AI: inevitabile o evitabile, questo (non) è il problema. Ciò che precede la trasparenza algoritmica[a]

Emanuela Tangari[*]

Abstract

L'articolo esplora la relazione tra Intelligenza Artificiale (IA) e fiducia, ponendo l'accento sulla trasparenza e la "trasparibilità" come elementi chiave per l'analisi di un utilizzo etico e responsabile, di cui il contesto medico si pone come caso d'uso privilegiato. Attraverso riferimenti a teorie filosofiche e psicologiche, si analizzano le sfide e le implicazioni delle decisioni autonome delle IA, mettendo in luce il loro impatto sul ragionamento umano. Viene preso in esame il progetto europeo MES-CoBraD per evidenziare i benefici e i limiti dell'applicazione dell'IA in medicina. Il tema centrale rimane la necessità di una trasparenza che superi la mera comprensione tecnica, per abbracciare una comprensione relazionale capace di sostenere una fiducia autentica e un utilizzo della ragione *tout court* nelle decisioni.

Parole chiave: Medicina e Tecnologie, Etica dell'Intelligenza Artificiale, Intelligenza Artificiale, Trasparenza Algoritmica, Fiducia e Affidabilità

Abstract

The article explores the relationship between Artificial Intelligence (AI) and trust, emphasizing transparency and "traceability" as key elements in analyzing the ethical and responsible use of AI, with the medical field serving as a prime use case. Drawing on philosophical and psychological theories, it examines the challenges and implications of AI-driven autonomous decisions, highlighting their impact on human reasoning. The European MES-CoBraD project is analyzed to showcase the benefits and limitations of AI applications in medicine. The central theme remains the necessity for transparency that goes beyond mere technical understanding, embracing

[a] This paper is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 965422. Saggio ricevuto in data 31/05/2024 e pubblicato in data 22/01/2025.

[*] Docente a contratto, Università di Roma "Tor Vergata", email: emanuela.angela.tangari@uniroma2.it.

a relational comprehension capable of fostering genuine trust and the application of reason in decision-making.

Keywords: Medicine and Technologies, Ethics of Artificial Intelligence, Artificial Intelligence, Algorithmic Transparency, Trust and Reliability

1. Introduzione

Nel 2013 veniva pubblicato un articolo dal titolo "The Inevitable Application of Big Data to Health Care"1; a distanza di quasi 10 anni, nel 2022 viene pubblicato "Artificial Intelligence for Health and Care Is Not Inevitable: Introduction and Critical Vocabulary"2. Indipendentemente dalla divergenza di prospettive e dalla posizione – e dalle teorie a supporto – che si vuole assumere, una cosa ci appare certamente inevitabile: la domanda aperta sull'Intelligenza Artificiale e sulle sue implicazioni, oggi il tema forse più presente nel dibattito filosofico, ingegneristico, matematico, sociale, giuridico.

Tra i campi d'applicazione più discussi – in cui da sempre la cooperazione tra esseri umani e tecnica è al centro – c'è senza dubbio quello della medicina. Non vogliamo qui indagare l'inevitabilità dell'utilizzo della AI in medicina; piuttosto, una premessa nota ma non così ovvia può essere quella di domandarsi che cosa possa significare non tanto una "intelligenza artificiale", ma piuttosto a quali condizioni le pratiche (mediche e non) sostenute dai sistemi di Intelligenza Artificiale o dai sistemi basati sugli algoritmi, o le norme, le decisioni – al di là che esse siano automatizzate o no – possano definirsi intelligenti, e soprattutto a partire da quali criteri e quali principi. Su tali principi è necessaria un'analisi che sposti l'attenzione da un piano formale e normativo a un piano fondamentale, cognitivo, che parta (e ritorni) alla condizione umana nel suo accadere.

Uno degli elementi caratteristici di tale condizione umana nel suo accadere è l'essere in relazione: il fatto che gli esseri umani esistono in un contesto di reciprocità, in un contesto esistenzialmente e socialmente correlato, e in una situazione storico-sociale in cui l'individualità sorge e si sviluppa. A ciò si aggiunge il fatto che il contesto storico-sociale in questione è un contesto permeato dalla tecnica-tecnologia; è quindi un contesto storico-tecno-sociale, in cui la tecnica non ricopre più solo il ruolo di artefatto, ma acquista lo statuto – almeno fenomenologico, anche se non si voglia considerare quello assiologico – di agente tra agenti, e non solo di un prodotto. E così perveniamo al fattore forse più rilevante della riflessione: a comporre il sostrato che regola il vivere comune, le relazioni, vi sono numerose dinamiche. Prendiamo qui in

[1] T.B. Murdoch, A. S. Detsky, *The inevitable application of big data to health care*, in «JAMA», 309, n. 13, 2013, pp. 1351-1352. https://doi.org/10.1001/jama.2013.393.

[2] R. Walker, *Artificial Intelligence for Health and Care Is Not Inevitable: Introduction and Critical Vocabulary,* in J. Dillard-Wright, J. Hopkins-Walsh, B. Brown (a cura di), *Nursing a Radical Imagination Moving from Theory and History to Ac-tion and Alternate Futures,* Routledge, London 2022.

esame una di queste, che si pone come centrale in riferimento al contesto storico-tecno-sociale.

La questione della fiducia e dell'affidabilità diventa uno snodo essenziale nel dibattito sulla tecnologia3, pur non trovando in questo dibattito il suo inizio; prende invece in causa problemi morali, virtù, predisposizioni, dinamiche sociali e intersoggettive che vengono molto prima delle loro applicazioni. Su che cosa si fonda la fiducia? quando possiamo dire che qualcosa o qualcuno è affidabile? e che rapporto c'è tra la categoria di affidabilità e l'esperienza – fenomenologicamente intesa – della fiducia? Non basta che un soggetto o un agente sia ritenuto (da altri o da noi stessi) affidabile affinché si generi l'*esperienza* della fiducia. Questo mostra l'intricato rapporto tra i due termini e apre ad un'altra dimensione fondamentale, quella dell'intenzionalità, delle volontà, dei fini. Nel campo dell'ingegneria tecnologica, sono proprio tali intenzioni – e dunque la costruzione e la strutturazione dei modelli, dei software, degli strumenti – a coinvolgere, a discesa, quegli elementi che vengono poi messi al centro della discussione: la trasparenza, la responsabilità, l'equità, il diritto alla privacy, l'inclusione, l'imparzialità, e le numerose altre sfere implicate.

La trasparenza nell'uso delle tecnologie digitali, specialmente in ambito medico, è fondamentale non solo per la fiducia degli utenti ma anche per garantire una partecipazione informata e consapevole da parte degli individui coinvolti: non si tratta dunque di un mero requisito tecnico ma anche un principio etico che deve guidare lo sviluppo e l'implementazione delle nuove tecnologie. La trasparenza, nel contesto delle tecnologie digitali, si riferisce alla disponibilità di informazioni sui processi interni di un sistema. Un sistema trasparente permette agli utenti di capire come le decisioni vengono prese, quali dati vengono utilizzati e quali criteri vengono applicati. La trasparenza è essenziale per garantire che i sistemi siano utilizzati in modo equo e responsabile. La trasparibilità – cui si farà cenno come fattore distinto dalla trasparenza e maggiormente in riferimento al funzionamento del ragionamento umano – si riferisce alla capacità di un sistema di rendere visibili i suoi processi interni e i suoi risultati. Questo include la possibilità di tracciare le fonti dei dati, comprendere le logiche di funzionamento degli algoritmi o dei processi psichici, e accedere ai dati che spieghino le decisioni. Nel contesto delle AI – ma anche del ragionamento umano –, della loro verifica e validazione, la trasparibilità diventa cruciale per descrivere il tema della fiducia. Nel campo delle tecnologie digitali, tale fiducia è strettamente legata alla trasparenza e alla spiegabilità, vale a dire alla possibilità di interagire con sistemi che operino in modo prevedibile, sicuro, comprensibile, responsabile, socialmente accettabile.

2. L' εὐδαιμονία aristotelica e l'euristica della scelta

Un passo indietro (o di lato): si è iniziato chiedendo quando la medicina, o un altro dominio, può definirsi "intelligente". La domanda può qui riguardare una certa

[3] O. O'Neill, *Trust and Accountability in a Digital Age*, in «Philosophy», 95, n. 1, 2020, pp. 3-17.

declinazione dell'intelligenza, cioè la razionalità: quando, cioè, una decisione può definirsi razionale. La storia della filosofia è costellata da tale problema, dalla filosofia antica a quella contemporanea, passando per la filosofia moderna e l'analisi della ragione, da Cartesio a Kant, senza conoscere la fine del problema. Già Aristotele nell'*Etica Nicomachea* descriveva la razionalità – la razionalità "pratica" – come mezzo, mediazione, per conseguire uno scopo o per soddisfare un desiderio. Prendiamo in prestito proprio la prospettiva aristotelica, particolarmente utile ad evidenziare quanto l'accezione che si attribuisce ad un termine o alla sua realizzazione sia significativa nella costruzione del giudizio "morale".

Sono necessarie due precisazioni: 1. qual è, per Aristotele, il fine ultimo di ogni azione. Questa domanda appare secondaria nella riflessione attuale sull'Intelligenza Artificiale: si trova dunque un ampio dibattito su che cosa sia quest'ultima e quali siano le sue implicazioni e i suoi impatti, ma la questione si allontana sempre di più dalla domanda originaria su quale sia l'"intelligenza" (cioè la ragione, il fine) che guida l'azione intelligente. 2. Il sillogismo che descrive l'azione razionale non può prescindere dall'elemento onnipresente in ogni movimento umano: quello della libertà, della volontà, o anche solo del libero arbitrio. Libertà di che cosa? verso dove si volge (dovrebbe volgere) la deliberazione? Verso la felicità, εὐδαιμονία. Questa εὐδαιμονία è l'accordo con l'ἀρετή, l'eccellenza: proprio dell'umano è quindi l'esercizio delle virtù, secondo Aristotele – etiche e dianoetiche –, l'esercizio razionale accordato all'eccellenza, e volto al bene. Tralasciando qui la riflessione sull'interpretazione descrittiva o normativa del modello deliberativo, è utile considerare, dopo aver sottolineato l'importanza del tema dei fini, un ulteriore aspetto nel discorso della razionalità, e quindi dell'intelligenza, sia essa umana o "artificiale": l'atteggiamento, il comportamento, la predisposizione specificamente umana nell'esercizio della razionalità. Se le azioni umane intelligenti o razionali seguissero un fine preciso e identificato (qualsiasi esso sia, e a maggior ragione se esso fosse il "bene"), non sarebbe possibile spiegare la gamma di contraddizioni, dissidi, controsensi che le scelte e le decisioni umane comportano.

Gli studi sulla "razionalità limitata" (bounded rationality) e sui meccanismi cognitivi che sono alla base delle decisioni umane e dei loro biases4 contraddicono l'idea di una razionalità logicamente intesa, ed evidenziano invece la potenza di una euristica irrazionale, tipicamente umana, che sottende l'agire individuale e collettivo. Secondo Daniel Kahneman5, meccanismi cognitivi come l'ancoraggio, la rappresentatività, la disponibilità, sono i motori dei giudizi ritenuti razionali, ma che poco hanno a che fare con il "sistema 2", quello logico, deliberativo, probabilistico, e molto di più ineriscono al "sistema 1", il pensiero "veloce", intuitivo. La stessa capacità computazionale e statistica della mente umana è soggetta necessariamente alla rappresentazione che una parte dell'io – o del sé – le restituisce: questo è il motivo per cui la memoria non raccoglie informazioni secondo la loro quantità, ma secondo la loro potenza: il

[4] A. Tversky, D. Kahneman, *Judgment under Uncertainty: Heuristics and Biases*, in «Science», 185, n. 4157, 1974, pp. 1124-1131.

[5] D. Kahneman, *Thinking, Fast and Slow*, Penguin Books, London 2012.

ricordo di un singolo evento può essere assai più vivido – e quindi più determinante e statisticamente più probabile in una futura scelta – di una somma di informazioni ripetute, ma ritenute meno incisive.

In chiave più analitica, Donald Davidson nel 1986 si dedicava alla delineazione del rapporto tra il fondamento dell'interpretazione e quello della conoscenza, affermando che la «conoscenza empirica […] nasce, piuttosto, dalla natura dell'interpretazione. Come interpreti dobbiamo trattare l'attribuzione autonoma di credenze, dubbi, desideri e preferenze come privilegiata; questo è un passo essenziale nell'interpretare il resto di ciò che una persona pensa e dice. La fondazione dell'interpretazione non è la fondazione della conoscenza, anche se una comprensione della natura dell'interpretazione può portare alla comprensione della natura essenzialmente veridica delle credenze»[6]; a distanza di anni, nel 2001, uno dei suoi lavori vede emergere un contrasto con la razionalità pratica basata sulla decision theory: «la teoria della decisione corrisponde alle nostre intuizioni su come le reali decisioni vengono prese, ed è parte del nostro apparato di senso comune per spiegare il comportamento intenzionale»7; ciò si fonda sul presupposto che le azioni che gli esseri umani compiono in maniera intenzionale sono guidate da credenze e desideri, che conferiscono valore all'azione stessa.

3. L'impatto della AI sul ragionare umano

Alla luce di questi brevi cenni su alcune delle prospettive concernenti la teoria della scelta e il funzionamento della razionalità umana, si può tornare a chiedersi che cosa accade quando l'agire (o il ragionare) umano può servirsi di una "intelligenza", di un ragionare – se così può dirsi – artificiale, che non possiede intenzionalità alcuna. Il problema sembra situarsi non solo nell'ipotetica decisione che un sistema decisionale autonomo può prendere, ma ancor prima – e in maniera ancora più determinante – nel modo in cui il ragionare umano può essere inficiato, alterato dalle decisioni autonome. Non si tratterebbe allora di un parallelismo, classicamente inteso, tra la decisione umana e quella artificiale, ma di un circuito in cui il sistema decisionale autonomo interferisce, "compromette" non solo e non necessariamente la singola decisione, ma il modo stesso di ragionare.

Qui appare in maniera predominante il tema della trasparenza, di seguito discusso, che non intendiamo allora solo nei riguardi delle singole decisioni o dello

[6] D. Davidson, *A coherence theory of truth and knowledge*, in E. LePore (ed), *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*, Blackwell, Cambridge 1986, p. 332: «[…] empirical knowledge […] springs, rather, from the nature of interpretation. As interpreters we have to treat self-ascriptions of belief, doubt, desire and the like as privileged; this is an essential step in interpreting the rest of what the person says and thinks. The foundations of interpretation are not the foundations of knowledge, though an appreciation of the nature of interpretation can lead to an appreciation of the essentially veridical nature of belief».

[7] D. Davidson, *Subjective, Intersubjective, Objective*, Oxford University Press, Oxford 2001, p. 126. https://doi.org/10.1093/0198237537.001.0001.

specifico processo del particolare strumento, ma nei riguardi dell'interazione tra le due intelligenze, tra le due razionalità: quella umana e quella artificiale.

Secondo Kahneman la stessa razionalità umana è assai complessa da descrivere; le scelte apparentemente razionali sono sovente deliberate dal sistema veloce, quello dell'intuizione, del ragionamento "pratico", che si basa o può basarsi su elementi anche ancestrali, inconsapevoli, non risolvibili da una logica linguisticamente descrivibile. Se è così complicato decifrare la natura del ragionamento umano, che ne è di quello (programmato – almeno finora – da individui umani, ma con una crescente dose di "autonomia") artificiale? Individuare, e tanto più descrivere, la *ratio* dei sistemi decisionali autonomi sembra impresa ardua, quanto più questi sistemi sono stratificati e complessi, non solo per la quantità di dati ma soprattutto per la loro tipologia e per i fattori discriminanti che dettano la scelta.

Fortemente interessante sarebbe indagare, da un punto di vista filosofico e psicologico, il modo in cui i sistemi decisionali artificiali compromettono la decisione umana, o meglio il modo in cui essi interagiscono col sistema decisionale umano; in che modo, quindi, la relazione umani-macchine modifica non tanto il risultato finale, ma il processo stesso in cui le decisioni umane procedono. Sembra quindi configurarsi, nell'universo del "Sistema 1 e 2" di Kahneman, un terzo sistema, che non si configura come sintesi o una negazione dei primi due e la cui rilevanza sta invece nel fatto di intercettare, modificare, penetrare intrinsecamente il funzionamento stesso dei primi due sistemi. Per questo il tema dell'Intelligenza Artificiale – in qualsiasi dominio di applicazione la si osservi – diventa decisivo, poiché non si pone al livello dei risultati o degli esiti, ma della strutturazione stessa del ragionamento, come anche della percezione umana, e della relazione degli esseri umani con il mondo, con gli altri, e con sé stessi. Sarebbe interessante per esempio osservare in che modo le AI mediche capaci di generare diagnosi o di proporre protocolli di cure si interpongano nel giudizio (presente e futuro) dei medici; in che modo cioè vadano a modificare non solo la decisione singola che di volta in volta viene presa, ma l'autocoscienza e l'abilità stessa dei professionisti, il rapporto tra le conoscenze acquisite e la capacità di decisione, l'esperienza di osservazione dei casi clinici, la strutturazione delle competenze.

Tralasciamo qui le questioni, note quanto importanti, del dibattito, che comprendono la considerazione della responsabilità8, dei pregiudizi, del controllo finale sui dati, o dell'equità delle scelte effettuate dai sistemi di Intelligenza Artificiale; di quanto, inoltre, siano realmente autonomi i sistemi di AI. Sulla scia di quanto accennato, soffermiamoci invece sull'aspetto della *trasparenza* e della spiegabilità delle decisioni, degli algoritmi che guidano le decisioni.

[8] K. Baum, S. Mantel, E. Schmidt, T. Speith, *From Responsibility to Reason-Giving Explainable Artificial Intelligence*, in «Philosophy & Technology», 35, n. 12, 2022. https://doi.org/10.1007/s13347-022-00510-w.

4. Trasparenza e "trasparibilità": una visione e non solo una ragione

La trasparenza è un tema centrale della discussione sulla creazione e sul funzionamento dei sistemi di AI9, specialmente di quelli decisionali, che possono per loro "natura" produrre discriminazioni e acuire divari già esistenti. La considerazione dei bias e degli algoritmi *black box* ripropone continuamente la necessità di lavorare ad una trasparenza che lasci intendere all'utente i criteri attraverso i quali il sistema giunge alla scelta10, e che tali criteri pongano alcuni fini o diritti umani – come l'equità, la "non-maleficenza", la responsabilità – come priorità del sistema stesso.

Di nuovo, emerge anche qui la difficoltà di rintracciare una definizione univoca e universale di "equità", per esempio, o di livello di spiegabilità che il sistema deve prevedere, o i destinatari per i quali esso deve risultare spiegabile. Si rinviene anche qui la difficoltà di pervenire ad un significato condiviso (e quindi applicabile) di trasparenza algoritmica, perché questo prevede in primo luogo la definizione di quali siano gli scopi e i soggetti coinvolti, e contemporaneamente dei criteri convenzionalmente condivisi sia per il processo di lettura delle previsioni/decisioni, sia soprattutto per un accordo, se possibile, su che cosa significhi l'interpretabilità. Se questa cioè ha a che fare con la possibilità di descrivere il processo di scelta e selezione, a partire dagli elementi forniti; e se tale descrizione dovrebbe garantire la comprensione, per esempio. Sorge allora la domanda su che cosa significhi comprendere, e ancor meglio su che cosa significhi comprendere per un essere umano (o per un gruppo, una società di individui). La comprensione, come la fiducia, non può essere facilmente disgiunta da fattori personali e spesso inconsci; tuttavia è proprio da questa comprensione – o comprensibilità – che prendiamo delle decisioni, che direzioniamo la scelta; la comprensione, come la fiducia che si accorda a qualcosa o qualcuno, sembra possedere i tratti di una intuizione, o di una "visione", propria degli individui umani, che non sempre e non facilmente è possibile descrivere con una logica algoritmica, matematica.

Si tratta di una conoscenza per così dire "morale", che non riguarda la conoscenza dei dati e delle informazioni ma di quelle intenzioni e fini cui lo strumento tende e da cui dipende. La trasparenza, dunque, come la fiducia, implica una relazione e non semplicemente una nozione, una disposizione di informazioni. Ciò è implicito nel concetto stesso di conoscenza, e dunque di comprensione e di spiegabilità: non basta che qualcosa sia spiegabile, o spiegato, affinché sia compreso. La conoscenza richiede anch'essa un fatto – il fatto di un rapporto che nasce tra il soggetto che conosce e l'oggetto conosciuto, nella sua natura e nel suo fine più profondo — e non semplicemente un dato (la potenziale spiegabilità o conoscibilità dell'oggetto).

---

[9] Non si porrà qui una distinzione particolare tra Machine Learning e Deep Learning; l'attenzione è invece su questioni di fondamento, su dilemmi e sfide per una riconfigurazione non solo tecnologica, ma anzitutto soggettiva, che ha a che fare con la professionalità e con il compito soggettivi.

10 W.J. von Eschenbach, *Transparency and the Black Box Problem: Why We Do Not Trust AI*, in «Philosophy & Technology», 34, 2021, pp. 1607-1622. https://doi.org/10.1007/s13347-021-00477-0.

Il tema problematico della trasparenza, e della trasparenza tecnologica, fonda su questo terreno le sue radici. La prima domanda, ancora una volta, è che cosa si intende per trasparenza. Quando nel processo conoscitivo e comprensivo umano qualcosa si ritiene trasparente; se sia una questione di accessibilità, di presa di possesso dei dati che costituiscono il contenuto dell'oggetto (sia esso una decisione, un algoritmo, un processo), o piuttosto di intelligibilità tecnica di questi. Forse entrambi questi fattori – insieme ad altri – contribuiscono alla percezione della trasparenza, fermo restando che 1. così come per il binomio fiducia-affidabilità, non è sufficiente che qualcosa sia *de facto* trasparente da un punto di vista tecnico perché sia anche comprensibile; 2. tale intelligibilità diviene problematica nel momento in cui l'elaborazione dei dati avviene attraverso l'apprendimento automatico, e non solo come esito della programmazione operata dall'essere umano: da qui l'idea di opacità o di inaccessibilità delle black-box11.

5. Un caso studio europeo: MES-CoBraD

Tra i numerosi casi di studio e di applicazione dell'AI in campo medico in cui le suddette questioni trovano spazio e necessità di essere attenzionate[12]. prendiamo in esame quello condotto nel progetto europeo MES-CoBraD Multidisciplinary Expert System for the Assessment & Management of Complex Brain Disorders), un interessante caso di Deep Learning utilizzato per la classificazione delle immagini (ad esempio espressioni geniche, scansioni MRI, ecc.). Coordinato dalla National Technical University of Athens, il progetto ha l'obiettivo di migliorare la precisione diagnostica e i risultati terapeutici nelle persone affette da disturbi cerebrali come i disturbi neurocognitivi (demenza), i disturbi del sonno e le crisi epilettiche (epilessia), nonché le loro interconnessioni.

L'Intelligenza Artificiale può porsi come strumento per migliorare e potenziare le funzioni cognitive umane dei medici che trattano pazienti affetti da malattie complesse13. Questi progressi richiedono un'analisi approfondita del modo in cui l'adozione di tali tecnologie sta influenzando i concetti di malattia, cure mediche, pratica clinica e la relazione di cura all'interno delle società. Di particolare rilievo sono i nuovi approcci basati sulle Reti Generative Avversarie14. che mirano ad ampliare il pool di dati medici disponibili per l'addestramento dei sistemi di apprendimento automatico capaci di riconoscere specifici tipi di cancro tramite immagini. Queste tecniche, note

---

[11] J. Burrell, *How the machine 'thinks': understanding opacity in machine learning algorithms*, in «Big Data & Society», 2016, pp. 1-12. https://dx.doi.org/10.2139/ssrn.2660674.

[12] J. Hatherley, *Limits of trust in medical AI*, in «Journal of Medical Ethics», 46, n.7, 2020, pp. 478-481. https://doi.org/10.1136/medethics-2019-105935.

[13] S.A. Bini, *Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: what do these terms mean and how will they impact health care?*, in «The Journal of Arthroplasty», 33, n. 8, 2018, pp 2358-2361. https://doi.org/10.1016/j.arth.2018.02.067.

[14] F.H.K. do Santos Tanaka, C. Aranha, *Data augmentation using GANs*, in «arXiv», 2019. https://doi.org/10.48550/arXiv.1904.09135.

come Data Augmentation15. presentano sfide significative dal punto di vista normativo, poiché i dati utilizzati per l'addestramento dei sistemi di Intelligenza Artificiale sono generati sinteticamente da altri sistemi di Intelligenza Artificiale.

La piattaforma MES-CoBraD è pensata per essere fruibile dalla più ampia comunità di utenti, ed essere conforme agli standard medici e sociali, legali ed etici suggeriti dalla Commissione Europea. Tra gli obiettivi del progetto vi è proprio la promozione del benessere e l'attenzione (e prevenzione) ai potenziali rischi, che comprende anche la condivisione dei dati reali (RWD) in una prospettiva di trasparenza e affidabilità, perseguite con la riduzione dei bias nei processi decisionali e garantendo che la responsabilità rimanga centrata sull'essere umano e sulle strutture sanitarie coinvolte, e incoraggiando i fruitori a seguire i codici internazionali di etica medica e delle pratiche per l'Intelligenza Artificiale.

Tra le altre ambizioni del progetto, vi è quella di supportare lo sviluppo e l'uso dell'AI per garantire che tutti possano essere curati e assistiti in strutture sanitarie che facciano uso di sistemi di Intelligenza Artificiale "sani", anche al di fuori del campo dell'imaging sanitario, attraverso un approccio "human-in-the-loop", attraverso il quale i professionisti medici e i pazienti possano essere coinvolti in un processo in cui la soluzione sia regolata, addestrata e testata su misura.

6. Trasparenza razionale, trasparenza artificiale

È doveroso chiedersi quale sia la distinzione tra la trasparenza (o meglio trasparibilità) artificiale e quella umana nei processi decisionali16. Secondo l'*argomento di uguale opacità*17 18, non dovrebbe sussistere una differenza radicale tra il modo in cui vengono giustificate le decisioni umane, in cui – data l'impossibilità di descrivere ogni passaggio della scelta umana – vi è una razionalizzazione a posteriori, e quello utilizzato dai sistemi ADM (Automated Decision-Making). Non si intende qui entrare nel merito di una critica dell'argomento di eguale opacità19 – molto interessante, perché pone tra gli altri il problema dell'autoregolazione e della modellazione del pensiero umano – ma si vogliono indicare i problemi di fondo e le implicazioni, e suggerirne una chiave di lettura.

Come accennato, l'opacità di alcuni sistemi di Intelligenza Artificiale è attribuita all'impossibilità di controllare o di seguire passo dopo passo il codice che produce un certo risultato, generando una sorta di "irriducibilità strutturale" di

---

[15] C. Shorten, T.M. Khoshgoftaar, *A survey on image data augmentation for deep learning*, in «Journal of big data», 6, n. 1, 2019, pp. 1-48. https://doi.org/10.1186/s40537-019-0197-0.

[16] J. Zerilli, A. Knott, J. Maclaurin, C. Gavaghan, *Transparency in algorithmic and human decision-making: is there a double standard?* in «Philosophy & Technology», 32, n. 4, 2019, pp. 661-683. https://doi.org/10.1007/s13347-018-0330-6.

[17] C. Buckner, *Black boxes or unfattering mirrors? Comparative bias in the science of machine behaviour*, in «British Journal for the Philosophy of Science», 74, n. 3, 2023, pp. 681-712.

[18] Burrell, *How the machine 'thinks'*, art. cit.

[19] U. Peters, *Explainable AI lacks regulative reasons: why AI and human decision-making are not equally opaque*, in «AI Ethics», 3, 2023, pp. 963-974. https://doi.org/10.1007/s43681-022-00217-w.

quest'ultimo e del suo processo20. Questo è direttamente proporzionale alla stratificazione e alla complessità del codice e della computazione algoritmica chiamata in causa nel sistema dell'AI e nel suo scopo, e spesso direttamente proporzionale anche alla sua efficienza. In questi casi diviene difficile individuare la correlazione tra i dati, le informazioni, e i risultati proposti, malgrado la natura intrinsecamente intellegibile di quei dati. Si dirà che lo stesso – su un altro livello – accade con la computazione umana, con i processi e con le scelte umane: abbiamo infatti notato come la razionalità o l'apparato decisionale non procede in modo lineare, logico; i movimenti psichici e della ragione si servono di dinamiche spesso inaccessibili alla coscienza o alla memoria umana, e tuttavia si dimostrano nella maggior parte dei casi *efficaci*.

D'altro canto, contrastare l'opacità del sistema decisionale (sia esso artificiale o soggettivo) necessita di considerare l'elemento dell'intenzionalità, tipico degli esseri umani, a cui inerisce la fiducia e la comprensibilità[21].

Valutare l'accuratezza delle previsioni o delle decisioni tra l'uno e l'altro sistema appare estremamente complesso: da un lato vi è la scarsa trasparenza dei sistemi decisionali autonomi per le ragioni già esposte; d'altro canto, però, le decisioni prodotte dall'Intelligenza Artificiale non possiedono il carattere di "autosabotaggio" o di condizionamento che la mente umana è capace di produrre e spesso (se non sempre) produce rispetto alle proprie scelte, inficiate da fattori psicologici, storici, esperienziali; l'opacità si colloca quindi negli esseri umani, e nel processo decisionale della mente, su un altro livello, ma è tuttavia presente e non facilmente calcolabile o prevedibile, data l'estrema complessità dei meccanismi psichici. A ciò si aggiunge un altro fattore discriminante tra i due sistemi: malgrado la possibile e frequente incoerenza tra la decisione umana, le sue spiegazioni, e i dati, gli elementi reali da cui la decisione è scaturita, il fatto stesso che i soggetti giustifichino una specifica decisione delineandone le ragioni produce una sorta di coerenza *post-hoc*, per la quale gli individui sono portati ad agire in linea con le spiegazioni fornite.

L'impegno alla coerenza – alla stregua dell'"effetto alone" descritto da Kahneman, per il quale la mente tende ad ignorare le informazioni che non sono allineate con la storia o con la percezione costruita – non può evidentemente modificare l'origine delle cause da cui le decisioni iniziali provengono, ma può certamente intervenire in questo processo e alterare l'automatismo di quelle cause. Da ciò deriverebbe, in ultima analisi, una maggiore trasparenza fondata su un movimento contemporaneamente passato e futuro in cui le intenzioni, le autoaffermazioni e le convinzioni (anche a posteriori) che gli individui umani attribuiscono alle scelte incidono concretamente sulla stabilità e sulla coerenza futura della decisione, e dunque della sua prevedibilità22.

---

20 P. Zellini, *La dittatura del calcolo*, Adelphi, Milano 2018.

21 A. Ferrario, M. Loi, E. Viganò, *Trust does not need to be human: it is possible to trust medical AI*, in «Journal of Medical Ethics», 47, 2021, pp. 437-438. https://doi.org/10.1136/medethics-2020-106922.

22 L. De Bruin, D. Strijbos, *Does confabulation pose a threat to first-person authority? Mindshaping, self-regulation and the importance of self-know-how*, in «Topoi», 39, 2020, pp. 151-161. https://doi.org/10.1007/s11245-019-09631-y.

7. Razionalità o ragionevolezza: un divario dialogico

Ancora, ma su un altro piano, è possibile mostrare che – per quanto opaca – la ragione e l'intuizione umana possiedono una sofisticatezza forse non sufficientemente spiegabile, ma ancora ragionevole e affidabile.

L'elemento problematico sottostante a queste complesse considerazioni e valutazioni sembra essere sempre il tipo di attribuzione di significato, di definizione condivisa o di valore che si associa agli elementi e ai termini presi in esame. È possibile, per esempio, paragonare la "spiegazione" fornita o ricavabile da un sistema di AI a quella che si può *esigere* da un essere umano? Se non ci si sofferma qui sulla richiesta – sull'aspettativa, strettamente umana in questo caso – che si ha nei confronti dell'uno e dell'altro sistema, qualsiasi comparazione risulta peregrina. Se la costruzione di un sistema di AI può poggiare su strutture designate a rispettare alcuni elementi essenziali (*ethics by desing*, etc.) che a loro volta rendono possibile una sorta di comparazione dei risultati, è altrettanto vero, o reale, che ad oggi le aspettative e le relazioni che le persone hanno con le Intelligenze Artificiali non sono (ancora?) le stessa di quelle che richiedono agli altri esseri umani. Questo si manifesta in termini di responsabilità e sensibilità che ci si aspetta. Rispetto agli esseri umani, non ci si limita a cercare razionalità o spiegabilità, ma anche una una *ragionevolezza* da intendersi come una comprensione profonda e completa dell'oggetto o soggetto coinvolto. Tale ragionevolezza è una caratteristica tipica delle interazioni umane, in quanto coinvolge proprietà sia personali che interpersonali, che emergono nella relazione tra individui.

Non lo sono, inoltre, dal punto di vista dell'interazione stessa: il rapporto che gli esseri umani intrattengono – e il campo medico ne è o ne dovrebbe essere un campo esemplificativo – è un rapporto dialogico, relazionale. Il rapporto dialogico richiede la possibilità di porre domande e ricevere risposte; ma, soprattutto, di porre *reciproche* domande.

Al di là, dunque, del funzionamento del sistema mentale in confronto a quello della AI, vi è dunque il funzionamento della relazione umana in confronto a quella che si può stabilire con i sistemi di AI; questi ultimi non domandano: o meglio, non domandano in modo *creativo*. Non vi è per l'AI la possibilità di porre domande affettive (se non per simulazione), e dunque *affettivamente intuitive* – quali sono quelle sullo stato emotivo e sul modo in cui ciascuno significa la propria esperienza – che nascono cioè dalla condizione stessa della relazione e sono impossibili al di fuori di essa.

Con questo arriviamo ad una maggiore delineazione della suddetta questione della ragionevolezza come capacità di osservazione della complessità dei fattori. Tale ragionevolezza, a differenza della razionalità, comprende quell'alveo di affetto (nel senso latino dell'*affectus*, dell'essere *affetti-da*) che genera una imprevedibile capacità di comprensione, impossibile alla sola razionalità. Da qui una abilità a porre domande che non derivano dalla sola associazione di dati e informazioni, ma da una esperienza che trova nella relazione stessa una sempre nuova vitalità, una ri-generazione continua

che si fonda sulla ragione affettiva, e che quindi risulta impossibile all'Intelligenza Artificiale la quale, per ora, può solo simulare.

È proprio la ragionevolezza a segnare anche la distinzione tra la comprensibilità, l'accettabilità delle decisioni che derivano da sistemi di AI e quelle prodotte dagli esseri umani, o tra un essere umano ed un altro; così come per il binomio affidabilità/fiducia, la ragionevolezza della decisione non dipende solo dalla coerenza intrinseca ai dati che la sottendono: una decisione può essere perfettamente logica se derivata dai suoi elementi, eppure parziale, insoddisfacente e, in ultima analisi, "disumana". Disumana non nell'accezione di "sbagliata" o "malevola", ma poco consona alla complessità – alla ragionevolezza, appunto – che gli esseri umani esigono e di cui sono capaci. Qual è allora l'elemento – o gli elementi – costitutivi di questa ragionevolezza? La possibilità che il risultato, di qualsiasi ordine e contesto esso sia, possieda appunto una "ragione". Una ragione che abbia i tratti non solo di una spiegazione, ma di un significato, di un motivo. La possibilità, dunque, di comprendere, o almeno intuire, qualcosa che non rientri nella categoria della casualità.

Una siffatta ragione non può essere altro che una ragione *poetica*, laddove per poetica si intende ciò che arriva al fondo di una esigenza, a cui solo l'esperienza – e non solo il linguaggio, il calcolo, la spiegazione può rispondere. Qui sembra esservi la radice della vera responsabilità: capace di rispondere sottraendosi al dominio della casualità, all'inquietudine che da essa si genera, di rendersi portavoce di una risposta, di una intuizione, di una proposta, non solo di un *come*, ma di un *perché* che sostenga l'esigenza di ragionevolezza della decisione, del fatto, del dato, del fine.