30 gennaio 2025, n. 33 [doc. web n. 10098477].

legale su scala sistemica, mentre i soggetti di minori dimensioni rimangono esposti a rischi significativamente più elevati e potenzialmente sproporzionati.

AI Act, giurisprudenza europea e "zone grigie" sulla sicurezza nazionale

L'AI Act adotta un approccio fondato sul rischio, vietando talune pratiche qualificate come di "rischio inaccettabile" — quali il social scoring generalizzato, la manipolazione comportamentale aggressiva, l'estrazione indiscriminata di immagini ai fini del riconoscimento facciale e i sistemi di valutazione emotiva in contesti sensibili — e imponendo requisiti particolarmente stringenti per i sistemi classificati come "ad alto rischio".

In tale ultima categoria rientrano, tra gli altri, i sistemi destinati a essere impiegati in ambiti quali infrastrutture critiche, istruzione, occupazione, accesso a servizi essenziali, applicazione della legge, migrazione e gestione delle frontiere, per i quali sono previsti obblighi in materia di gestione del rischio, governance dei dati, trasparenza, supervisione umana e registrazione delle attività.

Con specifico riferimento alla sorveglianza biometrica, l'AI Act vieta l'utilizzo generalizzato in tempo reale nei luoghi pubblici a fini di law enforcement, ammettendo esclusivamente deroghe circoscritte — quali la ricerca di vittime di reati gravi, la prevenzione di minacce terroristiche o la cattura di sospetti di reati di particolare gravità — comunque subordinate ad autorizzazione giudiziaria e al rispetto dei principi di necessità e proporzionalità.

La giurisprudenza europea richiamata in dottrina — dalle pronunce Digital Rights Ireland e Tele2 Sverige fino a Ligue des droits humains — ha già chiarito che forme di sorveglianza e conservazione dei dati su larga scala, prive di criteri selettivi, risultano incompatibili con gli articoli 7 e 8 della Carta dei diritti fondamentali dell'Unione europea, nonché con i principi sanciti dal Regolamento generale sulla protezione dei dati.

Tuttavia, l'AI Act esclude espressamente dal proprio ambito di applicazione i sistemi di intelligenza artificiale utilizzati esclusivamente per finalità militari, di difesa o di sicurezza nazionale, rimettendo agli Stati membri la disciplina di tali settori[76].

Tale esclusione — prevista dall'articolo 2, paragrafo 3, e richiamata nel Considerando 24 — è stata oggetto di rilievi critici in dottrina e da parte della società civile, che l'hanno qualificata come un

[76] R. RAZZANTE (a cura di), *Manuale sull'intelligenza artificiale*, Giappichelli, Torino, 2024.

potenziale “vuoto normativo”, suscettibile di trasformare il comparto militare-industriale in un’area caratterizzata da una tutela attenuata dei diritti digitali.

È stato altresì evidenziato come, anche al di fuori dell’ambito strettamente militare, l’Unione europea sia già confrontata con fenomeni riconducibili alla c.d. “digital mass surveillance”[77], in particolare nel contesto delle smart cities, ove l’impiego di sistemi di videosorveglianza avanzata, sensori IoT e tecnologie di tracciamento della mobilità rischia di determinare una progressiva normalizzazione di modelli di controllo pervasivo.

Le pronunce della Corte di giustizia e delle giurisdizioni nazionali in materia di riconoscimento facciale, smart cities e conservazione dei dati confermano l’esistenza di una marcata resistenza, nell’ordinamento europeo, nei confronti di forme di sorveglianza indiscriminata, resistenza che appare destinata a costituire un parametro interpretativo imprescindibile anche nella fase di attuazione dell’AI Act.

Italia tra strategia AI, difesa e protezione dei dati

Sul versante civile, l’Italia si è dotata, nel 2024, di una Strategia nazionale per l’Intelligenza Artificiale 2024–2026, la quale individua specifiche priorità nei settori della pubblica amministrazione, della sanità, della mobilità, della giustizia, dell’industria e dell’istruzione, prevedendo altresì misure in materia di governance, formazione, ricerca e sperimentazione attraverso sandbox regolatorie.

Il documento evidenzia espressamente l’esigenza di assicurare l’allineamento dell’impiego dell’intelligenza artificiale agli standard europei di tutela dei diritti fondamentali, nonché di promuovere un ecosistema competitivo, valorizzando il ruolo delle PMI, delle start-up e delle forme di collaborazione tra settore pubblico e privato.

[77] L. E. R. VEGA, L. SCAFFARDI, I. SPIGNO (a cura di*), I diritti fondamentali nell’era della digital mass surveillance*, Editoriale Scientifica, Napoli, 2021, nello specifico questo termine è generalmente utilizzato per riferirsi a sistemi di raccolta e/o accesso su larga scala. Tuttavia, la terminologia impiegata per descrivere tali operazioni è estremamente variegata. «Intelligence can be collected with technical means and at large scale. This surveillance technique is referred to in different ways, including “signals intelligence”, tegic survillance, “bulk investigatory powers”, “mass digital surveillance” and 'storage of data on a generalised basis», FRA (EUROPEAN UNION AGENCY FOR FUNDAMENTAL RIGHTS), *Surveillance by intelligence services: fundamental rights safeguards and remedies in the EU Volume I: Member States' legal framework*, 2017.

Sul piano della difesa, il Ministero competente ha adottato nel 2026 la strategia "AI and Defence"[78], nella quale l'intelligenza artificiale è qualificata come elemento "indispensabile" per il futuro della sicurezza nazionale, prevedendone un'integrazione progressiva nei processi di addestramento, pianificazione, analisi delle minacce, logistica e cooperazione industriale.

Tra i principi guida emergono, in particolare, il mantenimento di un controllo umano significativo sulle decisioni operative, la responsabilità di comando, il rispetto del diritto internazionale umanitario e il perseguimento della c.d. "sovranità digitale"[79], da realizzarsi mediante il rafforzamento delle capacità interne di sviluppo, validazione e gestione dei sistemi di intelligenza artificiale.

La dottrina evidenzia come, in parallelo, l'ordinamento italiano abbia progressivamente consolidato — anche attraverso l'applicazione del Regolamento Generale sulla Protezione dei Dati (GDPR), del Codice in materia di protezione dei dati personali e della giurisprudenza — una concezione della protezione dei dati quale "bene comune"[80], funzionale non soltanto alla tutela individuale, ma anche al corretto funzionamento delle istituzioni democratiche nell'ecosistema digitale.

In tale prospettiva, anche l'impiego dell'intelligenza artificiale in ambito difensivo deve essere interpretato alla luce di un necessario bilanciamento tra esigenze di sicurezza nazionale e salvaguardia dei diritti fondamentali, evitando che il richiamo alla sicurezza si traduca in una compressione indiscriminata degli spazi di dissenso o di autodeterminazione tecnologica degli operatori.

[78] Cfr. AI and Defence, Ministero della Difesa, https://www.difesa.it/assets/allegati/90197/ia_e_difesa_2026.pdf. Il documento intitolato "*AI e Difesa: Strategia di Difesa sull'Intelligenza Artificiale*" - fortemente promosso dal Ministro della Difesa, Guido Crosetto – rappresenta un quadro globale e ambizioso che definisce una linea d'azione da attuare immediatamente e completare entro un lasso di tempo limitato, finalizzato all'integrazione sistemica e sistematica dell'IA nel settore della difesa.

[79] A. SIMONCINI (a cura di), *Sistema dele fonti e nuove tecnologie. Il ruolo elle autorità indipendenti*, Giappichelli, Torino, 2023

[80] F. MODUGNO, *I "nuovi diritti" nella giurisprudenza costituzionale*, Giappichelli, Torino, 1995. La natura multiforme della riservatezza è stata evidenziata, mettendo in luce il suo rapporto con le diverse modalità di esercizio del potere, anche da S. RODOTÀ, *La «privacy» tra individuo e collettività*, in Pol. dir., 1974, il quale osserva che in ognuna delle istituzioni in cui si svolge la personalità umana «variano gli equilibri tra interessi privati e pubblici», con conseguente variazione delle modalità di esercizio del potere; e «chi non si culla nell'attesa di palingenesi, allora, deve essere in grado di compiere analisi e proporre soluzioni che tengano conto di tutte quelle variabili».

Tre piste operative per l'Europa (e per l'Italia)

A partire dalla vicenda che ha coinvolto Anthropic e OpenAI, nonché dal quadro teorico ricostruito in dottrina, emergono alcune direttrici di riflessione operativa rilevanti per il contesto europeo e italiano.

In primo luogo, appare opportuno circoscrivere l'impiego estensivo degli strumenti di sicurezza economica nei confronti di operatori nazionali. Misure quali la designazione quale "supply chain risk" o l'esercizio dei poteri speciali (golden power) sono state concepite per fronteggiare minacce connesse a interferenze estere, spionaggio o controllo di asset strategici; tuttavia, il caso statunitense evidenzia come un loro utilizzo nei confronti di imprese domestiche, motivato da divergenze relative a clausole etiche o a scelte di policy aziendale, possa tradursi in una forma di ritorsione regolatoria, idonea a disincentivare l'adozione di meccanismi volontari di limitazione e responsabilizzazione nell'impiego dell'intelligenza artificiale.

In secondo luogo, si pone l'esigenza di integrare in modo sistematico requisiti di sicurezza, etica e protezione dei dati all'interno delle procedure di procurement pubblico. Piuttosto che demandare ai singoli operatori la definizione di clausole etiche — con il rischio di generare conflitti asimmetrici con le amministrazioni — i capitolati di gara potrebbero prevedere standard minimi in materia di sicurezza, trasparenza, supervisione umana, protezione dei dati personali e rispetto dei diritti fondamentali. Un simile approccio risulterebbe coerente sia con l'impostazione del Regolamento generale sulla protezione dei dati, che impone principi di liceità, minimizzazione, proporzionalità e accountability by design, sia con la logica risk-based dell'AI Act[81].

Infine, risulta centrale il rafforzamento dei meccanismi di controllo e di ricorso avverso decisioni esecutive in materia di intelligenza artificiale[82]. L'intervento della giudice Rita Lin evidenzia l'importanza di un sindacato giurisdizionale effettivo su provvedimenti che incidono in modo significativo sull'attività degli operatori del settore, specialmente quando si intrecciano valutazioni di sicurezza nazionale e profili attinenti alla libertà di espressione. Nel contesto europeo, il ruolo delle Corti costituzionali, dei giudici amministrativi, delle autorità indipendenti — quali il Garante per la protezione dei dati personali — nonché, a livello sovranazionale, della Corte di giustizia dell'Unione

[81] Regolamento (UE) 2024/1689 del Parlamento europeo e del Consiglio, 13 giugno 2024, *Artificial Intelligence Act*, art. 5 (*Prohibited AI Practices*), disponibile online: https://artificialintelligenceact.eu/article/5/.

[82] RAZZANTE R., *AI e tutela dei diritti fondamentali*, in Dirittifondamentali.it, 2024.

europea e delle autorità di protezione dei dati, potrebbe essere ulteriormente valorizzato al fine di evitare che il richiamo alla sicurezza nazionale si trasformi in una clausola generale idonea a comprimere o neutralizzare ogni forma di controllo giurisdizionale, come già evidenziato in dottrina e dalla società civile[83].

Bibliografia

- NOVELLI C., FLORIDI L., GAUR A., *Two futures of AI regulation under the Trump administration*, 2025.
- DI BELLO B., *Ai pro. Progettare, creare e comunicare con l'intelligenza artificiale generativa*, Hoepli, Milano, 2024.
- FLORA M., MAGGI L., *Intelligenza artificiale generativa. Opportunità e sfide legali*, Pacini Editore, Pisa, 2024.
- MODUGNO F., *I "nuovi diritti" nella giurisprudenza costituzionale*, Giappichelli, Torino, 1995.
- RAZZANTE R. (a cura di), Manuale sull'intelligenza artificiale, Giappichelli, Torino, 2024.
- RAZZANTE R., *AI e tutela dei diritti fondamentali*, in Dirittifondamentali.it, 2024.
- RUFFOLO U., AMIDEI A., *Diritto dell'Intelligenza Artificiale. Proprietà industriale e intellettuale. CorpTech. Giustizia Predittiva. Transumanesimo. AI generativa. Metaverso*, LUISS University Press, Roma, 2024.
- SIMONCINI A. (a cura di), *Sistema delle fonti e nuove tecnologie. Il ruolo delle autorità indipendenti*, Giappichelli, Torino, 2023.
- VEGA L. E. R., SCAFFARDI L., SPIGNO I. (a cura di), *I diritti fondamentali nell'era della digital mass surveillance*, Editoriale Scientifica, Napoli, 2021.

[83] Per un approfondimento si veda, RAZZANTE R., *AI e tutela dei diritti fondamentali*, in Dirittifondamentali.it, 2024.

NEXTWORK360

Nextwork360 - Boosting Tech Company Growth

Nextwork360 è il network di riferimento in Italia di testate e siti B2B dedicati ai temi della Trasformazione Digitale e dell'Innovazione Imprenditoriale. Aggrega la più grande community di CIO, Innovation Manager, C-Level, imprenditori tech e ha la missione di diffondere la cultura digitale e imprenditoriale nelle imprese e pubbliche amministrazioni italiane e favorirne l'incontro con i migliori fornitori tecnologici attraverso eventi, progetti di comunicazione, lead generation e servizi inbound.

info@nextwork360.it

Nextwork360 S.r.l.

Via Moretto da Brescia 22, Milano