

ETICA DELLA RICERCA, BIOETICA,
BIODIRITTO E BIOPOLITICA
IV, 2025

SPIEGABILITÀ E INTELLIGENZA ARTIFICIALE

Prospettive tecnologiche, etiche e normative

a cura di
Ludovica Marinucci

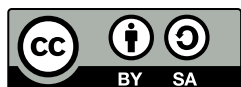
Abstract

In una prospettiva interdisciplinare, il volume collettaneo intende approfondire il concetto di ‘spiegabilità’ dell’Intelligenza Artificiale (IA), le sue implicazioni etico-giuridiche e le sfide tecnologiche, offrendo un dialogo tra ricercatori di diverse discipline (informatica, linguistica computazionale, logica, epistemologia, etica, bioetica, diritto, scienze cognitive e sociali). In particolare, sulla base di casi di studio, progetti di ricerca, approcci metodologici e interessi scientifici diversi, gli autori discutono se e in che modo il requisito chiave della spiegabilità possa essere effettivamente soddisfatto e implementato a livello tecnologico, cercando non solo di comprendere quali specifici obblighi di spiegabilità siano presi in considerazione dalla comunità di ricercatori ed esperti in ambito di IA, ma anche di prevedere quale grado di conformità etico-giuridica tali sistemi potranno possedere nel prossimo futuro e quali sono i relativi rischi attesi a livello individuale e sociale. Mostrando una pluralità di prospettive di ricerca, spesso in tensione, il volume invita a considerare la ‘spiegabilità’ come un concetto sfaccettato e stratificato che va considerato non come un requisito puramente tecnico ma come un insieme di pratiche epistemiche, etiche e normative.

Consiglio Nazionale delle Ricerche
Centro Interdipartimentale per l’Etica e l’Integrità nella Ricerca

Etica della Ricerca, Bioetica, Biodiritto e Biopolitica, vol. 4

ISSN (ed. digitale) 3103-2648
ISBN (ed. stampa) 978 88 8080 749 0
ISBN (ed. digitale) 978 88 8080 750 6
DOI 10.48220/eticabioeticabiodiritto-2025-4



La versione digitale è pubblicata in Open Access su www.edizioni.cnr.it.
Il presente lavoro è protetto dalla licenza [CC BY-SA 4.0](https://creativecommons.org/licenses/by-sa/4.0/)

Progetto grafico e impaginazione a cura di Tiziana Ciciotti
In copertina fotografia di Daniel Dalea su [Unsplash](https://unsplash.com/)

Pubblicato da
© **Cnr Edizioni, 2025**
Piazzale Aldo Moro, 7
00185 Roma
www.edizioni.cnr.it
bookshop@cnr.it

*ETICA DELLA RICERCA, BIOETICA,
BIODIRITTO E BIOPOLITICA*

Collana del
Centro Interdipartimentale per l'Etica e l'Integrità nella Ricerca del CNR



Diretta da

Cinzia Caporale, Elena Mancini, Ilja Richard Pavone

Comitato scientifico

Il Comitato scientifico della collana è composto dai membri della
Commissione per l'Etica e l'Integrità nella Ricerca del CNR
(www.cnr.it/it/ethics)

Comitato editoriale

Giorgia Adamo, Marco Annoni, Marco Arizza, Ludovica Durst, Andrea
Grignolio Corsini, Silvia Scalzini, Giulia Sciolli, Roberta Martina Zagarella

Comitato di redazione

Tiziana Ciciotti (*Responsabile*), Paola Grisanti, Emiliano Liberatori

Per informazioni: info@ethics.cnr.it

INDICE

<i>Prefazione</i> Cinzia Caporale	7
<i>Per un'Intelligenza Artificiale spiegabile: limiti, opportunità e sfide aperte</i> Ludovica Marinucci	11

PARTE I

Predizioni e scopi dell'Intelligenza Artificiale

<i>Prevedibilità senza spiegabilità: come il Deep Learning affronta le sfide del mondo naturale e della nostra mente</i> Guido Boella	21
<i>Quali sono gli scopi dell'Intelligenza Artificiale? Verso una spiegazione teleologica</i> Mattia Fumagalli e Roberta Ferrario	37

PARTE II

Logica, epistemologia ed etica dell'Intelligenza Artificiale

<i>BRIO: il ruolo della logica nella costruzione di un'Intelligenza Artificiale equa</i> Giuseppe Primiero	53
<i>Tra etica ed epistemologia: alcune note critiche sul principio della spiegabilità in Intelligenza Artificiale</i> Fabio Fossa, Giacomo Zanotti, Stefano Canali	67

PARTE III

Norme giuridiche e sociali ai tempi dell'Intelligenza Artificiale

<i>L'AI Act e la trasparenza, tra tracciabilità, spiegabilità e conoscibilità. Un nuovo tassello nell'ecosistema regolatorio della ricerca?</i>	
Carlo Casonato, Marta Fasan, Marta Tomasi	83
<i>Norme sociali e spiegazione del comportamento collettivo di umani e macchine</i>	
Giulia Andrighetto e Luca Tummolini	103
<i>Postfazione</i>	
Gilberto Corbellini	119
<i>Gli autori</i>	129

Prefazione

Cinzia Caporale

Centro Interdipartimentale per l'Etica e l'Integrità nella Ricerca, CNR

Tra le linee di ricerca sviluppate presso il Centro Interdipartimentale per l'Etica e l'Integrità nella Ricerca del CNR non vi è dubbio che l'analisi dei profili etico-giuridici dell'intelligenza artificiale (IA) è quella che nell'ultimo periodo è cresciuta maggiormente e in più breve tempo, fino a diventare centrale come materia in sé e pervasiva in tutti gli ambiti delle nostre ricerche. L'IA non si è, cioè, solo aggiunta alle nostre aree di interesse come un oggetto tra altri, ma ha agito come un fattore di riorientamento complessivo, inducendo una revisione delle categorie analitiche, dei linguaggi e delle responsabilità della ricerca stessa.

Il primo volume sul tema della collana "Etica della Ricerca, Bioetica, Biodiritto e Biopolitica" nasce da questa consapevolezza e indaga una questione centrale nella discussione etica e bioetica, quella della *spiegabilità* dei sistemi di IA. La scelta dell'argomento non risponde all'esigenza di isolare un requisito tecnico o normativo, ma all'intento di investigare una nozione che, più di altre, coglie le dissonanze tra le promesse e i limiti dell'IA nel rapporto con il sapere, la decisione e la responsabilità. Un problema teorico e pratico, la cui rilevanza attraversa in modo trasversale l'etica, la bioetica e il diritto.

Le diverse prospettive presenti evidenziano come la spiegabilità dell'IA non possa essere affrontata in modo univoco o riduzionistico. Al contrario, si tratta di un concetto stratificato, che cambia significato a seconda delle finalità e dei presupposti epistemici ed è caratterizzato da tensioni teoriche e pratiche che riflettono l'articolazione intrinseca dei sistemi di IA e dei contesti nei quali questi vengono disegnati, realizzati nel concreto e quindi utilizzati. Ed è in questa pluralità di approcci che si pone la necessità di disporre di forme di spiegazione sufficienti a rendere tali sistemi valutabili e, in ultima analisi, governabili.

Il volume mostra come la spiegabilità possa essere intesa, a seconda del punto di osservazione, come strutturalmente *impossibile*, logicamente ed epistemologicamente *necessaria* nonché *strumentale* sul piano etico-giuridico, oltre che relativa ai contesti d'uso e *critica* per la comprensione delle trasformazioni sociali, cognitive e culturali in divenire.

Dal punto di vista strutturale, l'impossibilità della spiegabilità rimanda ai 'limiti' di molte architetture di IA attualmente disponibili, in particolare di quelle basate su modelli opachi, ad alta dimensionalità e su processi di apprendimento non pienamente interpretabili nemmeno dagli stessi sviluppatori. In quest'ottica, la spiegabilità non è semplicemente un obiettivo tecnico ancora da raggiungere (sempre che a un certo punto ciò sia davvero possibile, e che continui ad esserlo al crescere della complessità dell'IA generativa), ma un problema che tocca la natura stessa di certi modelli computazionali. L'idea che ogni output algoritmico possa essere ricondotto a una spiegazione completa, lineare e intuitivamente comprensibile rischia di tradursi in un'aspettativa epistemicamente infondata e, sul piano pratico, fuorviante.

Il volume non elude questa difficoltà e certo non legittima l'opacità come valore, ma la assume come punto di partenza per liberarsi delle soluzioni semplificatrici o retoriche; quelle, purtroppo, maggiormente presenti nella letteratura etica e bioetica corrente e in particolare nei pareri di comitati e commissioni dedicati alla disamina di questi profili, il cui richiamo alla spiegabilità rischia di assolvere una funzione meramente rassicurante.

D'altro canto, la spiegabilità si rivela necessaria sul piano logico ed epistemologico. Senza una qualche forma di spiegazione, infatti, viene meno la possibilità stessa di comprendere, vagliare e discutere i risultati prodotti dai sistemi di IA. La spiegabilità diviene allora una condizione per l'attribuzione di significato alle decisioni automatizzate, per la loro integrazione in processi cognitivi e decisionali umani e per la costruzione di un sapere affidabile. In questa chiave, va evidenziato come la spiegazione non debba essere intesa esclusivamente come *trasparenza totale* del funzionamento interno del sistema, ma anche come definizione e accettazione di livelli diversi di comprensione, adeguati agli scopi conoscitivi e ai soggetti coinvolti. Essa non dovrebbe essere pensata come una proprietà assoluta dei sistemi di IA, valida indipendentemente dai contesti d'uso, ma come una relazione che coinvolge specifici domini applicativi, specifiche persone o gruppi e specifiche pratiche. Ciò che conta come *spiegazione adeguata* in un contesto medico può

essere radicalmente diverso da ciò che è richiesto in ambito giudiziario o amministrativo o scientifico: questa relatività non rappresenta una debolezza, bensì una condizione necessaria per rendere la spiegabilità effettivamente significativa e operante, una sua condizione di possibilità. Per di più, assumere l'opacità come dato di fatto non si traduce affatto nell'accettazione acritica delle decisioni algoritmiche, né nella delega automatica del giudizio morale a sistemi che non sono agenti morali.

È soprattutto sul piano etico e giuridico che la spiegabilità smette di essere un fine in sé e finisce per coincidere – secondo la visione più diffusa – con lo strumento ‘irrinunciabile’ per garantire nell’Era dell’IA principi e valori classici quali la responsabilità, l’equità, l’autonomia, la dignità della persona e la possibilità stessa di un controllo umano significativo, a sua volta considerato imprescindibile per assicurare la centralità dell’uomo. In ambito giuridico, poi, la spiegabilità diventa preconditione per l’esercizio dei diritti (come quelli alla non discriminazione e alla protezione da arbitrarietà, bias, asimmetrie di potere e scelte implicite), per la tutela dei diritti umani fondamentali e per la giustificabilità dell’uso dei sistemi di IA in ambiti che incidono sulla libertà e sulle vite delle persone (una decisione che non può essere spiegata è, se non altro sul piano morale, una decisione difficile da legittimare). Tuttavia, occorre non confondere l’efficacia regolativa con una garanzia di legittimazione morale in senso forte. Le richieste di spiegabilità formulate dal diritto europeo possono costituire strumenti di *governance* e *accountability* istituzionale, senza per questo risolvere le criticità etiche più profonde che l’IA solleva.

Il volume, con le problematicità che esplicita, è utile anche a questo fine, sottrarsi all’idea che la spiegabilità possa risolversi in un requisito astratto o irrealistico per tornare a essere calibrata in funzione delle diverse finalità regolative e del minimo etico desiderabile. *Distingue frequenter*, verrebbe da dire.

Al contrario, cercare una definizione univoca o una soluzione definitiva al problema della spiegabilità ci distoglierebbe dalla consapevolezza che questa debba essere affrontata come una sfida reale, complessa e interdisciplinare.

Le trasformazioni sociali, cognitive e culturali che l’IA produce e amplifica mettono in discussione automatismi e potere epistemico, e introducono nuove forme di delega decisionale e mutamenti nei criteri di autorità e *trust*. È possibile che nel futuro si attenuino valori considerati come largamente condivisi in determinate aree geopolitiche e culturali, e che la base normativa comune – ad esempio quella europea – non sia sufficientemente

robusta da reggere l'impatto dei sistemi di IA che via via vengono sviluppati e si affermano. Nel pluralismo ineliminabile che connota il mondo globalizzato e comunque le democrazie, appare problematico attribuire alle sole procedure giuridiche la capacità di produrre un consenso etico sostantivo. La spiegabilità, in questo scenario, non può essere assunta come garanzia universale, ma come uno degli strumenti attraverso cui rendere esplicite, e quindi discutibili, le scelte di valore che connotano lo sviluppo e l'uso dei sistemi di IA.

La crisi che deriva da questo squilibrio è un fatto strutturale. La spiegabilità non è soltanto una sfida tecnica, ma una prova di maturità etica per le istituzioni chiamate a governare l'IA nel mondo contemporaneo e per la stessa ricerca scientifica, chiamata a interrogare criticamente i propri presupposti e le proprie responsabilità.

Per un'Intelligenza Artificiale spiegabile: limiti, opportunità e sfide aperte

Ludovica Marinucci

Centro Interdipartimentale per l'Etica e l'Integrità nella Ricerca, CNR

Le preoccupazioni sulla 'spiegabilità' dell'IA non sono un fenomeno nuovo. Già nell'ambito della progettazione dei c.d. sistemi esperti (*expert systems*) sono stati studiati diversi schemi esplicativi, ad esempio presentando esplicitamente all'utente le regole impiegate nelle catene di ragionamento. La spiegabilità è così divenuta uno dei criteri di definizione dei sistemi esperti stessi: «One of the defining criteria of expert systems is their ability to 'explain' their operation¹». L'avvento di sistemi decisionali 'opachi', basati in particolare sulle reti neurali profonde (*Deep Neural Network*), non ha fatto altro che aumentare l'attenzione verso i rischi che sono legati al loro utilizzo per effettuare previsioni in contesti critici prendendo decisioni non giustificabili. L'opposto della natura opaca di tali modelli è, quindi, la sfida della 'spiegabilità', ovvero la ricerca di una comprensione diretta del loro funzionamento e della produzione di un certo risultato.²

Questa tensione tiene acceso il dibattito sui limiti e le opportunità di sviluppare la c.d. *eXplainable AI* (XAI), espressione originariamente coniata nel contesto del programma DARPA (*Defense Advanced Research Projects Agency*)³

¹ Bruce G. Buchanan e Reid G. Smith, "Fundamentals of expert systems", Annual review of computer science, 3, 1 (1988), p. 43. DOI: 10.1146/annurev.cs.03.060188.000323; si veda anche: William Swartout, Cecile Paris e Johanna Moore, "Explanations in knowledge systems: design for explainable expert systems", IEEE Expert, 6, 3 (1991): 58-64. DOI: 10.1109/64.87686.

² Zakary C. Lipton, "The mythos of model interpretability", Queue, 16, 3 (2018) 30:31-30:57. DOI: 10.1145/3236386.3241340.

³ David Gunning e David W. Aha, "DARPA's Explainable Artificial Intelligence (XAI) Program (2017)", AI Magazine, 40, 2 (2019): 44-58. DOI: 10.1609/aimag.v40i2.2850; David Gunning et al., "DARPA's explainable AI (XAI) program: A retrospective", Applied AI Letters, 2, 4 (2021). DOI: 10.1002/ail.2.61.

e ormai presente in tutti i principali documenti e conferenze che si occupano di IA.

A livello tecnologico, i lavori in ambito di XAI ambiscono alla creazione di tecniche che consentano un compromesso tra le prestazioni di un modello e la sua trasparenza.⁴ Questo si traduce nello sviluppo di modelli che mantengono un elevato livello di prestazioni di apprendimento (ad esempio, accuratezza delle previsioni) anche attraverso le correzioni delle loro carenze, rese possibili dalla migliore comprensione dei loro processi.⁵ In ciò, oltre agli sviluppatori, giocano un ruolo sempre più rilevante gli utenti e *stakeholder* di tali sistemi, per i quali le ‘spiegazioni’ sui loro *output* devono essere adeguate non solo al contesto di utilizzo ma anche alle loro conoscenze.⁶ Per questo motivo, la vera sfida dovrebbe consistere nel costruire sistemi che forniscono spiegazioni differenziate e situate, invece di rincorrere definizioni generalizzabili e spesso incoerenti perché non solo ignorano chi sono gli specifici destinatari della spiegazione (tecnici, cittadini, regolatori, ricercatori, ecc.), ma cercano anche di conciliare motivazioni allo sviluppo della XAI, tra cui fiducia, causalità, tollerabilità cognitiva e informatività, che sono spesso incompatibili. Ad esempio, la fiducia di un cittadino può non dipendere dalla comprensione causale del sistema bensì dalla credibilità dell’organizzazione che lo produce. Oppure, spiegazioni troppo informative, dettagliate o tecniche potrebbero risultare confondenti e ostacolare il processo decisionale.⁷

Tuttavia, i tentativi di fare ordine tra le varie definizioni, e la relativa terminologia utilizzata in modo intercambiabile o con sottili differenze di

⁴ Luca Longo, Randy Goebel, Freddy Lecue et al., “Explainable artificial intelligence: Concepts, applications, research challenges and visions”, in *Machine Learning and Knowledge Extraction. CD-MAKE 2020*, Lecture Notes in Computer Science, vol. 12279, Cham: Springer International Publishing, 2020. DOI: 10.1007/978-3-030-57321-8_1.

⁵ Zhu, Jichen, Antonios Liapis, Sebastian Risi et al., “Explainable AI for designers: A human-centered perspective on mixed-initiative co-creation”, in *2018 IEEE conference on computational intelligence and games (CIG)*, Maastricht, Netherlands: IEEE press, 2018. DOI: 10.1109/CIG.2018.8490433.

⁶ Robert R. Hoffman, Shane T. Mueller, Gary Klein et al., “Explainable AI: roles and stakeholders, desirements and challenges”, *Frontiers in Computer Science*, 5 (2023). DOI: 10.3389/fcomp.2023.1117848.

⁷ Denys O. Chergykalo e Dmitry A. Klyushin, “Fundamental Fallacies in Definitions of Explainable AI: Explainable to Whom and Why?”, in *Explainable AI: Foundations, Methodologies and Applications*, a cura di Mayuri Mehta et al., Cham: Springer International Publishing, 2023, pp. 25–42. DOI: 10.1007/978-3-031-12807-3_2.

significato che variano a seconda dell'autore, hanno il merito di aver messo in evidenza una questione centrale: la spiegabilità non è solo un'esigenza tecnologica ma rappresenta soprattutto la risposta a richieste epistemiche, etiche, sociali e giuridiche che convergono nell'importanza della spiegabilità e, per estensione, della trasparenza dei processi automatizzati. La loro inclusione nel novero dei requisiti etici dei principali documenti e linee guida sull'etica dell'IA, tra cui l'*Ethics Guidelines for Trustworthy AI* (2019), e successivamente in quelli normativi previsti nel recente *Artificial Intelligence Act* (2024) è avvenuta grazie al riconoscimento della natura ibrida dell'*explainability* – altro termine, meno frequentemente utilizzato, per intendere la 'spiegabilità' – la quale includendo sia la dimensione epistemologica dell'*intelligibility* ('come funziona?') sia quella etica dell'*accountability* ('chi è responsabile del suo funzionamento?') è «the crucial missing piece of the jigsaw when we seek to apply the framework of bioethics to the ethics of AI⁸». Qui vediamo la spiegabilità diventare un quinto principio etico fondamentale in quanto necessario a rendere operativi gli altri quattro principi classici della bioetica (beneficenza, non-maleficenza, autonomia e giustizia) di cui costituisce, quindi, la condizione di possibilità nel contesto dell'IA. Questa impostazione è stata condivisa e recepita a livello giuridico: la spiegabilità, intesa come condizione per la tutela dei diritti fondamentali, per la partecipazione democratica e per la contestabilità delle decisioni automatizzate, è divenuta requisito di legittimità dei sistemi di IA.⁹

I dibattiti filosofici relativi alla possibilità di considerare la 'spiegabilità' come un principio epistemico implicito piuttosto che un quinto principio etico autonomo capace di imporre di per sé un dovere morale¹⁰, oppure

⁸ Luciano Floridi, Josh Cows, J., Monica Beltrametti et al, "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations", *Minds & Machines*, 28 (2018): 689–707. DOI: 10.1007/s11023-018-9482-5; si veda anche: Luciano Floridi e Josh Cows, "A Unified Framework of Five Principles for AI in Society", *Harvard Data Science Review*, 1, 1 (2019). DOI: 10.1162/99608f92.8cd550d1.

⁹ Elisa Spiller, "Il diritto di comprendere, il dovere di spiegare. Explainability e intelligenza artificiale costituzionalmente orientata", *BioLaw Journal-Rivista di BioDiritto*, 2 (2021): 419-432. DOI: 10.15168/2284-4503-832.

¹⁰ João F. N. B. Cortese, Fabio G. Cozman, Marcos P. de Lucca-Silveira et al., "Should explainability be a fifth ethical principle in AI ethics?", *AI and Ethics*, 3, 1 (2023): 123–134. DOI: 10.1007/s43681-022-00152-w.

come un principio etico di secondo ordine collegato agli altri¹¹, non mettono tuttavia in discussione il suo ruolo fondamentale per la corretta applicabilità degli altri principi etici in ambito di IA. Ad esempio, il preservare l'autonomia, intesa come il potere dell'essere umano di decidere se delegare (*decide-to-delegate*¹²) e cosa alla macchina, sottende la necessità di avere informazioni adeguate e complete sul suo funzionamento per poter prendere liberamente tale decisione. In tale contesto, la questione centrale sembra essere se sia possibile una vera comprensione di cosa stia facendo la macchina e del perché abbia formulato un certo *output*. Questo richiama di nuovo il problema relativo ai potenziali destinatari della spiegazione, che possono interagire con lo stesso sistema in momenti diversi. In tal caso, è necessario decidere di volta in volta se seguire un approccio che cerca di presentare una spiegabilità minima adeguata alle esigenze comuni a tutti gli interessati, oppure se perseguire la strategia, di certo più impegnativa, di differenziare le spiegazioni in base ai diversi attori in gioco.¹³ Se guardiamo all'ambito della ricerca scientifica, con particolare riferimento agli studi empirici che coinvolgono partecipanti umani nelle varie fasi di sviluppo di sistemi di IA, se ne possono identificare almeno tre categorie: i ricercatori/sviluppatori, i partecipanti agli studi empirici e i comitati etici che approvano tali studi. In tale contesto, il principio dell'*explainability* in quanto elemento necessario per la stesura del consenso informato acquista un suo valore intrinseco consentendo il rispetto dell'autonomia dei partecipanti alle ricerche.¹⁴ Tuttavia, nonostante la proliferazione di svariati approcci, le attuali tecniche della XAI sono capaci di rendere solo approssimazioni dei modelli molto diverse dalle spiegazioni umane (filosofiche, psicologiche, sociologiche, ecc.),¹⁵ rendendo complesso

¹¹ Jessica Morley, Luciano Floridi, Libby Kinsey et al., "From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices", *Science and engineering ethics*, 26, 4 (2020): 2141–2168. DOI: 10.1007/s11948-019-00165-5.

¹² Luciano Floridi, Josh Cows, J., Monica Beltrametti et al., "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations", *op. cit.* p. 698.

¹³ Christian Herzog, "On the risk of confusing interpretability with explicability", *AI and Ethics*, 2, 1 (2022): 219–225. DOI: 10.1007/s43681-021-00121-9.

¹⁴ Frank Ursin, Cristian Timmermann e Florian Steger, "Explicability of artificial intelligence in radiology: Is a fifth bioethical principle conceptually necessary?", *Bioethics*, 36, 2 (2022): 143–153. DOI: 10.1111/bioe.12918.

¹⁵ Brent Mittelstadt, Chris Russell e Sandra Wachter. "Explaining explanations in AI", in *Proceedings of the conference on fairness, accountability, and transparency*, New York: Association for Computing Machinery, 2019, pp. 279–288. DOI: 10.1145/3287560.3287574.

tanto l'ottenimento di un vero consenso informato quanto la valutazione da parte dei comitati etici della spiegabilità stessa, nonché dei rischi e delle responsabilità derivanti dall'utilizzo del sistema di IA.¹⁶

Per superare i limiti della spiegabilità tecnica e arrivare a un'adeguata spiegabilità etico-legale, l'adozione di un approccio interdisciplinare nelle ricerche in ambito di IA sembra una delle strategie più convincenti. Su questa linea, il presente volume offre un dialogo tra ricercatori di diverse discipline (informatica, linguistica computazionale, logica, epistemologia, etica, bioetica, diritto, scienze cognitive e sociali) che, sulla base di casi di studio, progetti di ricerca, approcci metodologici e interessi scientifici diversi, discutono se e in che modo il requisito chiave della spiegabilità possa essere effettivamente soddisfatto e implementato a livello tecnologico, cercando non solo di comprendere quali specifici obblighi di spiegabilità siano presi in considerazione dalla comunità di ricercatori ed esperti in ambito di IA, ma anche di prevedere quale grado di conformità etico-giuridica tali sistemi potranno possedere nel prossimo futuro e quali sono i relativi rischi attesi a livello individuale e sociale. Il volume, quindi, mostrando una pluralità di prospettive di ricerca, spesso in tensione, invita a considerare la 'spiegabilità' come un concetto sfaccettato e stratificato che va considerato non come un requisito puramente tecnico ma come un insieme di pratiche epistemiche, etiche e normative.

Nel primo saggio, Guido Boella esplora la relazione tra 'prevedibilità' dei fenomeni e 'spiegabilità' dei modelli di *Deep Learning* i quali ottengono ottimi risultati proprio perché riproducono e rispecchiano la complessità caotica dei fenomeni che modellano, rendendo di fatto impossibile offrire spiegazioni esplicite del loro funzionamento interno. Per far fronte a questo limite epistemologico strutturale dell'IA connessionista, nel secondo saggio Mattia Fumagalli e Roberta Ferrario propongono l'idea di un *co-design* di progettisti e utenti dei sistemi di IA volto a chiarire, attraverso gli strumenti della modellazione ontologica, le 'finalità' di tali artefatti, ossia perché un sistema esiste e cosa deve fare, come elemento centrale per poter attribuire le responsabilità tanto della loro appropriata progettazione quanto del

¹⁶ Sarah Bouhouita-Guermech, Patrick Gogognon e Jean-Christophe Bélisle-Pipon, "Specific challenges posed by artificial intelligence in research ethics", *Frontiers in artificial intelligence*, 6 (2023): 1149082. DOI: 10.3389/frai.2023.1149082.

loro corretto utilizzo. Su questa linea, nel terzo saggio, Giuseppe Primiero riformula la ‘spiegabilità’ come un problema di ‘fiducia’ verso sistemi che, pur restando internamente opachi, possono essere valutati e resi affidabili attraverso gli strumenti della logica, proponendo in particolare i criteri formali della metodologia nata nel contesto del progetto BRIO (*Bias, Risk and Opacity in AI*). Tale prospettiva è integrata dal quarto saggio nel quale Fabio Fossa, Giacomo Zanotti e Stefano Canali interpretano la spiegabilità come un ‘valore strumentale’ che serve a tutelare autonomia, equità e responsabilità e, focalizzandosi sull’ambito medico, invitano a un pluralismo esplicativo dei sistemi opachi che li renda davvero comprensibili adattandosi a diverse categorie di utenti (pazienti, medici, tecnici, ecc.).

Il quinto e il sesto saggio offrono ulteriori prospettive affrontando aspetti normativi e sociali legati allo sviluppo e all’utilizzo dell’IA. In particolare, Carlo Casonato, Marta Fasan e Marta Tomasi analizzano l’AI Act mostrando, con particolare riferimento alla ricerca scientifica, come trasparenza, tracciabilità e spiegabilità si integrino nel nuovo ecosistema regolatorio europeo. Specularmente, Giulia Andrighetto e Luca Tummolini affrontano il tema della spiegabilità per studiare il comportamento collettivo in sistemi misti umani-macchine, mostrando come nuovi tipi di norme sociali stiano emergendo dall’interazione tra agenti umani e artificiali, spesso in modi non previsti dai progettisti. La spiegabilità, quindi, è vista non solo dalla prospettiva dall’alto delle norme giuridiche e dei vincoli istituzionali ma anche dal basso dei comportamenti reali e delle norme sociali emergenti grazie all’interazione con le macchine. Su questa linea, la Postfazione di Gilberto Corbellini decostruisce le narrazioni allarmistiche sull’AI, invitando invece a guardare la capacità dell’IA di amplificare e rendere più visibili *bias* umani, ridefinire dinamiche emotive e cognitive e trasformare abitudini individuali e sociali. Di conseguenza, sostiene Corbellini, più che sulla trasparenza tecnica occorre focalizzarsi sulla capacità di gestire gli effetti cognitivi e sociali che conseguono alla nostra interazione, sempre più pervasiva, con sistemi di IA.

Il volume, tenendo insieme questa stratificazione di prospettive e voci diverse, suggerisce quale possa essere un approccio proficuo, soprattutto per la ricerca scientifica, per affrontare le complesse sfide poste dall’IA.

RINGRAZIAMENTI

Sono grata a Cinzia Caporale per lo stimolo e la fiducia, senza cui il volume non sarebbe nato, e a Tiziana Ciciotti e Marco Arizza per la paziente cura editoriale. Rivolgo inoltre un forte ringraziamento agli autori del volume, in ordine alfabetico, Giulia Andrighetto, Guido Boella, Stefano Canali, Carlo Casonato, Gilberto Corbellini, Marta Fasan, Roberta Ferrario, Fabio Fossa, Mattia Fumagalli, Giuseppe Primiero, Marta Tomasi, Luca Tummolini e Giacomo Zanotti, nonché a Paolo Dario, Nicola Lettieri, Monica Palmirani, Vieri Giuliano Santucci, Salvatore Sapienza e Teresa Scantamburlo, che hanno partecipato al dibattito tenutosi nel corso del *webinar* da cui ha preso le mosse il volume.

PARTE I

PREDIZIONI E SCOPI DELL'INTELLIGENZA ARTIFICIALE

Prevedibilità senza spiegabilità: come il *Deep Learning* affronta le sfide del mondo naturale e della nostra mente¹

Guido Boella

Università di Torino

STORIA DI UN PROBLEMA COMPLESSO

Da sempre gli esseri umani cercano un ordine nell'imprevedibilità del nostro mondo naturale e sociale, e l'Intelligenza Artificiale (IA) è solo l'ultimo strumento per realizzare questo nostro sogno.

Dalla nascita dell'intelligenza artificiale nell'estate del 1956 al Dartmouth College, nella città di Hanover, nel New Hampshire (US), i ricercatori dell'IA si sono focalizzati per decenni sull'idea di costruire macchine capaci di imitare l'intelligenza umana. All'inizio, l'approccio per raggiungere l'obiettivo era di codificare in un programma i modelli costruiti per spiegare il funzionamento delle nostre capacità mentali. Ad esempio, per creare macchine intelligenti si sono utilizzate le stesse regole della logica in cui riteniamo si articoli il nostro pensiero. Nonostante negli anni si siano raggiunti molti risultati scientifici, questa metodologia, chiamata 'IA simbolica' non ha portato ai successi sperati, culminando in lunghi periodi di stasi conosciuti come 'inverni dell'IA'.

L'approccio nel costruire 'modelli della realtà' è ispirato all'idea base della scienza moderna che spiega i fenomeni del mondo (nel caso dell'IA, la nostra mente) e fa previsioni (replicare il comportamento intelligente) scoprendo e studiando le leggi che li regolano. Dal 1600, l'avvento della scienza moderna ha mostrato che è possibile utilizzare leggi e modelli matematici per spiegare il mondo della fisica e fare previsioni. «L'Universo è scritto nella lingua della matematica», diceva Galileo Galilei nel suo *Saggiatore* del 1623: dove non arrivano i simboli della matematica «è un aggirarsi

¹ Una versione preliminare di questo articolo è comparsa sul MagIA - Magazine Intelligenza Artificiale: <https://magia.news/lintelligenza-artificiale-non-ha-paura-del-caos/>

vanamente per un oscuro labirinto»², immagine quasi da Inferno dantesco.

Il metodo scientifico, mezzo secolo dopo Galileo, raggiunge la sua maturità con il grande matematico e fisico inglese Isaac Newton nei suoi *Principia mathematica philosophiae naturalis* del 1687: la sua legge di gravitazione universale spiega ad un tempo il moto di caduta verso terra dell'aneddotica mela (che cadde, forse non in testa, a Woolsthorpe Manor, da un albero che ancora oggi è conservato nel villaggio del Lincolnshire, in Inghilterra) e il moto circolare degli astri, ritenuti ancora qualche decennio prima 'corpi celesti', cioè di natura quasi divina, e da quel momento ridotti a 'pietre rotolanti' nello spazio. Ma da genio quale era, Newton capisce anche che il mondo non è totalmente paragonabile al meccanismo di un orologio, come sostengono i meccanicisti. Ad esempio, il Marchese Laplace esprime questa visione deterministica attraverso l'esperimento mentale del famoso 'demone', che ha preso il suo nome: gli esseri umani non riescono a prevedere tutto con le leggi della fisica, ma solo perché ignorano le condizioni iniziali di ogni elemento. Un'intelligenza superiore – il demone, appunto – che conoscesse la posizione, la velocità di ogni corpo e le forze che agiscono su di esso, potrebbe spiegare sia tutto il passato che prevedere l'intero futuro. Per Newton la visione interamente meccanicista è sbagliata prima di tutto perché incompatibile con la sua visione teologica, basata sull'idea che la natura non sia autosufficiente e che non possa fare a meno di Dio. Pochi sanno, perché fu lo stesso Newton a mantenere il massimo riserbo, che durante tutta la vita si dedicò non solo alla scienza ma anche a studi di teologia (i suoi scritti religiosi sono stati messi all'asta dagli eredi nel 1936, poi donati alla Biblioteca Nazionale di Israele), storia sacra e alchimia. L'economista John Maynard Keynes, che acquistò parte degli scritti, non a caso, lo definì 'l'ultimo degli alchimisti'.

Newton aveva capito che l'Universo, anche se fosse stato regolato interamente da leggi semplici, sarebbe stato difficile da prevedere. L'intuizione della difficoltà del problema spiega il rifiuto nell'aiutare l'astronomo inglese Edmund Halley, mentore e finanziatore, nel calcolare l'orbita dell'omonima cometa, apparsa pochi anni prima, per cercare di prevederne il ritorno. A Newton era chiara la difficoltà nel dover ricalcolare l'orbita ad ogni

² Galileo Galilei, "Il Saggiatore" (1623), in *Opere*, ed. Nazionale a cura di Antonino Favaro, Firenze: Giunti-Barbera, 1966, vol. VI, p. 232.

spostamento della cometa, perché influenzata nel suo percorso non solo dalla massa del sole, ma anche da quella dei pianeti a cui passa vicino, pianeti che si muovono a loro volta attorno al sole e, quindi, la loro influenza sulla cometa cambia di conseguenza. Oggi questo tipo di calcoli si possono fare con i supercomputer. Newton, però, aveva già intuito che c'era un ulteriore livello di complessità, che rende difficile spiegare e predire l'universo con un modello matematico o da simulare anche con l'aiuto di un calcolatore.

Negli scritti di Newton troviamo già traccia di quello che alcuni secoli dopo verrà chiamato 'il problema dei tre corpi'. Il termine è oggi popolare grazie a una serie tratta dall'omonimo romanzo di fantascienza del cinese Liu Cixin³ nel quale una specie aliena è costretta a emigrare verso la Terra perché il suo pianeta orbita attorno a un sistema composto da tre soli e questa combinazione rende totalmente imprevedibile calcolarne l'orbita e, quindi, quella del pianeta e il susseguirsi delle stagioni. Il problema dei tre corpi è un caso particolare di 'sistema caotico', concetto studiato dal matematico e meteorologo Edward Lorenz. L'uomo vive circondato da molti fenomeni che possono essere descritti come 'sistemi caotici': il meteo (siamo appunto partiti da quello), il cambiamento climatico, il corpo umano⁴, le popolazioni biologiche, gli ecosistemi, il traffico, il sistema economico finanziario, ecc. Un sistema caotico è caratterizzato principalmente dal fatto che una minima (piccola a piacere) differenza dello stato iniziale porta, dopo un limitato orizzonte temporale, all'evoluzione del sistema verso esiti completamente diversi ad altre scale più grandi, con effetti non proporzionali alla differenza di condizioni iniziali. Questa proprietà – dipendenza sensibile dalle condizioni iniziali – è descritta da Lorenz con una celeberrima metafora: «Può il battito di ali di una farfalla in Brasile scatenare un tornado in Texas?»⁵. Certo che no, ma rende bene l'idea della sproporzione che può esistere fra cause ed effetti in un sistema caotico in cui tutto influisce su tutto.

³ Cixin Liu, *Il problema dei tre corpi*, trad. (dall'inglese) di Benedetta Tavani, Milano: Mondadori, 2017 [Trad. in inglese dall'originale cinese a cura di Ken Liu, *The Three-Body Problem*, New York: A Tor Book – Tom Doherty Associates, 2014].

⁴ Anastasia Korolj, Hau-Tieng Wu, and Milica Radisic, "A Healthy Dose of Chaos: Using fractal frameworks for engineering higher-fidelity biomedical systems". *Biomaterials*, 219, 119363 (2019). DOI: 10.1016/j.biomaterials.2019.119363.

⁵ Edward Lorenz, "Predictability: Does the Flap of a Butterfly's Wings in Brazil Set off a Tornado in Texas?", Transcript of a lecture given to the 139th meeting of the American Association for the Advancement of Science, 1972.

Quindi, come faccio a sapere se un sistema è caotico? Esso deve essere caratterizzato da tre condizioni:

- i. Agenti o influenze multipli: più agenti o fattori che interagiscono tra loro influenzano gli esiti.
- ii. Agenti adattabili: gli agenti nel sistema devono essere in grado di adattarsi alle circostanze esterne, di cambiare i loro comportamenti con i cambiamenti nel loro ambiente.
- iii. Informazioni localizzate: gli agenti devono considerare solo informazioni locali (non conoscono informazioni sull'intero sistema) mentre prendono le loro decisioni.

Quest'ultimo è il motivo per cui gli scacchi sono complessi, ma non caotici. Pensate ai sistemi composti dalle cellule che formano il nostro corpo, ai veicoli nel traffico, agli agenti di borsa, ecc. La teoria del caos è strettamente legata alla teoria dei sistemi complessi. I sistemi complessi sono composti da molte parti interconnesse che interagiscono in modi non lineari, mostrano comportamenti emergenti che non possono essere facilmente previsti dalle proprietà dei singoli componenti e sviluppano spontaneamente una struttura o un comportamento ordinato senza un controllo esterno centralizzato.

LE PREVISIONI DEL *DEEP LEARNING*: DAL METEO AL LINGUAGGIO

Finora, ci eravamo rassegnati a non poter prevedere tutto con i nostri modelli matematici, sapevamo che al massimo potevamo fare simulazioni approssimate ma solo fino ad un certo orizzonte temporale, come accade con le previsioni del tempo, che sono affidabili solo per pochi giorni.

Invece, negli ultimi anni i nuovi sviluppi dell'IA hanno fatto emergere un altro paradigma, il c.d. *Machine Learning*, l'apprendimento automatico. I sistemi di *Machine Learning* – e, in particolare, quelli che usano l'apprendimento profondo, il *Deep Learning* – apprendono a classificare o a generare informazioni partendo solamente da esempi, in grande quantità, e senza richiedere che venga codificata a mano in un programma la nostra conoscenza, come invece avveniva nel paradigma precedente dell'IA. E questo nuovo approccio inizia a mostrare una capacità migliore della nostra di fare previsioni nei sistemi caotici e complessi. Abbiamo adesso un nuovo

attore in grado non solo di risolvere il problema dei tre corpi⁶, ma anche, di recente, di fare previsioni in ambiti di forte impatto, come il meteo e la salute.

Google Deepmind, nel novembre 2023, ha presentato il sistema di IA chiamato GraphCast che, se non riesce ancora a tenere conto dei battiti di ali delle farfalle, prevede centinaia di variabili meteorologiche per i prossimi 10 giorni con una risoluzione di 0,25° a livello globale in meno di 1 minuto. GraphCast supera significativamente le prestazioni dei sistemi deterministici operativi più accurati sul 90% dei 1.380 obiettivi di verifica e le sue previsioni supportano una migliore previsione di eventi gravi, incluso il monitoraggio dei cicloni tropicali, i fiumi atmosferici e le temperature estreme⁷.

È necessario, però, fare attenzione: essendo basato sulla tecnologia del *Deep Learning*, GraphCast fa le previsioni del tempo senza conoscere cosa siano ‘sole’, ‘nuvole’, ‘pioggia’, ‘mari’ o ‘monti’, e senza conoscere le leggi fisiche della meteorologia. GraphCast lavora sulle serie storiche: in circa un mese impara a fare previsioni analizzando le serie storiche di alcuni decenni delle diverse variabili meteorologiche in tutto il globo terrestre: temperatura, pressione, umidità, precipitazioni, ecc. Ad oggi, le previsioni del tempo vengono fatte (ad esempio, nell’European Centre for Medium-Range Weather Forecasts (ECMWF) di Bologna) da decine tra i migliori fisici mondiali, che costruiscono un modello matematico della Terra fatto girare per ore su un supercomputer; GraphCast riesce, invece, a fare le previsioni in un minuto con una sola TPU (un tipo di processore utilizzato per l’IA).

Sempre Google Deepmind con *Isomorphic Labs* l’8 maggio scorso 2024⁸ ha presentato *AlphaFold 3*, un sistema di *Deep Learning* che non solo predice la struttura di proteine, DNA e RNA, ma anche le loro interazioni con una accuratezza del 50% migliore rispetto ai sistemi esistenti. Questi fenomeni sono ulteriori esempi di sistemi complessi e AlphaFold 3 apre nuove prospettive sulla comprensione dei meccanismi alla base della vita e sull’individuazione di nuovi principi farmacologici.

⁶ Breen, Philip G. et al., “Newton versus the machine: solving the chaotic three-body problem using deep neural networks”, *Monthly Notices of the Royal Astronomical Society*, 494, 2 (2020): 2465-2470. DOI: 10.1093/mnras/staa713.

⁷ Remi Lam et al., “Learning skillful medium-range global weather forecasting”, *Science*, 382, 6677 (2023): 1416-1421. DOI: 10.1126/science.adi2336.

⁸ John Abramson et al., “Accurate structure prediction of biomolecular interactions with AlphaFold 3”, *Nature*, 630 (2024): 493-500. DOI: 10.1038/s41586-024-07487-w.

Il successo della nuova IA basata sul *Deep Learning* non riguarda soltanto il mondo della fisica e della biologia, ma sorprendentemente anche i fenomeni della nostra mente, come il linguaggio.

Vi siete mai chiesti come faccia il linguaggio umano, pur essendo solo una sequenza lineare di simboli, a rappresentare un mondo composto da complessi accadimenti naturali e sociali, incluse le sofisticate interazioni umane, tutti fenomeni ad alta dimensionalità? Vediamo se e come a questa domanda sono riusciti a rispondere i linguisti computazionali che usano l'intelligenza artificiale per studiare il linguaggio e insegnarlo alle macchine.

Negli ultimi sessant'anni noi linguisti computazionali, usando il paradigma dell'IA simbolica abbiamo provato a creare sistemi artificiali che fossero in grado di 'comprendere' il linguaggio umano e di 'parlare' con noi e come noi: nonostante i grandi progressi scientifici, i risultati pratici però sono stati minori delle aspettative. In un primo momento, abbiamo provato a modellare il linguaggio con grammatiche formali che codificano le regole di composizione di una frase. Regole che per semplicità hanno un carattere locale (ad esempio, mettono assieme articolo e nome, con in mezzo, opzionalmente, uno o più aggettivi per formare un sintagma nominale da comporre poi con un verbo, ecc.). Regole così semplici da essere quasi sempre addirittura 'libere dal contesto', usando la definizione di Noam Chomsky, uno dei padri della linguistica computazionale, inventore delle grammatiche formali.

Per modellare il significato di una frase, abbiamo introdotto formalismi simbolici di rappresentazione della conoscenza, come le reti semantiche e le ontologie formali, che però hanno cristallizzato la semantica del linguaggio in concetti di geometrica precisione, in contrasto con quanto ci dicevano già dagli anni Settanta scienziati cognitivi come Eleanor Rosch e George Lakoff, secondo cui i concetti hanno confini molto imprecisi e difficili da formalizzare con una formula della logica. Ma, come si usa dire, "la guerra si fa con i soldati che si hanno": all'epoca non c'erano altri strumenti per raggiungere l'obiettivo, e comunque queste ricerche hanno portato molti avanzamenti e gettato le basi per i progressi successivi.

Una volta capita la difficoltà di procedere in questo modo troppo rigido, noi linguisti computazionali, a fine secolo, abbiamo aggiunto modelli statistici del linguaggio per introdurre più flessibilità: abbiamo costruito modelli del

linguaggio basati sulle frequenze di co-occorrenza delle parole (*bigrammi*, *trigrammi*, ..., *n-grammi*). Ad esempio, ‘io vedo’ è una coppia di parole più frequente dell’improbabile bigramma ‘io casa’, e quindi una frase che usi la seconda espressione probabilmente non è corretta. Tuttavia, anche con questa tecnica si rimane in un approccio locale, fermi all’interno di una frase.

I modelli statistici hanno rappresentato un passo avanti importante anche verso la capacità di gestire grandi quantità di dati linguistici in modo non solo più flessibile, ma anche più scalabile, cioè non limitandosi ad un singolo dominio o a una singola lingua. Questi modelli hanno permesso di automatizzare l’estrazione di conoscenze linguistiche, rendendo possibile l’analisi di corpora di dimensioni notevoli che altrimenti sarebbero stati impraticabili da trattare costruendo manualmente grammatiche, lessici, ontologie, ecc. Così facendo, però, si è rinunciato ad una nozione di significato più profonda, ancorché più rigida, riducendo la semantica alla co-collocazione fra parole (c.d. *Distributional Semantics*): ‘gatto’ e ‘micio’ hanno lo stesso significato non perché si riferiscono entrambi ad una specie di simpatico felino nel mondo reale, ma perché compaiono spessissimo in frasi simili.

Con queste tecnologie, tuttavia, siamo al più arrivati a creare assistenti vocali, come Alexa e Siri, con le loro frustranti conversazioni. Un risultato così limitato è anche dovuto al fatto di non essere riusciti a dare una risposta alla domanda riguardo alla capacità del linguaggio, apparentemente lineare, di rappresentare la complessità del mondo.

Poi, il 30 novembre 2022 è arrivata una nuova generazione di modelli probabilistici, basati su nuove tecniche di *Machine Learning*: i c.d. *Transformers*, le reti neurali trasformative), ovvero GPT, e di rincorsa tutti gli altri *Large Language Model* (LLM), come LLama, Claude, Gemini, Mistral, Minerva, Llamantino, Ernie, Pangu, Wudao, ecc. Si tratta di modelli che hanno imparato a ‘parlare’ da soli, quasi come un essere umano, senza che nessuno gli abbia ‘spiegato cosa siano’ (o per meglio dire ‘nessuno li abbia programmati codificando’) le regole grammaticali, i concetti che costituiscono il significato delle parole, le strutture argomentative, le intenzioni comunicative e neanche come esprimiamo le emozioni con l’intonazione di una frase.

Allo stesso tempo gli LLM non ‘comprendono’ alcunché: tutto quello che fanno è, dato il testo in *input*, restituire l’insieme di parole che ne costituisce la continuazione più probabile alla luce della conoscenza statistica

imparata nella fase di apprendimento digerendo tutti i testi del Web in ogni lingua. Al contrario, l'apprendimento del linguaggio da parte degli esseri umani è ricco di interazioni sociali e contestuali, non necessita di enormi quantità di esempi (povertà dello stimolo) ed è molto rapido, caratteristiche che hanno portato Noam Chomsky a sostenere che la grammatica sia innata nella mente umana e base condivisa da tutti i linguaggi umani. Gli LLM, però, hanno delle performance linguistiche eccezionali, generano risposte coerenti e contestualmente appropriate, che li hanno resi il principale successo commerciale dell'AI (dopo gli algoritmi di personalizzazione della pubblicità).

LLM E LINGUAGGIO NATURALE: PER UNA MUTUA (IN)SPIEGABILITÀ

Qual è il segreto del successo degli LLM nell'elaborare il linguaggio umano? Cosa hanno capito che noi linguisti computazionali non avevamo capito? Saranno in grado rispondere alla nostra domanda iniziale, o aiutarci a dare una risposta? Ancora non lo sappiamo con esattezza, ma cominciano ad emergere degli indizi, legati di nuovo alle teorie del caos e dei sistemi complessi. Gli LLM possono essere uno strumento prezioso per studiare il linguaggio stesso. Questa affermazione non va equivocata: non vogliamo qui suggerire che studiando il funzionamento interno degli LLM riusciremo a capire qualcosa di più del linguaggio umano. Un tale approccio ci potrebbe fare ricadere circolarmente nel peccato originale, la fallacia iniziale 'cartesiana' di noi linguisti, alla base della metodologia dei primi decenni di tentativi di analisi del linguaggio umano: la pretesa che con la nostra autocoscienza avremmo potuto accedere ai meccanismi della mente che sono alla base del linguaggio umano, definendo le regole della grammatica e il significato dei concetti con la logica. Per decenni un peccato originale di *hybris* ha portato la ricerca verso una strada (anche se l'unica percorribile) che ci ha fatto smarrire nel tentativo di comprendere il fenomeno linguaggio.

Tuttavia, diversamente che nel passato quando studiavamo la nostra mente e cervello, possiamo 'aprire la scatola e mettere le mani dentro' gli LLM. I ricercatori stanno cominciando a identificare quali pattern di attivazione di neuroni artificiali corrispondono ad una certa parola, e a modificare i pesi della rete neurale per stimolare o inibire una risposta in una certa direzione (come

stanno facendo ad *Anthropic* sul loro LLM Claude⁹). Purtroppo, analizzare e testare trilioni di parametri è probabilmente un compito troppo complesso per poter diventare una soluzione generale e, come vedremo più avanti in questo articolo, forse anche questa strada è fallace per principio e non per le limitazioni della tecnologia attuale: fallace allo stesso modo della pretesa di studiare come la mente umana comprenda il linguaggio con dei modelli locali come le grammatiche.

Grazie all'utilizzo degli LLM, i linguisti oggi possono generare statistiche sulle frequenze con cui le parole possono comparire in un qualsiasi testo. Prima degli LLM si potevano fare solo approssimazioni (brutali, tipo sostituire una parola con la sua lunghezza per semplificare il problema). Oggi un LLM, per costruzione, dato come contesto un testo lungo anche un milione di parole, restituisce non solo la parola più probabile che costituisce la prosecuzione di un dato contesto, ma può darci la probabilità in quel contesto di tutte le parole del dizionario. Come dire, gli LLM diventano una specie di 'acceleratore di particelle' per permettere ai linguisti, come ai fisici, di studiare una realtà invisibile ad occhio nudo (in questo caso invisibile non perché troppo piccola da vedere, ma perché troppo grande da abbracciare in un solo sguardo).

Usando gli LLM come strumenti di analisi, i linguisti stanno rafforzando la consapevolezza che il linguaggio umano ha proprietà particolari, ma fondamentali, che abbiamo ignorato per lungo tempo. Prima di tutto, e sorprendentemente, il linguaggio umano ha una natura frattale. I frattali sono geometrie complesse che mostrano una auto-similarità a diverse scale: se si guarda una parte di un frattale, si vedrà una figura simile alla figura più grande da cui è stata presa. I frattali sono figure geometriche speciali perché, pur essendo contenuti in uno spazio finito, possono avere un livello di dettaglio infinito. Un frattale può avere una 'dimensione' che non è un numero intero. Ad esempio, una linea ha una dimensione di 1, un quadrato ha una dimensione di 2, ma una linea frattale può avere una dimensione frazionaria (come 1,5), riflettendo la sua complessità in una linea monodimensionale che riempie però una superficie intera ripetendosi

⁹ Adly Templeton et al., "Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet", Transformer Circuits Thread (2024). <https://transformer-circuits.pub/2024/scaling-monosemanticity/>.

all'infinito. I frattali sono un fenomeno ben presente in natura: si pensi alla struttura degli alberi, delle foglie (ben evidente è la frattalità della struttura delle foglie di felce), dei cavolfiori, delle montagne, delle coste, dei fiocchi di neve, ecc. Ma hanno natura frattale anche i fenomeni sociali come il traffico, l'economia, la finanza, l'urbanizzazione, ecc.

La natura frattale del linguaggio non si riduce solo al fatto che le stesse proprietà (dipendenza, coerenza, strutturazione gerarchica) si ritrovano a scale diverse nel linguaggio (dal livello del testo, del paragrafo a quello della frase fino alle parole stesse), ma questa auto-similarità è affiancata anche dall'esistenza di correlazioni su scale arbitrariamente lunghe nei testi¹⁰, fenomeno ancora largamente non compreso. Le dipendenze a lungo raggio si riferiscono alla situazione in cui i punti dati in una serie temporale sono correlati per lunghi periodi. L'esistenza di una correlazione fra punti lontani significa che le osservazioni distanti nel tempo hanno ancora un'influenza significativa l'una sull'altra: le parole di una frase non sono solo correlate fra loro, ma sono correlate anche con altre strutture ad una scala più ampia, come il paragrafo che contiene la frase o l'intero testo in cui compare.

Ci sono diversi indici che ci segnalano e ci permettono di misurare le proprietà scalari del linguaggio umano che mostrano il suo carattere frattale, come la distribuzione statistica delle parole in un testo dettate dalla legge di Zipf e dalla legge di Heaps¹¹. Il rispetto di entrambe le leggi dimostra che il linguaggio non segue un pattern lineare semplice, ma piuttosto un comportamento complesso e ricco di dettagli, proprio come fanno i frattali. Quando proviamo però a generare un testo con i metodi tradizionali basati sulla teoria dell'informazione o sul trattamento statistico del linguaggio naturale, approcci che si basano prevalentemente su metodi di correlazioni

¹⁰ Eduardo G. Altmann, Giampaolo Cristadoro, and Mirko Degli Esposti, "On the origin of long-range correlations in texts", *Proceedings of the National Academy of Sciences*, 109, 29 (2012): 11582-11587. DOI: 10.1073/pnas.1117723109.

¹¹ La legge di Zipf stabilisce che le parole più comuni in un testo sono molto più frequenti delle parole meno comuni. La seconda parola più comune appare circa la metà delle volte della parola più comune, la terza parola un terzo delle volte, e così via (George Kingsley Zipf, *Human Behaviour and the Principle of Least-Effort*, Cambridge: Addison Wesley Press, 1949). Invece, la legge di Heaps stabilisce che il numero di parole uniche in un testo cresce man mano che aumenta la sua lunghezza, ma cresce a un ritmo decrescente. All'inizio si trovano molte parole nuove, ma poi si trovano sempre meno parole nuove man mano che si continua a leggere (Gustav Herdan, *Type-token mathematics: A textbook of mathematical linguistics*, The Hague: Mouton, 1960).

locali, a breve raggio, non troviamo traccia nel testo prodotto degli effetti delle leggi di Zipf e Heaps. Quindi stiamo producendo un testo degenerato, meno ricco di dettagli, diverso da quello prodotto dagli umani quando parlano. Il linguaggio, quindi, presenta due caratteristiche: (i) frattalità e (ii) dipendenze a lungo raggio, le quali sempre troviamo nelle attività sociali (traffico, finanza, ecc.) e nel mondo naturale (anche il corpo umano ha caratteri frattali, e la dipartenza dal giusto livello di frattalità è associata a malattie). Questi fenomeni sono parte di sistemi complessi e caotici, e infatti i concetti di frattalità e dipendenza a lungo raggio hanno una stretta relazione con il concetto di sistemi complessi e caotici.

Come dicevamo sopra, essendoci una relazione fra teoria del caos e l'IA, le tecnologie di *Deep Learning* – su cui si basano anche gli LLM – paiono cogliere meglio la complessità di alcuni sistemi caotici, come i fenomeni naturali studiati dalla meteorologia o dalla biologia, permettendo di fare previsioni migliori di quelle che riusciamo a fare con i nostri metodi approssimati. Se si usa un modello neurale per generare un testo le due leggi (i, ii) valgono ancora nel testo prodotto. Questo vuol dire che i modelli basati su reti neurali – e fra questi i recenti LLM – riescono meglio a riprodurre il linguaggio come lo parliamo noi umani dei modelli che abbiamo costruito pensando di aver capito il funzionamento del linguaggio. Usando come strumento gli LLM, i ricercatori¹² sono riusciti a provare ulteriormente non solo che il linguaggio è auto-simile, mostrando complessità a tutti i livelli di granularità, senza una lunghezza di contesto caratteristica particolare, ma anche la presenza di dipendenze a lungo raggio, caratteri che sono cominciati ad emergere solo recentemente con le reti neurali ricorrenti (RNN) e LSTM¹³ che precedono tecnologicamente gli LLM. E dimostrano anche che le dipendenze a breve termine nel linguaggio, come nei paragrafi, riflettono le dipendenze su scale più ampie, come interi documenti.

Allo stesso tempo, la dimostrazione dei caratteri frattali presenti nel linguaggio ci aiuta a fare luce su come riescano a funzionare gli LLM,

¹² Ibrahim Alabdulmohsin, Vinh Q. Tran, and Mostafa Dehghani, “Fractal Patterns May Illuminate the Success of Next-Token Prediction”, Proceedings of 38th Conference on Neural Information Processing Systems (NeurIPS), 2024. DOI: 10.48550/arXiv.2402.01825.

¹³ Shuntaro Takahashi, and Kumiko Tanaka-Ishii, “Do neural nets learn statistical laws behind natural language?”, PloS one, 12, 12 (2017): e0189326. DOI: 10.1371/journal.pone.0189326.

sistemi che rimangono ancora oggetti molto misteriosi, anche se costruiti grazie all'ingegno degli esseri umani. Gli LLM, infatti, 'imparano' da soli sottoponendosi ad infiniti indovinelli di completamento di una frase a cui hanno tolto una parola secondo il meccanismo della c.d. *Next Token Prediction* (predizione della parola successiva); e questo indovinello viene ripetuto su tutto il testo presente sul web che gli LLM hanno scaricato e letto, in tutte le lingue. Questa fase di apprendimento permette alla fine agli LLM non solo di predire la prossima parola di una nuova frase mai vista prima, ma anche, dato un contesto formato da un testo lungo decine di migliaia di parole (fino ad un milione di parole in LLM come *Gemini 1.5* pro di Google), di produrre in output un testo di migliaia di parole che è la continuazione più probabile del testo dato in input.

Perciò, gli LLM riescono a scalare quanto hanno appreso in un contesto locale, generalizzando le informazioni del *training set*, in modi che superano le semplici probabilità di una sequenza di parole, ad un contesto molto più ampio che non hanno considerato nella fase di apprendimento. Ma proprio perché il linguaggio è 'auto-simile' a tutti i livelli, la conoscenza appresa a livello locale per fare previsioni sulla prossima parola ritorna utile al LLM anche ad un livello più ampio: l'auto-similarità implica che i modelli nel linguaggio a livello di paragrafo riflettano i modelli osservati a livello di testo intero.

Viceversa, poiché il linguaggio mostra fenomeni dettagliati e complessi ad ogni livello di granularità, non è sufficiente affidarsi solo al contesto locale di una frase per prevedere con correttezza il prossimo *token*. Tuttavia, gli LLM riescono ad applicare i modelli da loro appresi a livello locale (la prossima parola) anche a livelli di granularità superiori; cioè, comprendono fenomeni di più alto livello, come la direzione dell'argomento e il contesto più ampio e l'intenzione del parlante. Gli LLM riescono a bilanciare tra contesti a breve e lungo termine, e potrebbe essere questa la ragione del loro successo: riescono a gestire la frattalità e complessità insita nel linguaggio, che è la preconditione per poter rappresentare un mondo ad alta dimensionalità e complesso. Abbiamo, quindi, trovato la risposta alla nostra domanda iniziale. Ma ci rimangono ancora due punti da affrontare per chiudere l'argomento.

PREVEDIBILITÀ SENZA SPIEGABILITÀ

Perché non possiamo pensare di poter guardare dentro un LLM per cercare di studiare il linguaggio umano? Il primo motivo è che non abbiamo alcuna garanzia che gli LLM ‘comprendano’ il linguaggio nello stesso modo in cui lo comprendiamo noi: potrebbero stare solo ‘simulando’ quanto facciamo, anche se con una perfezione stupefacente. E se anche potessimo provare che la loro ‘comprensione’ avviene in maniera simile alla nostra, c’è un secondo motivo, di principio: è un errore metodologico pensare di studiare gli LLM per estrarre finalmente un modello esplicito, un insieme di regole grammaticali e formule logiche per capire come funziona il linguaggio umano, come avrebbero voluto fare i linguisti computazionali fino a qualche decennio fa con la nostra mente. Il motivo di principio è che le stesse reti neurali profonde (*Deep Neural Network*) – la tecnologia di apprendimento automatico che è alla base degli LLM – presentano analogie con i fenomeni frattali e complessi¹⁴. E questo isomorfismo fra la macchina e il fenomeno che la macchina riproduce rende una illusione pensare di studiare la macchina per capire i fenomeni linguistici: il livello di complessità della macchina rimane lo stesso della nostra mente, analogo a quello del linguaggio e dei fenomeni che il linguaggio permette di descrivere: un livello di complessità sproporzionato rispetto ai nostri modelli espliciti basati su regole locali.

In conclusione, rimane ancora la domanda: *ma come facciamo, quindi, noi umani a comprendere un linguaggio che non è solo una sequenza lineare di parole (ancorché strutturata gerarchicamente in strutture grammaticali), ma rispecchia la frattalità, complessità e alta dimensionalità del mondo?* La risposta in realtà la sapevamo da tempo e forse noi linguisti computazionali dovremmo fare autocritica per aver creduto di potere affrontare un problema complesso come il linguaggio con degli strumenti analitici relativamente semplici e focalizzati sull’ambito locale della singola frase come le grammatiche, le reti semantiche, le ontologie e la logica formale, ma anche i modelli statistici basati su n-grammi. È vero che per decenni non abbiamo avuto alternative. Anche solo utilizzare delle grammatiche più potenti e attente al contesto (*context-sensitive* nella categorizzazione di Noam Chomsky) era ai limiti della portata

¹⁴ Jascha Sohl-Dickstein, “The boundary of neural network trainability is fractal” (2024). DOI: 10.48550/arXiv.2402.06184.

della capacità computazionale disponibile di allora, dato che una grammatica sensibile al contesto rende esponenziale la complessità del processo di analisi grammaticale della frase.

Tuttavia, sapevamo già che nel linguaggio c'era qualcosa di più, consapevolezza che abbiamo cercato di dimenticare per via delle limitazioni tecnologiche. Per anni nel corso di Sistemi Cognitivi nella Laurea magistrale in Informatica che tenevo all'Università di Torino, da un lato, spiegavo come costruire sistemi di elaborazione del linguaggio naturale con le regole dell'AI simbolica (grammatiche, logica, ontologie, ecc.), e, dall'altro, dicevo che le nostre capacità linguistiche sono parte della nostra "conoscenza tacita", concetto definito nel 1966 dal filosofo, economista e psicologo ungherese Michael Polanyi¹⁵, fratello del più famoso economista Karl Polanyi. La conoscenza tacita rappresenta quella parte della mente che riguarda il saper fare ed è acquisita tramite la pratica. Essa caratterizza attività motorie (come manipolare gli oggetti, andare in bicicletta), ma anche la nostra percezione, la nostra intelligenza emotiva e, fatto che abbiamo sottovalutato, anche capacità di base come parlare e ragionare. Noi impariamo a parlare e a ragionare prima di andare a scuola, prima di imparare le regole della grammatica, ma soprattutto anche senza mai dare un esame di logica. Viene definita 'tacita' perché scarsamente accessibile alla coscienza e non descrivibile in maniera esplicita con il linguaggio e, quindi, difficilmente formalizzabile con la matematica o la logica. È, tuttavia, la conoscenza tacita che ci permette di comprendere il linguaggio e di parlarlo, anche se è preclusa alla nostra capacità esplicita di capire come funziona. Dobbiamo accettare la nostra umiltà metodologica: non è sempre possibile costruire un modello formale di come funziona la nostra mente.

Allo stesso modo, ChatGPT e simili imparano a parlare senza avere bisogno di conoscere la grammatica: impara a tentativi, come i bambini, ma ha prestazioni infinitamente superiori ad ogni sistema di elaborazione del linguaggio naturale basato su regole grammaticali, logica e ontologie o n-grammi e le loro frequenze. Per la prima volta con gli LLM abbiamo riprodotto in maniera utilizzabile alcune capacità della nostra 'conoscenza tacita', ma abbiamo aperto il vaso di Pandora di un mondo che, se diventa

¹⁵ Michael Polanyi, *The Tacit Dimension*, Garden City, NY: Doubleday Anchor, 1966.

più prevedibile grazie alle macchine, allo stesso tempo ci ricorda di non essere sempre ‘spiegabile’ in termini di modelli espliciti. I successi del *Deep Learning* nel mondo della natura e nel mondo della nostra mente ci dicono che poter fare previsioni migliori sui fenomeni caotici e nei sistemi complessi ha un costo: rinunciare alla possibilità di comprendere come viene fatta la previsione. Per anni il *Machine Learning*, l’apprendimento automatico che è alla base dell’attuale ‘primavera dell’IA’ e a cui appartiene la tecnologia del *Deep Learning*, è stato criticato per essere una *Black Box*, cioè un modello ‘non trasparente’, che non si può analizzare per spiegare come ha preso una decisione o ha classificato un certo caso o fatto una determinata previsione.

Paradossalmente comincia ad emergere che, oltre al fatto che è impossibile dare spiegazioni nei sistemi caotici e complessi, non è neanche necessario farlo per fare previsioni. In tali sistemi, infatti, non si possono fare previsioni tramite modelli matematici, ma solo tramite approssimazioni o simulazioni. Anziché tentare di «abbracciare la complessità del mondo che va oltre la comprensione umana»¹⁶, gli esseri umani hanno finora cercato di ridurla, mentre questa complessità supera di gran lunga le leggi e i modelli che l’uomo inventa per spiegare l’universo, modelli formati da leggi e regole locali semplificate. Sia i modelli tradizionali che quelli creati dal *Machine Learning* sono rappresentazioni del mondo: il primo tipo è una rappresentazione che abbiamo creato basandoci sulla nostra comprensione; il secondo tipo di modelli è generato da una macchina che abbiamo creato: al confronto di quelli precedenti, questi nuovi modelli hanno una scala, contenuti e una struttura incommensurabili, ma sono modelli difficili se non impossibili da esplorare.

Nessuno ancora sa spiegare perché la tecnologia del *Deep Neural Network*, alla base dell’apprendimento automatico moderno, riesca a cogliere i battiti di ali di farfalla nei fenomeni caotici e le complessità dei sistemi nel mondo naturale e nella nostra cognizione. *Deep Learning* significa avere una rete neurale con una struttura a molti livelli, e questa profondità permette alla rete di apprendere una gerarchia di caratteristiche, dalle astrazioni a basso livello a quelle ad alto livello, di modellare le molteplici scale di interazione

¹⁶ David Weinberger, *Caos quotidiano. Un nuovo mondo di possibilità*, trad. di Massimo Durante, Torino: Codice Edizioni, 2020 [Orig. *Everyday chaos: Technology, complexity, and how we're thriving in a new world of possibility*, Harvard Business Review Press, 2019].

presenti nei sistemi complessi. Sistemi di *Deep Learning* come i *Transformers* (alla base degli *LLM*) utilizzano meccanismi di auto-attenzione per valutare l'importanza delle diverse parti della sequenza di input (le parole del testo), catturando dipendenze a prescindere dalla loro distanza. Questo li rende altamente efficaci per compiti come il modellamento del linguaggio, la traduzione e la previsione di sequenze complesse.

Ma è possibile che sia anche la dimensione enorme di questi sistemi (composti da centinaia di miliardi se non trilioni di parametri) a permettere di identificare più relazioni sottostanti nei dati, analizzando molti più dati su una scala molto più ampia di quanto possiamo fare noi – costretti nelle nostre aspettative di come i dati debbano essere collegati fra loro –, cogliendo invece interdipendenze più complesse anche grazie alla natura frattale di questi sistemi. Sistemi di tale taglia e con caratteristiche frattali possono creare così approssimazioni migliori, catturare schemi intricati, dipendenze e caratteristiche ad alta dimensionalità di sistemi complessi.

Stiamo uscendo dal labirinto, ma rimaniamo nell'oscurità dell'assenza di 'spiegazioni'.

Quali sono gli scopi dell'Intelligenza Artificiale?

Verso una spiegazione teleologica

Mattia Fumagalli

Libera Università di Bolzano

Roberta Ferrario

Istituto di Scienze e Tecnologie della Cognizione, CNR

INTRODUZIONE

Vari tipi di sistemi di Intelligenza Artificiale (IA) sono ormai presenti in molti ambiti. Troviamo queste tecnologie in campi come la finanza, la sanità, la gestione dei rischi, il sistema giudiziario, ma anche in settori più legati all'intrattenimento, come la generazione di immagini e la raccomandazione di contenuti multimediali. Attualmente queste tecnologie sono accessibili non solo a tecnici e specialisti, ma anche ad utenti non esperti/e, che possono utilizzarle in vari modi, ad esempio attraverso i loro smartphone. Tuttavia, nonostante questa ampia diffusione, i dubbi sull'affidabilità dell'IA rimangono forti¹. Buona parte di tale diffidenza deriva dall'opacità tipica dei processi decisionali che stanno alla base di queste tecnologie, che spesso riutilizzano conoscenze implicite immagazzinate in vasti dataset. In questo senso, diverse iniziative sono state intraprese per migliorare la comprensione da parte degli/lle utenti, attraverso la crescita dell'area di ricerca chiamata *eXplainable AI* (XAI)². Quest'ultima ha un duplice obiettivo: da un lato, incrementare la comprensione umana dei sistemi IA e, dall'altro, spiegare le motivazioni alla base delle loro decisioni, rispettando requisiti di 'interpretabilità' e 'affidabilità'.

¹ Amina Adadi and Mohammed Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)", IEEE access, 6 (2018): 52138-52160. DOI: 10.1109/ACCESS.2018.2870052.

² Waddah Saeed and Christian Omlin, "Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities", Knowledge-Based Systems, 263 (2023): 110273. DOI: 10.1016/j.knosys.2023.110273.

La necessità crescente di spiegazioni è testimoniata dall’emanazione di nuove normative, soprattutto all’interno dell’Unione Europea³, che richiedono che i principi che regolano le decisioni dei sistemi di IA siano espliciti, rendendo questi ultimi ‘spiegabili’. Tuttavia, la letteratura scientifica non sembra avere raggiunto un accordo su una definizione condivisa di ‘spiegazione’ e, per questo motivo, uno dei nodi cruciali è lo sviluppo di un quadro di riferimento per un’IA spiegabile. Alcune strategie, per esempio, si concentrano su una spiegazione a livello globale del funzionamento dell’IA, ricorrendo a tecniche di semplificazione o visualizzazione, come l’uso di alberi decisionali⁴. Altre, invece, puntano su spiegazioni localizzate, offrendo giustificazioni per ciascuna decisione presa dall’IA in relazione al contesto specifico⁵.

In questo articolo proponiamo la ‘spiegazione teleologica’ e le inferenze che da essa possono derivare come perno di una nuova prospettiva attraverso la quale affrontare il dibattito sulle spiegazioni nell’ambito della XAI, soprattutto grazie al suo valore normativo. Inizialmente descriveremo questa forma di spiegazione confrontandola con altre tipologie considerate rilevanti in letteratura. In seguito, analizzeremo le spiegazioni teleologiche nel contesto specifico dell’IA. Infine, mostreremo come tali spiegazioni, fondate su specifiche nozioni di ‘funzione’ e ‘scopo’, siano fondamentali per reintrodurre un criterio normativo capace di valutare se e in che modo un sistema o, più in generale, un artefatto di IA possa funzionare in maniera errata o inadeguata.

IDENTIFICARE LO SCOPO DI UN ARTEFATTO

Iniziamo considerando un esempio di artefatto classico di uso comune, come una sedia. Questo oggetto semplice, solitamente composto da una seduta, uno schienale e talvolta dei braccioli, ha uno scopo specifico, che è quello di offrire un posto su cui sedersi. Tuttavia, possiamo assumere che la sedia possa essere impiegata anche per altri obiettivi, come bloccare

³ Artificial Intelligence Act (2024). <https://www.artificial-intelligence-act.com/>.

⁴ Eoin M. Kenny et al., “Explaining Black-Box Classifiers Using Post-Hoc Explanations-by-Example: The Effect of Explanations and Error-Rates in XAI User Studies”, *Artificial Intelligence*, 294, (2021): 103459. DOI: 10.1016/j.artint.2021.103459.

⁵ Riccardo Guidotti et al., “A survey of methods for explaining black box models”, *ACM computing surveys (CSUR)*, 51, 5 (2018): 1-42. DOI: 10.1145/3236009.

una porta o ostacolare un passaggio, a seconda delle esigenze e del contesto. Detto ciò, risulterebbe comunque strano valutare la qualità di una sedia basandosi sulla sua capacità di bloccare una porta. Potremmo dire che una sedia non sia adatta per bloccare e tenere una porta chiusa, ma al contempo riconoscere che rimane una buona sedia, perfetta per sedersi. Inoltre, se la sedia non fosse efficace per bloccare la porta, non incolperemmo né il venditore né il designer della sedia; diversamente, se oscillasse ogni volta che la si usasse per sedersi, ci sentiremmo autorizzati/e a manifestare il nostro disappunto. Questo esempio evidenzia come valutare un artefatto semplice come una sedia sia solitamente un'operazione naturale e intuitiva, priva di ambiguità.

Ora prendiamo in considerazione un artefatto nel dominio dei sistemi di IA, come un motore di ricerca, come ad esempio Google, Bing, e simili, o un *Large Language Model* (LLM), come GPT, PaLM, LLaMa, ecc. Siamo di fronte a una situazione altrettanto chiara come quella della sedia? È altrettanto facile individuare lo 'scopo' di questi strumenti? Possiamo dire che un motore di ricerca funziona male se propone tra i primi risultati link sponsorizzati? Ha senso criticare un LLM perché fallisce in operazioni di ragionamento? O sarebbe come criticare una sedia perché non riesce a mantenere chiusa la porta? In questo contesto, data la complessità degli artefatti in questione, la loro evoluzione continua e i molteplici usi nei più svariati contesti, parlare di uno scopo principale sembra poco realistico. Non a caso, alcuni studiosi, riferendosi all'IA o ai sistemi informativi, utilizzano metafore che li paragonano a risorse, come l'elettricità⁶, piuttosto che a semplici artefatti.

Questo approccio appare ragionevole, specie per molti dei nuovi sistemi di IA, che possono essere personalizzati e utilizzati da chiunque per scopi *ad hoc*, talvolta nemmeno previsti in fase di progettazione. Basti osservare i vari tentativi di definire un approccio olistico alla valutazione degli LLM, in cui i compiti presi in esame spaziano tra i più disparati, come traduzione, generazione di codice, estrazione di dati, raccomandazioni in ambiti diversi. Tuttavia, possiamo fare riferimento anche a prodotti apparentemente più trasparenti e comuni, come i social network o, appunto, i motori di ricerca. Spesso si tende a considerare queste nuove tecnologie

⁶ Andrew Ng, *AI is the new electricity*. O'Reilly Media, Incorporated, 2018. <https://www.oreilly.com/radar/ai-is-the-new-electricity/>.

come neutrali, tanto che, ad esempio, l'intervento di moderatori nei social network è spesso mal tollerato o messo in discussione⁷.

Detto questo, determinare lo scopo di artefatti di IA dipenderebbe dal contesto d'uso o, più specificamente, dal compito per cui li si sta utilizzando. Ad esempio, un motore di ricerca potrebbe non essere ottimale per fornire informazioni, ma risultare eccellente per raccogliere dati dagli/le utenti o mostrare pubblicità per gli/le investitori/trici. Un LLM potrebbe essere ideale per generare codice, ma non per fornire raccomandazioni in ambito legale. Se seguiamo questa ipotesi, resta aperta – e molto complessa – la questione di capire come identificare un uso corretto per ciascuno di questi strumenti. La tesi che vorremmo sostenere, senza voler semplificare eccessivamente, è che, se l'uso viene valutato sempre rispetto a un contesto specifico, si rischia di giungere alla conclusione che non esiste un uso propriamente corretto o scorretto di questi prodotti, e quindi ogni uso è ammissibile. In altre parole, la loro neutralità ne impedisce la 'falsificabilità'. Ma non è proprio ciò che cerchiamo di evitare quando vogliamo garantire un controllo delle tecnologie in uso e la responsabilità nel loro utilizzo?

SPIEGAZIONE TELEOLOGICA

Le persone tendono ad attribuire scopi specifici agli oggetti. Tipicamente, queste attribuzioni assumono la forma di affermazioni come "le sedie servono per sedersi" o "le fotocamere degli smartphone servono per scattare foto e registrare video". Inoltre, spesso si ritiene che il fine di un oggetto spieghi *perché sia fatto in quel modo*. Le sedie, ad esempio, hanno sempre una seduta e uno schienale, una caratteristica normale per le sedie, proprio perché questa struttura risponde alla loro funzione di sostenere la schiena della persona e offrire una seduta confortevole. Si può dire quindi che una sedia ha una seduta e uno schienale proprio per sostenere chi si siede. Questo rappresenta un tipico esempio di 'spiegazione teleologica'⁸. La funzione della sedia, cioè permettere di sedersi comodamente, spiega perché

⁷ Shawn Walker, Dan Mercea, and Marco Bastos, "The Disinformation Landscape and the Lockdown of Social Platforms", *Information, Communication & Society*, 22, 11 (2019): 1531-1543. DOI: 10.1080/1369118X.2019.1648536.

⁸ Sehrang Joo, Sami R. Yousif, and Joshua Knobe, "Teleology beyond explanation", *Mind & Language*, 38, 1 (2023): 20-41. DOI: 10.1111/mila.12393.

la sedia abbia le sue caratteristiche e perché queste includano sempre una seduta e uno schienale.

L'importanza di questo tipo di spiegazione⁹ si deve principalmente al suo ruolo normativo. Per poter identificare e analizzare deviazioni e malfunzionamenti è necessario introdurre un concetto di corretto funzionamento. Pensiamo ancora alle caratteristiche delle sedie: sono progettate per sostenere e offrire confortevolezza alla persona seduta. Si può dire che una sedia non funzioni bene, o che sia difettosa, proprio se non rispetta questo scopo. Ad esempio, se una sedia non ha una struttura stabile, si allontana dalla norma del buon funzionamento e potrebbe essere considerata inadeguata o con un difetto di progettazione. Allo stesso modo, se una sedia è scomoda o mal conformata, non risponde al suo scopo di offrire confortevolezza, rivelando una deviazione dal suo corretto funzionamento.

Le spiegazioni teleologiche, concentrandosi sul fine e sul funzionamento corretto degli oggetti, permettono una valutazione normativa delle loro caratteristiche. Anche se questo tipo di ragionamento è cruciale in diversi contesti, soprattutto quando è richiesto di valutarne la funzionalità, in altri ambiti viene considerato meno affidabile. Un motivo è che, rispetto alle spiegazioni causali, le spiegazioni teleologiche invertono l'ordine temporale tra *explanandum*, cioè il fenomeno da spiegare (ad esempio, 'i veicoli di emergenza usano sirene e luci lampeggianti'), ed *explanans*, cioè l'elemento esplicativo (ad esempio, 'per segnalare la loro presenza e aprirsi un varco nel traffico, riducendo il tempo di percorrenza'). Gli eventi menzionati nell'*explanans* sono successivi a quelli dell'*explanandum*, cosa che non accade nelle spiegazioni causali, dove le cause precedono sempre gli effetti. Tale struttura introduce una certa incertezza nelle spiegazioni teleologiche, poiché molti fini ed effetti funzionali potrebbero non realizzarsi mai. Pensiamo al contesto delle scienze naturali, dove i fenomeni vengono spiegati cercandone le cause, che precedono sempre gli effetti. Qui, il ruolo centrale dell'esplicazione non è normativo, ma predittivo, ovvero l'individuazione delle cause di un fenomeno è legata all'importanza di poterne prevedere il comportamento in futuro. In ambito scientifico, si può dire che per comprendere un fenomeno dobbiamo essere in grado di prevedere la sua manifestazione. In questo senso, tornando all'esempio della sedia, le caratteristiche dell'artefatto non vengono spiegate

⁹ George Frederick Schueler, *Reasons and purposes: Human rationality and the teleological explanation of action*, Oxford: Clarendon Press, 2003.

guardando al fine, ma al contrario, il fatto che la sedia sia comoda è spiegato dalle sue caratteristiche (seduta e schienale, ad esempio). In altre parole, il fatto che le sedie abbiano una seduta e uno schienale spiega il loro essere adatte a fornire supporto e a permettere di sedersi comodamente.

D'altro canto, quando è possibile risalire ai/lle progettisti/e, l'incertezza attribuita alla spiegazione teleologica non sussiste. Al contrario, il ruolo della spiegazione teleologica diventa essenziale per comprendere perché un oggetto abbia o debba avere certe caratteristiche.¹⁰ In questo caso, il fine per cui un oggetto è stato progettato diventa la sua 'funzione propria', e fa sempre riferimento alle intenzioni di chi l'ha progettato, realizzato o posizionato con l'obiettivo di produrre l'effetto previsto, ossia la funzione dell'artefatto¹¹.

In questo senso, le spiegazioni teleologiche possono essere interpretate come un tipo di spiegazione causale. Tornando all'esempio della sedia, se spieghiamo la sua forma dicendo che deve servire a sedersi, l'*explanans* si riferisce a qualcosa che segue invece di precedere l'*explanandum*. Tuttavia, questo problema è apparente, poiché l'*explanans* non guarda semplicemente a un effetto futuro, ma implicitamente si riferisce a qualcosa di precedente: l'intenzione del/la progettista di 'riprodurre' la sedia con lo scopo specifico di permettere alle persone di sedersi comodamente.

FUNZIONE PROPRIA

Secondo la prospettiva teleologica (principalmente intenzionalista¹²) che prendiamo qui maggiormente in considerazione, la 'funzione propria' di un artefatto si riferisce allo scopo specifico per cui è stato progettato o creato da un agente. Questa funzione può essere vista come l'effetto di una selezione intenzionale che di solito rimane coerente nel tempo con l'uso per cui è stata originariamente concepita, indipendentemente dal contesto. Ad esempio, una sedia usata intenzionalmente per colpire qualcuno diventa

¹⁰ Mattia Fumagalli, and Roberta Ferrario, "Representation of Concepts in AI: Towards a Teleological Explanation", Proceedings of the Joint Ontology Workshops 2019 (JOWO), vol. 2518, 2019. <https://ceur-ws.org/Vol-2518/paper-CAOS2.pdf>

¹¹ Si noti che, in un'accezione più ampia di artefatto, possono essere considerati artefatti anche oggetti rinvenuti nell'ambiente e utilizzati per uno scopo, come un sasso che viene utilizzato come fermacarte. In questo caso, il riferimento è alle intenzioni di chi usa l'oggetto per ottenere un certo effetto.

¹² Marc Artiga, "A dual-aspect theory of artifact function", *Erkenntnis*, 88, 4 (2023): 1533-1554. DOI: 10.1007/s10670-021-00414-9.

un'arma, poiché l'utente l'ha scelta per quello scopo in quella situazione, ma possiamo comunque assumere che la sua funzione rimanga quella di permettere alle persone di sedersi, poiché questo è l'effetto per cui è stata progettata e acquistata. Tuttavia, questa coerenza funzionale è più difficile da stabilire in artefatti complessi come quelli utilizzati nell'ambito dell'intelligenza artificiale, dove spesso esistono molteplici intrecci di funzioni che si rifanno a vari agenti (progettista, utente...), ciascuno con scopi diversi. In questo senso, si può osservare che il concetto di funzione di un artefatto varia a seconda del contesto. Ad esempio, talvolta le intenzioni dell'utente possono avere la precedenza su quelle del/la progettista. Proprio per questo caso specifico degli artefatti più complessi, è utile chiarire cosa possa significare 'funzione' in una prospettiva teleologica e differenziare tra diversi tipi di funzione.

Considerando che un'analisi dettagliata su come il concetto di funzione sia utilizzato in diversi ambiti¹³ supera lo scopo di questo articolo, è necessario chiarire le ipotesi che adottiamo. Qui interpretiamo la funzione in senso strettamente eziologico, ovvero *la funzione di qualcosa è ciò per cui quella cosa è stata selezionata*. In altre parole, in questo contesto, assumiamo che attribuire una funzione a un artefatto o a una sua caratteristica significhi riferirsi agli effetti che spiegano perché quella caratteristica sia stata selezionata. Seguendo l'approccio proposto anche da Marc Artiga¹⁴, possiamo assumere che F sia una funzione propria di un artefatto α se:

- i. α realizza efficacemente F quando viene utilizzato da alcuni agenti;
- ii. α è progettato come un artefatto di tipo A , dove gli artefatti di tipo A vengono riprodotti perché realizzano F .

Per la nostra discussione questa descrizione, anche se approssimativa, è sufficiente per chiarire che, per identificare se una funzione di un artefatto è una 'funzione propria', occorre esaminare principalmente il tipo di artefatto a cui è associata, ovvero la sua storia e perché quel tipo di artefatto viene solitamente riprodotto¹⁵, per soddisfare quale scopo. Anche questa definizione fa riferimento sia alla progettazione sia all'uso dell'artefatto.

¹³ All'interno della prospettiva teleologica sono presenti diverse sfumature.

¹⁴ *Ibidem*.

¹⁵ Il ruolo svolto dalla riproducibilità per gli artefatti è analogo a quello della selezione naturale per gli organismi: solo le caratteristiche che permettono all'organismo di sopravvivere in modo più efficiente nell'ambiente vengono selezionate dall'evoluzione e riprodotte nella specie.

In questo senso, possiamo dire che *permettere alle persone di sedersi* è una funzione propria della sedia α , poiché (i) α realizza efficacemente F – permettere di sedersi – quando viene usata da qualcuno, e (ii) α è progettata come un artefatto del tipo SEDIA e le SEDIE sono artefatti riprodotti per *permettere alle persone di sedersi*. Di conseguenza, le condizioni (i) e (ii) permettono di identificare il corretto funzionamento. Ad esempio, se α soddisfa una funzione F_i quando viene usata da un agente, ma l'artefatto appartiene a un tipo riprodotto per una funzione diversa F_j , è illegittimo dire che α funziona correttamente. In modo analogo, se l'artefatto appartiene a un tipo riprodotto per F_i e non soddisfa F_i quando viene usato da alcuni agenti, è ragionevole dire che l'artefatto non funziona correttamente. Infine, quando α non soddisfa una funzione F_i se usato da un agente, ma appartiene a un tipo riprodotto per una funzione diversa F_j , è illegittimo dire che α non funzioni correttamente.

Ora, queste ipotesi con l'esempio della sedia funzionano facilmente. In questo caso, infatti, abbiamo una chiara dipendenza dal tipo di appartenenza, in cui ogni artefatto individuale ha una funzione propria associata al tipo a cui appartiene¹⁶. Una sedia è fatta per sedersi e le sedie, come tali, sono fatte per sedersi. Ciò non esclude che l'artefatto sia usato o progettato secondo una funzione che non è, o non è completamente, associata al tipo a cui appartiene. Ad esempio, una sedia può essere usata per bloccare una porta o progettata come un modello da esposizione d'arte, senza l'intento che qualcuno ci si sieda. Tuttavia, anche in questi casi in cui la sedia non viene usata o non può essere usata per sedersi, possiede ancora quella funzione, poiché appartiene al tipo SEDIA. Si suppone, in ogni caso, che quella sedia abbia la funzione propria del suo tipo.

Diversamente, nel caso di artefatti per i quali non esiste una chiara dipendenza dal tipo, i criteri per identificare le funzioni proprie non possono essere semplici come quelli che abbiamo fornito. Cosa succede se utilizzo con successo un artefatto che non posso confrontare con altri artefatti esistenti o associare a un tipo specifico di artefatto? E se utilizzassi un artefatto che non posso paragonare ad altri artefatti esistenti o associare a una specifica categoria di artefatti? E se un artefatto fosse progettato per svolgere una funzione particolare che nessun altro tipo di artefatto esistente potrebbe soddisfare?

¹⁶ Simon J. Evnine, *Making objects and events: A hylomorphic theory of artifacts, actions, and organisms*, Oxford: Oxford University Press, 2016.

NEGOZIAZIONE DI FUNZIONI

Oggi, come abbiamo già avuto modo di illustrare, molti degli artefatti che usiamo hanno caratteristiche molto diverse rispetto a quelli più tradizionali, come le sedie. Questi nuovi oggetti sono molto più difficili da definire e spesso appaiono come oggetti semi-finiti che possono, in qualche modo, essere completati dagli/le utenti. Diverse tecnologie etichettate come *Sistemi Informativi* o *Sistemi di IA* rappresentano casi esemplari in tal senso. La mancanza di un accordo su cosa sia un sistema di IA dimostra chiaramente come per questi oggetti manchi una relazione di dipendenza di tipo, fondamentale per stabilire il corretto funzionamento secondo la definizione fornita, come avviene nel caso di artefatti come mobili o utensili fisici.

Quindi, sembra esserci una questione aperta: *qual è la funzione propria di un sistema di IA?* Per illustrare questo punto, consideriamo le sfide affrontate nel tentativo di definire regolamentazioni (AI Act) per l'IA all'interno dell'Unione Europea (UE). Questa mancanza di dipendenza di tipo è anche evidenziata dalla difficoltà nel determinare cosa certi tipi di oggetti artificiali e informatici siano 'supposti fare'.

Prendiamo, per esempio, la 'funzione di trovare informazioni' che corrispondano a una ricerca. È questa una funzione peculiare di artefatti come i motori di ricerca o di artefatti come i *chatbot*? Oppure, qual è la funzione degli algoritmi di apprendimento profondo? Sono riprodotti per 'riassumere testi', per la 'visione artificiale', per il 'riconoscimento vocale'? Ancora, qual è la funzione delle ontologie in ambito informatico? Sono riprodotte per 'abilitare l'interoperabilità' o per 'abilitare servizi di ragionamento'? Diremmo che in tutti questi casi non c'è una risposta completamente sbagliata o corretta. Questi tipi di artefatti sono spesso progettati per consentire funzioni multiple e hanno naturalmente funzioni che derivano da altri tipi di artefatti o che si sovrappongono con quelle di altri tipi. In questo caso, le funzioni non sono legate a una dipendenza di tipo, ma dipendono in gran parte dal contesto e dalle intenzioni degli agenti coinvolti nel loro uso e progettazione.

Come possiamo parlare quindi di una 'funzione propria' in questo scenario più complesso? Una proposta, che accenniamo in questa sede, ma che abbiamo in programma di sviluppare in un lavoro futuro, può essere fornita con una modifica delle condizioni presentate nella sezione

precedente, ammettendo che il ‘corretto’ funzionamento di un artefatto in IA può essere valutato in base a un ‘accordo’ tra le intenzioni dell’utente/ utilizzatore/trice dell’artefatto e quelle del/la progettista/ideatore/trice¹⁷. Questa proposta si basa sul fatto che non esiste una definizione concordata di IA, e che l’IA possa essere precisata solo pragmaticamente. Se adottiamo una definizione pragmatica di sistema di IA, ovvero come “qualsiasi sistema comunemente definito come sistema di IA o sistema di apprendimento automatico”, possiamo considerare gli LLM, per esempio, come un sottoinsieme di questi sistemi. Se un/a utente utilizza effettivamente un LLM ‘per tradurre un testo’ e tale funzione era prevista dal/la progettista, questo è sufficiente per affermare che quest’ultima è una funzione propria.

In realtà, la situazione è molto più complessa. La relazione tra l’insieme di ‘sistemi di IA’ e l’insieme di ‘funzioni di IA’ è multi-a-molti. Una funzione di classificazione, ad esempio, può essere implementata efficacemente tramite un albero decisionale o una rete neurale. Allo stesso modo, la funzione di restituire informazioni date una specifica *query* può essere gestita sia da *chatbot* sia da motori di ricerca. Ancora, la stessa rete neurale può essere utilizzata per funzioni multiple, come ‘classificazione del testo’ o ‘previsione delle serie temporali’. Inoltre, l’insieme delle funzioni che possono essere associate ai sistemi di IA è ampio e non sempre facile da definire. Quante funzioni possono essere associate agli LLM? Riassumere, tradurre, prevedere, ecc.? Cosa significa simulare, raffinare e orientare conversazioni simili a quelle umane o generare idee commerciali?

Infine, un punto molto importante. Questi artefatti sono spesso usati per funzioni non previste dai/lle loro progettisti. Ad esempio, i motori di ricerca possono essere manipolati per scopi di marketing e gli LLM possono essere utilizzati per simulare ragionamenti o generare contenuti specifici di dominio, come quello medico, per cui non sono addestrati. Inoltre, a volte i/le progettisti/e creano questi artefatti con funzioni che non sono destinate a essere adottate direttamente dall’utente, ad esempio, per raccogliere diversi tipi di dati o per fornire visibilità a investitori/trici e pubblicitari/e.

In questo scenario articolato, nell’identificare il funzionamento proprio di un artefatto di IA, non si ha a che fare con una sola funzione,

¹⁷ Si noti che il riferimento di tipo che alcuni artefatti hanno può essere visto anche come il risultato di un accordo stabilito nel tempo. Il fatto che per alcuni artefatti ciò non sia possibile è naturale e dipende spesso dal loro grado di novità e complessità.

ma con più funzioni che possono essere state attribuite dai/lle progettisti/e o dagli/lle utenti. Tutte le funzioni che soddisfano entrambe le condizioni condivise sia da utenti e disegnatori/trici saranno funzioni proprie, ovvero potranno essere usate per valutare il corretto funzionamento dell'artefatto.

Le funzioni rimanenti, cioè quelle progettate ma non utilizzate e quelle adottate pur non essendo state progettate, possono essere considerate improprie. Le intenzioni dei/lle progettisti/e e degli/lle utenti, dunque, influenzano naturalmente il numero di funzioni proprie di un dato artefatto che, grazie a un processo di negoziazione tra tali agenti, può anche variare nel tempo e attraverso interazioni successive. Un/a progettista può consentire o vietare l'uso di funzioni che non aveva considerato, ma che sono state 'scoperte' da alcuni/e utenti. Allo stesso modo, un utente può o meno essere a conoscenza di una funzione progettata e decidere se ammettere o meno funzioni che non ha utilizzato.

CONCLUSIONE

Per concludere, *in che senso possiamo dire che la spiegazione teleologica possa tornare utile nel contesto dell'IA?* Abbiamo visto che secondo la logica della spiegazione teleologica, le caratteristiche di un artefatto vengono spiegate attraverso lo scopo per cui è stato progettato. In questo modo, la spiegazione assume un carattere normativo piuttosto che predittivo e consente l'attribuzione di una funzione propria all'artefatto. Tuttavia, gli artefatti di IA possono essere utilizzati per molteplici scopi e permettono agli/lle utenti maggiore libertà riguardo al modo in cui possono essere impiegati. Per questo motivo, abbiamo sfruttato la spiegazione teleologica come una strategia di identificazione delle funzioni proprie che prevede una fase di negoziazione delle funzioni, in cui sia il/la progettista(i/e) che l'utente(i) rendono esplicite le proprie funzioni intese, cioè quelle che sono state espressamente progettate e quelle che vengono utilizzate rispettivamente.

In questo senso, la nostra proposta può essere vista come complementare a una strategia di progettazione, poiché l'attribuzione di funzioni dovrebbe idealmente iniziare nella fase di progettazione. Naturalmente, la negoziazione può avvenire anche dopo che l'artefatto è stato prodotto e utilizzato. Tuttavia, in tal caso, l'aggiunta o la rimozione di alcune funzioni può costituire la progettazione di un nuovo artefatto.

Se tale artefatto sia semplicemente una versione aggiornata dell'artefatto originale o un artefatto genuinamente nuovo, è una questione ontologica che richiede una discussione apposita. Inoltre, il processo di negoziazione mediante il quale vengono identificate le funzioni proprie va considerato un processo dinamico, di cui gli/le utenti sono elementi costitutivi.

In questo contesto, la spiegazione non va considerata come costitutiva solo della fase di progettazione (*ex-ante*), ma anche come una forma di valutazione che può essere iterativamente informata dal feedback fornito attraverso l'uso continuo degli/le utenti. Questo aspetto iterativo consentirebbe di rilassare o restringere il set di funzioni attribuite all'artefatto (funzioni recentemente attribuite o funzioni emergenti).

Da un lato, l'idea proposta di riutilizzo delle dinamiche della spiegazione teleologica dovrebbe consentire di spiegare il corretto funzionamento di un artefatto facendo riferimento a quelle funzioni accordate sia dai/le progettisti/e che dagli/le utenti, permettendo così di stabilire se un artefatto di IA funzioni correttamente o sia difettoso (ad esempio, valutando l'artefatto considerando solo le funzioni su cui c'è accordo). Dall'altro, la fase di negoziazione può essere utilizzata come valutazione del grado di trasparenza e controllo dell'artefatto e, attraverso cicli iterativi di negoziazione, può garantire che alcune funzioni sulle quali non c'è accordo siano permesse o vietate.

Un obiettivo ulteriore di questa breve discussione è quello di suggerire come una nuova versione adattata della spiegazione teleologica può essere sfruttata per promuovere e valutare (i) il controllo degli artefatti in questione, prevenendo e identificando usi scorretti del sistema, e (ii) la trasparenza, favorendo la consapevolezza dell'utente sulle funzioni del sistema. Tuttavia, possiamo elencare molte altre motivazioni alla base della nostra proposta. Una che ci preme sottolineare è che, nel contesto della spiegazione teleologica, la responsabilità degli agenti coinvolti nella progettazione/utilizzo dell'artefatto è centrale. Essi/e svolgono infatti un ruolo attivo nella determinazione del corretto funzionamento dell'artefatto, in modo tale da poter identificare una sorta di *co-design* dell'artefatto stesso.

In termini più tecnici, tale *co-design* può essere visto anche come il risultato della 'elicitazione dei requisiti (funzionali)' e della specificazione dell'artefatto attraverso un processo continuo e iterativo di negoziazione.

RINGRAZIAMENTI

Roberta Ferrario si è avvalsa del supporto del progetto BRIO, finanziato dal Ministero dell'Università e della Ricerca (MUR) attraverso lo schema PRIN (Progetto nr. 2020SSKZ7R), del progetto SMARTEST (Progetto nr. 202223E8Y4X), finanziato dal MUR attraverso lo schema PRIN e del progetto FAIR, (Progetto nr. PE 00000013), finanziato dal MUR.

PARTE II

LOGICA, EPISTEMOLOGIA ED ETICA DELL'INTELLIGENZA ARTIFICIALE

BRIO: il ruolo della logica nella costruzione di un'Intelligenza Artificiale equa

Giuseppe Primiero

Università degli Studi di Milano

INTRODUZIONE

Il problema della valutazione della trasparenza di sistemi computazionali la cui struttura non sia accessibile (per esempio, per motivi legati alla privacy o al diritto industriale) o, anche laddove accessibile, non analizzabile per una questione essenziale di complessità strutturale, è un tema ormai centrale. Tipicamente, nella letteratura relativa alla c.d. *eXplainable Artificial Intelligence* (XAI), questa domanda è formulata in termini che potremmo definire ‘ontologici’: quali proprietà deve avere, o devono essere ricostruite, perché un sistema possa essere considerato trasparente? Gli approcci tipici sono quelli che distinguono trasparenza dettata da *low-level*, *middle-level* e *high-level features* nel modello¹.

Nel caso della nostra ricerca, partiamo invece da una domanda di natura ‘epistemologica’: in che condizioni possiamo fidarci dei risultati di un sistema computazionale opaco? Stiamo quindi proponendo uno slittamento metodologico: trattare un problema ontologico in termini epistemologici che traduce il problema XAI in un problema TAI (*Trustworthy Artificial Intelligence*). E nel cercare le condizioni di affidabilità che un sistema deve garantire, stiamo in realtà formulando una domanda di tipo ‘etico’, di profonda rilevanza per i contesti umani, ovvero: *quali proprietà consideriamo essenziali rispetto ad una decisione del nostro sistema computazionale, tale che il loro rispetto garantisca la nostra fiducia nei suoi confronti?* Questo problema etico, che

¹ Andrea Apicella et al., “Explanations in Terms of Hierarchically Organised Middle Level Features”, Proceedings of the 2nd Italian Workshop on Explainable Artificial Intelligence, co-located with 20th International Conference of the Italian Association for Artificial Intelligence (AIxIA 2021), 2021, pp.44-57, <https://ceur-ws.org/Vol-3014/paper4.pdf> e i riferimenti bibliografici ivi presenti.

potremmo caratterizzare come un tema di RAI (*Responsible AI*), si traduce per noi nel problema della fiducia: un sistema che si comporta in maniera equa può essere considerato affidabile. Nella letteratura questa nozione di ‘equità’ è formulata rispetto a categorie protette (cioè una categoria di utenti definita da una proprietà particolare protetta dalle autorità per ragioni politiche o di policy) secondo diverse definizioni²:

- i. parità demografica: agenti con e senza attributi protetti dovrebbero ottenere le stesse ricompense attese.
- ii. equità controfattuale: sia in uno scenario fattuale che in uno scenario controfattuale, dove l’unica differenza è se gli attributi protetti valgono per un agente, gli agenti dovrebbero ottenere le stesse ricompense attese.
- iii. parità statistica condizionata: all’interno di un gruppo di agenti caratterizzati da un fattore legittimo che influenza le ricompense, gli agenti con e senza attributi protetti dovrebbero ottenere le stesse ricompense attese.

Partiti da una domanda di natura ontologica (*cosa è un sistema trasparente?*) siamo passati ad una domanda di natura epistemologica (*come facciamo a sapere che un sistema è affidabile?*) per arrivare ad una domanda di natura etica (*quali sono le proprietà che un sistema considerato equo deve soddisfare?*). Tuttavia, esiste un ultimo passaggio possibile, ovvero la traduzione di questo problema ontologico-epistemologico-etico nei termini di una soluzione ‘logica’. Dal punto di vista logico-formale, la domanda iniziale di trasparenza diventa: *come verificiamo che un sistema computazionale opaco produca dei risultati equi sufficienti a far considerare il sistema affidabile da parte di un utente?* In ultima analisi, cerchiamo di rispondere ad una domanda di logica formale, di verifica e di implementazione di tale verifica in un processo di controllo.

Nel seguito di questo breve articolo, desidero illustrare la metodologia del progetto BRIO (*Bias, Risk and Opacity in AI*)³, un insieme di risultati teorici, della loro applicazione e implementazione anche con il progetto

² Alessandro Castelnovo et al., “The zoo fairness metrics in machine learning”. CoRR, abs/2106.00467, 2021.

³ Progetto BRIO. <https://sites.unimi.it/brio>.

SMARTEST (*Simulation of Probabilistic Systems for the Age of the Digital Twin*)⁴ per rispondere a questa domanda logica.

DUE LIVELLI DI RESPONSABILITÀ PER L'INTELLIGENZA ARTIFICIALE

L'idea che l'Intelligenza Artificiale (IA) debba essere qualificata in termini di responsabilità sta velocemente consolidandosi nella letteratura e nella pratica. La nozione di 'IA responsabile' varia e risponde a diversi criteri, includendo principi generali tra i quali: il rispetto della privacy, l'identificazione e la mitigazione di pregiudizi, la trasparenza, la rendicontabilità delle decisioni, la predizione di comportamenti che si allontanano da standard determinati e considerati accettabili, la valutazione del rischio⁵. Comune a tutti questi problemi specifici è l'identificazione delle figure e delle strutture atte a garantire la soddisfacibilità dei criteri scelti.

In questo contesto, troppo ampio per essere analizzato qui nel dettaglio, ci facciamo una domanda specifica, relativamente al problema della verifica di un sistema computazionale opaco rispetto all'equità dei suoi risultati. Per spiegare questo più specifico aspetto, il problema della responsabilità può essere considerato da due punti di vista. Da una parte, c'è il problema di allenare dei sistemi di apprendimento automatico in maniera tale che producano risultati equi, e di identificare chi è responsabile di questa fase. C'è inoltre un secondo problema, che è quello di controllare che un algoritmo di IA in funzione produca risultati equi. *Chi è responsabile, chi controlla che la fase di allenamento e la fase validazione siano state adeguatamente svolte e risultino sufficientemente complete per un uso sicuro dei sistemi di IA?*

Questa distinzione è interessante tanto da un punto di vista tecnico, quanto da quello della divisione della responsabilità e della definizione di strumenti non solo di natura tecnica ma anche di natura legale. Alla luce di questa distinzione, ci sono essenzialmente due livelli di responsabilità. Il primo livello consiste nel chiedersi di chi sia la responsabilità della creazione di un sistema equo: dalla selezione dei dati di sviluppo, al design delle funzioni di ottimizzazione e degli algoritmi, oltre che della fase di allenamento e validazione. Questo livello appartiene essenzialmente allo sviluppatore, inteso

⁴ Progetto SMARTTEST. <https://sites.unimi.it/smartest/>.

⁵ Virginia Dignum, *Responsible Artificial Intelligence*, Switzerland: Springer, 2019; Sray Agarwal e Shashin Mishra, *Responsible AI*, Switzerland: Springer, 2021.

come la categoria complessa di tutte le figure che sono coinvolte nel processo di costruzione di un sistema di apprendimento automatico.

Il secondo livello è quello sul quale noi ci concentriamo. Riguarda la responsabilità di controllo sui risultati di un sistema di apprendimento automatico che permetta di accertare che la responsabilità di primo livello sia stata soddisfatta, ovvero che un sistema sia stato allenato e sviluppato in maniera tale da garantire risultati equi. La verifica in questo secondo senso, e in particolare nel contesto del controllo di comportamenti equi e della mitigazione del *bias*, inteso come pregiudizio, da parte di sistemi di IA, può essere implementata in tutte le fasi della costruzione del modello e del ciclo di vita dei sistemi di ML:

- i. pre-processing*: consiste nel trattare i dati di addestramento per minimizzare la presenza di bias nell'output;
- ii. in-processing*: si affrontano i *bias* che possono emergere dal processo di apprendimento, inclusi sia quelli che originano dai dati sia quelli dovuti ai meccanismi di addestramento;
- iii. post-processing*: si considerano i risultati di un modello e si applicano alcune trasformazioni per eliminare i *bias*⁶.

Di chi sia la responsabilità di questa fase di controllo dipende principalmente da una scelta politica rispetto alla regolamentazione dello sviluppo e distribuzione anche economica dell'IA. Nello scenario internazionale corrente si individuano diversi ruoli associati a distinte strategie politiche, e l'*EU AI Act*⁷ ha certamente qualificato la strategia europea come quella che più delle altre identifica nel ruolo del legislatore il responsabile per la fase di controllo⁸.

La nostra ricerca è dunque pensata per supportare l'analisi necessaria alla responsabilità di secondo livello, quella del controllo o verifica. Questa caratterizzazione ci pone dunque in un ruolo nel quale non possiamo assumere conoscenza o accesso ai modelli e ai dati usati nella fase di allenamento e tuning (responsabilità di primo livello). Il modello, in quanto spesso sviluppato a

⁶ Max Hort et al., "Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey", *ACM Journal on Responsible Computing*, 1, 2 (2024): 1-52. DOI: 10.1145/3631326.

⁷ Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024. <http://data.europa.eu/eli/reg/2024/1689/oj>.

⁸ Emmanuel Pernot-Leplay, "Global AI & Tech Policies, The AI Regulation Dilemma: A China, EU & U.S. Comparison". <https://pernot-leplay.com/ai-regulation-china-eu-us-comparison/>; Andrea Baronchelli, "Shaping new norms for AI", *Philosophical Transactions of the Royal Society B* 379.1897 (2024): 20230028. DOI: 10.1098/rstb.2023.0028.

livello industriale, è di fatto inaccessibile al controllore. E in questa condizione assumiamo perciò la condizione di opacità completa rispetto a questi elementi.

Sotto queste condizioni, sviluppiamo un metodo *post-hoc*, ovvero una strategia che si sviluppa a posteriori e che assume come altamente probabile (se non inevitabile nella maggioranza dei casi) la condizione di 'inspiegabilità' dei modelli.

In presenza, dunque, di un sistema opaco, sfruttiamo però la capacità di osservare e valutare il comportamento del modello in analisi sui dati di test attuali. Prendiamo dunque come oggetto di analisi la relazione input-output del sistema per estrarre delle osservazioni sul suo comportamento e formulare delle analisi.

LA STRATEGIA BRIO

La strategia di verifica *post-hoc* BRIO assume un principio teorico molto semplice: giudicare l'affidabilità di un comportamento di un sistema di IA richiede la valutazione della sua robustezza in termini statistici insieme al rispetto di un criterio di valutazione. Chiameremo nel seguito questo principio la "*Affidabilità-con-BRIO = Robustezza rispetto ad un criterio di valutazione*".

Questo principio di affidabilità riformula la classica nozione di verifica formale di sistemi computazionali deterministici. Per questi ultimi, storicamente, il criterio di valutazione è stata la traduzione formale della specifica del sistema, ovvero del suo comportamento atteso formulato in maniera coerente, completa ed esaustiva.

Rispetto alla specifica formale, si controlla la correttezza del comportamento teorico del programma come determinato dalla corrispondente macchina astratta con gli strumenti deduttivi della logica formale. E poi, rispetto alla traduzione della specifica formale nella sua implementazione, si verifica il comportamento pratico come osservato in fase di *testing* e secondo criteri induttivi⁹.

⁹ Per una breve riproposizione della distinzione tra correttezza matematica, correttezza fisica e classi di errori computazionali si veda ad esempio: Nicola Angius et al., "The Philosophy of Computer Science", The Stanford Encyclopedia of Philosophy (Summer 2024 Edition), a cura di Edward N. Zalta e Uri Nodelman. <https://plato.stanford.edu/archives/sum2024/entries/computer-science/>; Giuseppe Primiero, *On the foundations of computing*. Oxford: Oxford University Press, 2019.

Nel caso dei sistemi di IA opachi, alcuni adattamenti sono ovviamente necessari. In primo luogo, si guarda a sistemi non-deterministici, il cui output osservato è statisticamente caratterizzato all'interno di una classe di output possibili e per una batteria di esecuzioni numericamente determinata. In secondo luogo, la specifica non è disponibile, non è solitamente completa e comunque di non facile estrazione. Questo è ancora una volta dovuto al problema dell'opacità dei sistemi di IA, in particolare in relazione alla loro (i) 'spiegabilità', cioè l'impossibilità di determinare cause con metodi controfattuali e interventisti; e (ii) 'affidabilità', cioè la verificabilità rispetto ad un fenomeno atteso¹⁰.

La nostra verifica di 'Affidabilità-con-BRIO' usa dunque due termini di confronto specifici: (i) un'espressione formale che denoti il comportamento osservabile di un sistema non-deterministico in un numero finito di esecuzioni, possibilmente sotto l'assunzione che la distribuzione di proprietà del suo modello sia non-nota; (ii) un modello trasparente del comportamento atteso, normativamente o eticamente desiderabile per il modello osservato sui dati offerti in input. Abbiamo così un modello opaco, del quale possiamo osservare il comportamento, e richiediamo la formulazione di un modello trasparente che descriva sotto le stesse condizioni iniziali un comportamento desiderabile. In altre parole: dati gli stessi dati in entrata, cosa ci aspettiamo o cosa vorremmo che il sistema facesse? Quali sono i criteri che riteniamo essenziali, necessari, non negoziabili in output? Il secondo modello trasparente sarà dunque il metro di valutazione del primo modello osservato, e il processo di verifica formale consisterà nel misurare una distanza tra i due modelli. Tale misura varrà ovviamente sotto assunzione di un margine di accettabilità, cioè consisterà nel misurare quanto i risultati del modello opaco divergano dai risultati di un modello trasparente sotto le stesse condizioni di comportamento iniziale.

La costruzione del modello teorico di riferimento e la definizione di una funzione di misura dell'affidabilità del modello opaco come distanza del comportamento osservato da quello atteso sono stati illustrati in modelli logici da un punto di vista computazionale, teoretico-dimostrativo

¹⁰ Per queste due nozioni si veda in particolare: Alberto Termine e Giuseppe Primiero, "Causality Problem in Machine Learning Systems", in *The Routledge Handbook of Causality and Causal Methods*, a cura di Federica Russo, Phyllis Illari, Routledge, 2024.

e semantico¹¹. Questi sistemi formali ci permettono di avere una definizione precisa della nozione di 'affidabilità' risultante dalla robustezza di un output che in aggiunta soddisfa un criterio di valutazione (ad esempio, etico), tale da poter applicare gli strumenti di verifica formale appropriati. In questo modo ci permettono:

- i. la simulazione del comportamento del programma sotto osservazione,
- ii. il ragionamento sulla sua affidabilità,
- iii. la descrizione delle possibilità di ottenere conoscenza dell'affidabilità del programma,
- iv. la formalizzazione del ragionamento in presenza di dati pregiudizievole.

Su questa base, il passo successivo consiste nello sviluppo di un applicativo che possa essere di supporto ad un utente finale che voglia effettivamente farsi delle domande sul comportamento del sistema sotto osservazione.

L'APPLICATIVO BRIO

L'esperienza di ricerca formale teorica di BRIO è stata convertita in un applicativo di supporto all'analisi dell'equità di sistemi IA opachi¹². L'analisi che è possibile eseguire con il tool BRIO si articola al momento in tre parti: (i) rilevazione del *bias*, (ii) misurazione del rischio, (iii) amplificazione del *bias*.

Il modulo di BRIO dedicato alla rilevazione delle violazioni dell'equità su categorie considerate protette prende in input: (i) le previsioni di un

¹¹ Francesco A. Genco e Giuseppe Primiero, "A Typed Lambda-Calculus for Establishing Trust in Probabilistic Programs". DOI: 10.48550/arXiv.2302.00958; Fabio Aurelio D'Asaro et al., "Checking Trustworthiness of Probabilistic Computations in a Typed Natural Deduction System", *Journal of Logic and Computation*, 2025; exaf003. DOI: 10.1093/logcom/exaf003.; Ekaterina Kubyshkina e Giuseppe Primiero, "A possible worlds semantics for trustworthy non-deterministic computations", *International Journal of Approximate Reasoning*, 172 (2024): 109212. DOI: 10.1016/j.ijar.2024.109212; Chiara Manganini e Giuseppe Primiero, "Reasoning with and about Bias", in *Perspectives on Logics for Data-driven Reasoning. Logic, Argumentation & Reasoning*, a cura di Juergen Landes e Hykel Hosnu, vol. 35, Springer, 2024. DOI: 10.1007/978-3-031-77892-6_7.

¹² Per i dettagli teorici sul funzionamento del tool: Greta Coraglia et al., "BRIOxAlkemy: a Bias Detecting Tool", *Proceedings of the 2nd Workshop on Bias, Ethical AI, Explainability and the role of Logic and Logic Programming*, co-located with the 22nd International Conference of the Italian Association for Artificial Intelligence (AI*IA 2023), 2024, pp. 44–60; Greta Coraglia et al., "Evaluating AI fairness in credit scoring with the BRIO tool". DOI: 10.48550/arXiv.2406.03292.

modello di IA codificate come un insieme di punti in una distribuzione con relative caratteristiche, (ii) un insieme di parametri, compresa la designazione di una o più caratteristiche sensibili, (iii) una distribuzione di riferimento, che può essere calcolata automaticamente sul dataset di input del modello, o fornita esternamente.

Il risultato dell'applicativo è una quantificazione della possibilità che il modello di IA in esame sia iniquo rispetto alle caratteristiche designate quando si confrontano le previsioni e la distribuzione di riferimento. Il sistema può condurre due tipi di analisi, consistenti nel confrontare il comportamento del sistema di IA rispetto a un comportamento desiderabile, e a classi sensibili correlate alla stessa caratteristica. Se la seconda analisi segnala un comportamento potenzialmente pregiudizievole, è possibile condurre un controllo successivo su alcune (o tutte) le sottoclassi delle classi sensibili considerate. Questo secondo controllo è volto a verificare se il *bias* riscontrato a livello delle classi possa essere spiegato da caratteristiche degli individui diverse dalla caratteristica sensibile in questione. L'output di questo modulo consiste in una lista di violazioni riscontrate e in una misurazione della percentuale di violazioni ottenute rispetto al numero massimo di violazioni che potrebbero riscontrarsi.

L'ambiente BRIO dispone poi di un modulo dedicato alla misurazione del rischio associato alle violazioni dell'equità da parte dei sistemi di IA. La misura del rischio viene ottenuta aggregando i risultati di tutti i test svolti dal modulo precedentemente descritto e che rilevano violazioni dell'equità. Il modulo prende in input una serie di risultati di test diversi, relativi a caratteristiche sensibili potenzialmente diverse, e restituisce un valore aggregato nell'intervallo reale unitario $[0,1]$, che rappresenta quanto sia elevato il rischio che il sistema di IA testato si comporti in modo ingiusto. Nel calcolo di questa misura, è anche possibile scegliere se concentrarsi sulla parità di gruppo o sulla parità individuale. Intuitivamente, concentrarsi sulla parità di gruppo significa considerare più grave una discriminazione basata su pochissime informazioni: per esempio, una scelta fatta solo sulla base del valore della caratteristica sensibile sarà una discriminazione di gruppo. Concentrarsi sulla parità individuale, d'altra parte, significa considerare più grave una discriminazione tra due individui che hanno molti valori in comune ma un valore diverso relativo alla caratteristica sensibile. BRIO offre l'opzione di scegliere tra le due.

Infine, con BRIO siamo in grado di esaminare ulteriormente il livello di *bias* mostrato da un modello in termini della sua propagazione dal set di test su cui viene valutato. Utilizzando BRIO per misurare il *bias* rispetto a una data caratteristica sensibile prima e dopo che venga elaborata dal modello, confrontiamo i due valori per capire quanto venga amplificato o diminuito. Tale procedura è particolarmente auspicabile in un contesto in cui sono disponibili e rilevanti serie temporali, e in cui i modelli vengono frequentemente aggiornati, per monitorare il comportamento a lungo termine dei processi utilizzati.

IMMAGINARE UN MODELLO EQUO

Un aspetto cruciale della metodologia BRIO riguarda l'immaginazione del modello *post-hoc* che è usato come criterio di valutazione. Immaginare cosa vogliamo che il modello ci dica, e come ci immaginiamo che il mondo debba essere, è un compito difficile. Ci è richiesto uno sforzo di immaginazione etica e politica, per decidere cosa sia una IA equa, condizione e non conseguenza dell'uso dell'IA. Questo richiama al ruolo dell'immaginazione nel ragionamento morale e a come questo è stato trattato nella letteratura.

Senza pretesa di completezza, un primo approccio di interesse per noi riguarda l'intendere in un certo modo una cosa distinta dal sé, e contrapporlo all'intendere il sé come altro¹³: in questo senso la metodologia BRIO pone prima un parametro di valutazione distinto dal sé, cioè distinto dalla realtà prodotta dal modello di IA, per poi interpretarlo come ciò che si vorrebbe il sé fosse, ovvero quello che vorremmo il modello di IA desse in output.

Un secondo approccio è più pragmatico: immaginare le conseguenze nell'uso degli strumenti e in particolare della tecnologia¹⁴. Da questo punto di vista, la metodologia BRIO è uno strumento per investigare le conseguenze della tecnologia, separando chiaramente i metodi che richiedono di essere investigati (quelli opachi dell'IA moderna), da quelli capaci di investigare (quelli trasparenti che combinano metodi statistici e simbolici).

¹³ Bernard Williams, "Moral Luck. Philosophical Papers 1973-1980", *Philosophical Quarterly*, 33, 132 (1983): 288-296.

¹⁴ Mark Coeckelbergh, *Imagination and principles: an essay on the role of imagination in moral reasoning*. New York: Palgrave-Macmillan, 2007.

Un terzo approccio che riteniamo rilevante è quello che vede l'IA stessa come atto immaginativo¹⁵: esplorare le condizioni di ammissibilità dei risultati di un modello di IA può diventare uno strumento per esplorare la mitigazione e la ricostruzione su basi eticamente affidabili di nuovi strumenti tecnologici.

In questo contesto, BRIO rappresenta un metodo di verifica formale delle conseguenze dell'IA, dopo l'immaginazione e possibilmente prima dell'uso.

ULTERIORI SVILUPPI

La metodologia BRIO e l'applicativo di verifica che lo implementa sono in fase di sviluppo da parte di MIRAI¹⁶, *spin-off* dell'Università degli Studi di Milano nato con l'obiettivo di offrire strumenti di supporto ad una IA responsabile. Molteplici sono le linee di potenziali ulteriori sviluppi. Qui desidero illustrarne solo due.

La prossima fase di evoluzione del tool di verifica BRIO riguarderà l'implementazione di un modulo per estrapolare dall'analisi delle violazioni di equità delle indicazioni utili per considerare la trasparenza del comportamento del modello sotto osservazione. Lo strumento di base è una metrica battezzata *Correction Distance*¹⁷. Essa esprime l'inverso della quantità di informazione aggiuntiva sufficiente perché un modello sotto osservazione produca una classificazione contraria a quella offerta su un certo individuo, pesata sull'accuratezza del modello. In altre parole, questa misura ci dice se (e in quel caso con quanta informazione su caratteristiche non ancora valutate per l'individuo in questione) il modello potrebbe cambiare la predizione attualmente fornita, passando dunque da una predizione scorretta ad una corretta o viceversa (nel caso di classificazione binaria), o semplicemente ad una delle altre possibili classificazioni (nel caso di modelli *multi-target*).

¹⁵ Jim Davies, "Artificial Intelligence and Imagination", in *The Cambridge Handbook of the Imagination*, a cura di Anna Abraham, Cambridge Handbooks in Psychology, Cambridge University Press; 2020, pp. 162-172.

¹⁶ MIRAI spin-off. <https://mirai.systems/>.

¹⁷ Chiara Manganini e Giuseppe Primiero, "Reasoning with and about Bias", *op. cit.*

Con l'implementazione di questo nuovo modulo sarà possibile un'analisi puntuale del valore che ciascuna proprietà contribuisce nel determinare un certo valore della predizione, fungendo a tutti gli effetti come una analisi della rilevanza delle *features* del modello. Su questa base sarà possibile sviluppare ulteriori metriche per fare inferenze controfattuali.

La successiva fase di sviluppo è quella su cui si concentra l'analisi teorica del progetto SMARTEST (*Simulation of Probabilistic Systems for the Age of the Digital Twin*).

Questa nuova fase progettuale mira ad offrire risultati più generali per la definizione di 'Affidabilità-con-BRIO'. Un potenziale limite della strategia di verifica sviluppata finora è, infatti, che sistemi non-deterministici possono presentare diversi livelli di affidabilità rispetto a diversi output.

Per analogia, è come pensare che dato un dado, si possa osservare un comportamento affidabile rispetto al numero 5, e risulti inaffidabile nel produrre il numero 6 quando computato confrontandolo con un dado non truccato. Sebbene questa sia una conclusione perfettamente sensata nel contesto in cui ragioniamo, è plausibile volere costruire delle classi di equivalenza sui risultati di affidabilità ottenuti, per potere inferire in quali condizioni il sistema nel suo complesso possa dirsi affidabile, nel senso di soddisfare a criteri di validità considerati sufficientemente forti.

Inoltre, le varie classi di affidabilità dovrebbero esprimere una gradazione oltre un valore binario: un sistema che è 'perlopiù' affidabile è preferibile ad uno 'quasi mai' affidabile, almeno in contesti non critici.

Per fare questo intendiamo sfruttare i risultati tecnici offerti dalle bi-simulazioni probabilistiche. Con questo strumento formale possiamo caratterizzare risultati la cui incertezza rientra in un range ammissibile per il modello considerato equo. Oppure potrebbero essere risultati la cui incertezza corrisponde a quella che attribuiamo al nostro modello equo, senza tuttavia essere esaustivi, cioè un'incertezza solo parzialmente ammissibile.

Infine, potremmo trovarci di fronte a risultati la cui incertezza potrebbe essere in alcuni casi parziale, in altri eccedente rispetto a quella considerata ammissibile dal modello equo.

Su questa base si distinguono diverse nozioni di validità e quindi anche diverse classi di comportamenti rispetto all’Affidabilità-con-BRIO¹⁸.

CONCLUSIONI

L’obiettivo concettuale primario di BRIO (e poi di SMARTTEST) può essere formulato come segue: creare strumenti teorici e tool pratici per ripensare l’approccio, intrinsecamente di larga scala, dei nostri sistemi computazionali (e di IA in particolare), ed offrire mezzi per la valutazione critica alle comunità che di tali sistemi fanno uso.

In particolare, il metodo BRIO richiede di immaginare esplicitamente quei criteri e quei valori che una comunità ritiene inalienabili nell’uso della tecnologia al di là dell’ottimizzazione, dell’intervento predittivo, e della pratica puramente soluzionista delle nostre infrastrutture tecno-sociali attuali. Questi criteri e valori, di natura etica, epistemologica ed estetica devono essere esplicitamente formulati, messi a verifica e sfruttati anche legalmente come strumenti di valutazione della tecnologia in un contesto di uguaglianza e giustizia sociale, al di là della massimizzazione dell’efficienza.

Nel formulare una strategia di contrasto all’egemonia della tecnologia del controllo realizzata tramite sistemi di IA profondamente opachi, statisticamente ingiusti e discriminatori, i metodi formali sono uno strumento attivo di resistenza che aumenta le nostre capacità di riflessione, supporta la necessaria opposizione al controllo algoritmico, e ci aiuta nel processo di cura collettiva e di raggiungimento dell’eguaglianza sociale.

¹⁸ Chiara Manganini e Giuseppe Primiero, “Defining Formal Validity Criteria for Machine Learning Models”, in *Philosophy of Science for Machine Learning*, a cura di Giorgia Pozzi e Juan Manuel Duran, Springer, forthcoming. L’idea di definire classi di uguaglianza per la verifica di sistemi non-deterministici deriva dall’applicazione della nozione di copia approssimata introdotta in: Nicola Angius e Giuseppe Primiero, “The logic of identity and copy for computational artefacts”, *Journal of Logic and Computation*, 28, 6 (2018): 1293-1322, e utilizza lo strumento della bsimulazione su strutture probabilistiche; cfr. ad esempio: Christel Baier et al., “Bisimulation and Simulation Relations for Markov Chains”, *Electronic Notes in Theoretical Computer Science*, 162 (2006): 73-78. Nel caso di BRIO, queste nozioni sono sviluppate in un contesto formale teoretico-dimostrativo.

RINGRAZIAMENTI

Questa ricerca è stata finanziata dai Progetti: PRIN2020 BRIO (*Bias, Risk and Opacity in AI*, 2020SSKZ7R), PRIN2020 SMARTTEST (*Simulation of Probabilistic System in the Age of the Digital Twin*, 20223E8Y4X) del MUR e dal Progetto “Dipartimenti di Eccellenza 2023-2027” attribuito dal MUR al Dipartimento di Filosofia “Piero Martinetti” dell’Università degli Studi di Milano.

Tra etica ed epistemologia: alcune note critiche sul principio della spiegabilità in Intelligenza Artificiale

Fabio Fossa, Giacomo Zanotti, Stefano Canali

Politecnico di Milano

INTRODUZIONE

Il principio della spiegabilità figura in numerosi documenti di indirizzo volti a chiarire quali siano i valori da rispettare in vista di uno sviluppo responsabile delle tecnologie di Intelligenza Artificiale (IA) in ambito sia internazionale che europeo – contesto, quest’ultimo, su cui in particolare ci concentreremo. Tuttavia, come vedremo, progettare sistemi spiegabili pone una serie di difficoltà che esulano dalla dimensione propriamente tecnica della questione, su cui dunque bisogna riflettere facendo ricorso ad altri strumenti concettuali.

Scopo di questo contributo è indagare da una prospettiva filosofica alcuni problemi sollevati dall’adozione del principio della spiegabilità nel campo dell’IA. La seconda sezione esplora da una prospettiva teorica lo statuto normativo del valore della spiegabilità, mettendone in evidenza la natura strumentale e le relative implicazioni pratiche. La terza sezione, adottando un punto di vista più marcatamente epistemologico, mette in luce come un sistema spiegabile non sia necessariamente comprensibile per le persone coinvolte, sottolineando l’urgenza di un pluralismo di strategie esplicative. Nella quarta e ultima sezione le considerazioni svolte vengono concretizzate con un’analisi dell’utilizzo di sistemi IA nell’ambito medico e delle complesse questioni etiche ed epistemologiche che ne conseguono.

La discussione intende rimarcare come l’adozione del principio di spiegabilità non ponga solo problemi tecnici rilevanti, ma richieda una trattazione puntuale dei relativi aspetti etici ed epistemologici. La realizzazione di sistemi di IA affidabili dipende in buona misura anche da questo genere di indagini.

LA SPIEGABILITÀ E IL SUO CONTESTO: QUESTIONI ETICHE

Dal punto di vista della filosofia morale, il valore della spiegabilità presenta una peculiarità degna di nota se confrontato con altri principi a cui spesso ci si appella in vista della definizione di linee guida etiche per la progettazione e l'utilizzo dell'IA. Si consideri il caso europeo. Nel contesto del noto documento *Orientamenti etici per un'IA affidabile*¹, il valore della spiegabilità confluisce nel più ampio principio dell'esplicabilità. Insieme al rispetto dell'autonomia umana, alla prevenzione dei danni e all'equità, il principio dell'esplicabilità completa il quadro valoriale che informa l'approccio europeo alla c.d. *Trustworthy AI*, cioè all'IA degna di fiducia. Ne riportiamo qui la definizione che alle istanze classiche di apertura della scatola nera algoritmica unisce questioni di trasparenza aziendale e istituzionale:

L'esplicabilità è fondamentale per creare e mantenere la fiducia degli utenti nei sistemi di IA. Tale principio indica che i processi devono essere trasparenti, le capacità e lo scopo dei sistemi di IA devono essere comunicati apertamente e le decisioni, per quanto possibile, devono poter essere spiegate a coloro che ne sono direttamente o indirettamente interessati. Senza tali informazioni, una decisione non può essere debitamente impugnata. Non sempre è possibile spiegare, tuttavia, perché un modello ha generato un particolare risultato o decisione (e quale combinazione di fattori di input vi ha contribuito). È il cosiddetto caso della “scatola nera”, i cui algoritmi richiedono un'attenzione particolare. In tali circostanze, possono essere necessarie altre misure per garantire l'esplicabilità (ad esempio, la tracciabilità, la verificabilità e la comunicazione trasparente sulle capacità del sistema), posto che il sistema nel suo complesso rispetti i diritti fondamentali. Il grado di esplicabilità necessario dipende in larga misura dal contesto e dalla gravità delle conseguenze nel caso in cui il risultato sia errato o comunque impreciso².

Le considerazioni appena riportate sulla rilevanza della spiegabilità nel caso di sistemi di IA che svolgono compiti significativi dal punto di vista morale mettono in evidenza le ragioni che portano ad assegnarle lo statuto di principio fondamentale. Tuttavia, il suo rango di principio non corrisponde appieno allo statuto normativo di cui godono i principi del rispetto

¹ Commissione europea, Direzione generale delle Reti di comunicazione, dei contenuti e delle tecnologie, *Orientamenti etici per un'IA affidabile*, Ufficio delle pubblicazioni, 2019. DOI: 10.2759/640340.

² *Ivi*, pp. 14-15.

dell'autonomia, della prevenzione dei danni e dell'equità. L'importanza di questi tre principi, infatti, non è data da motivazioni ulteriori. Al contrario, essi rappresentano diritti fondamentali e richiedono assoluto rispetto in quanto tali. Affinché un sistema di IA che limiti uno di questi principi possa essere ritenuto degno di fiducia, dunque, è necessaria una giustificazione che mostri come la limitazione sia motivata dal rispetto di un altro principio, ritenuto nella fattispecie più rilevante dal punto di vista morale.

Non è questo il caso della spiegabilità. Non si tratta di un principio che, in alcuni casi, può giustificare una limitazione concomitante dell'autonomia, della prevenzione dei danni o dell'equità. Al contrario, la definizione del principio chiarisce che la sua rilevanza etica è direttamente proporzionale alla misura in cui esso promuove l'autonomia individuale e quindi tutela responsabilità e giustizia. In altre parole, la spiegabilità non è un valore che ha in sé la propria ragion d'essere e dunque è da perseguire in senso assoluto. Esso vale, piuttosto, solo nella misura in cui permette agli esseri umani coinvolti di esercitare la propria autonomia, soddisfare le responsabilità che ne derivano, e raddrizzare eventuali torti. E solo in questa misura è da perseguire.

Per casi simili, in etica applicata si usa distinguere tra valori intrinseci e valori strumentali³. Un valore è intrinseco quando ha in sé il motivo della sua rilevanza morale. Ogni sua limitazione richiede di essere giustificata dall'affermazione contestuale di un altro valore intrinseco, che nella situazione di cui ci si occupa è ritenuto degno di precedenza. I principi del rispetto dell'autonomia, della prevenzione dei danni e dell'equità sono di questo tipo. Essi esplicitano diversi aspetti della dignità umana, la quale nel contesto sia della cultura morale che della normativa europea esige rispetto incondizionato. Un valore è, invece, strumentale quando la sua rilevanza morale non risiede in sé stesso, ma in ciò che il suo rispetto produce o comporta in una data situazione. È questo il caso della spiegabilità: il suo valore non è assoluto, ma dipende dal contesto – dalla misura in cui, in una data situazione dove umani e sistemi di IA interagiscono, si rende necessaria la tutela dell'autonomia, della responsabilità e della giustizia.

³ Ibo Van de Poel, "Values in engineering design", in *Philosophy of Technology and Engineering Sciences*, a cura di A. Meijers, Burlington-Oxford-Amsterdam: North Holland, 2009, pp. 973–1006. DOI: 10.1016/B978-0-444-51667-1.50040-9.

La differenza tra la spiegabilità e gli altri principi dell'IA degna di fiducia è meritevole di nota per quanto ne consegue sul piano pratico. Il peculiare statuto normativo dei valori strumentali motiva considerazioni diverse rispetto al caso dei valori intrinseci. Mettere in pratica il valore della spiegabilità, infatti, non significa perseguirla a meno che altri principi più rilevanti richiedano la precedenza. Comporta invece far sì che i valori che la giustificano – autonomia, responsabilità, giustizia – siano adeguatamente perseguibili nelle situazioni di riferimento. Di conseguenza, l'applicazione della spiegabilità è inseparabile da considerazioni contestuali – considerazioni, cioè, relative alle condizioni pratiche in cui gli individui si trovano ad interagire con sistemi di IA e ai modi in cui il comportamento autonomo, responsabile e giusto ne viene condizionato e mediato.

Mettere in pratica la spiegabilità richiede, quindi, riflessioni e misure che rispondano a molteplici e mutevoli condizioni effettive. Perciò, il problema principale da affrontare in questo senso consiste nella tutela simultanea del comportamento autonomo, responsabile e giusto in situazioni in cui l'interazione con sistemi di IA ne media l'esercizio. Le difficoltà sorgono quando le tre dimensioni suddette non possono essere soddisfatte allo stesso modo o grado. Il problema, più che tecnico, è di carattere etico. Le condizioni di esercizio di uno dei valori che la spiegabilità dovrebbe tutelare possono infatti rappresentare ostacoli all'esercizio degli altri. Mettere in pratica la spiegabilità significa, quindi, imparare a gestire in modo soddisfacente situazioni in cui autonomia, responsabilità e giustizia possono essere servite sincreticamente soltanto in modo imperfetto.

La sfida è complessa, tanto più quando si muove dal piano astratto della riflessione sui valori al piano concreto della loro implementazione. Si consideri, ad esempio, un possibile contrasto tra supporto dell'autonomia e promozione della responsabilità. Per comportarsi in modo autonomo è necessario avere accesso alle informazioni relative alla decisione che si vuole prendere o all'azione che si vuole svolgere. Tuttavia, quantità eccessive di informazioni o modalità di comunicazione confuse possono contribuire all'adozione di comportamenti irresponsabili.

Sul piano dell'implementazione, questo problema di spiegabilità si manifesta, ad esempio, nel caso dei veicoli autonomi in cui l'utente è responsabile

di supervisionare il funzionamento del sistema⁴. *Quali informazioni devono essere condivise con l'utente per supportare il suo esercizio di autonomia e responsabilità?* Il rispetto dell'autonomia sembrerebbe richiedere che il sistema metta l'utente al corrente di tutte le informazioni che esso elabora circa l'ambiente circostante, cosicché l'utente possa prendere il controllo del veicolo qualora un ostacolo non sia rilevato o sia categorizzato erroneamente (ad esempio, una bicicletta sia equivocata con un'automobile). Una simile configurazione del sistema, tuttavia, potrebbe portare l'utente a prendere il controllo del veicolo anche in situazioni in cui errori di percezione artificiale come quelli appena accennati non pongano effettivi rischi in termini di sicurezza. Ciò è problematico, in quanto la manovra in cui l'utente riprende il controllo del veicolo è nota per comportare rischi non indifferenti. Il suo svolgimento immotivato, quindi, può essere valutato come irresponsabile.

In questo senso, si potrebbe argomentare che un certo livello di opacità⁵ del sistema potrebbe aiutare a mettere l'utente nelle condizioni di esercitare meglio la propria responsabilità. E tuttavia, la limitazione dell'autonomia dell'utente che ne conseguirebbe è evidente. Mettere in pratica la spiegabilità in un simile contesto comporta decisioni complesse su come veicolare contemporaneamente i diversi valori intrinseci che questa, data la sua natura strumentale, è deputata a mediare – decisioni che necessariamente intersecano la dimensione tecnica dei sistemi coinvolti, ma che presentano un profilo prima di tutto morale.

La peculiarità dello statuto normativo della spiegabilità ne rende, quindi, particolarmente difficile e controversa la gestione. Motivo per cui è importante metterla a fuoco e lavorare a una presa di consapevolezza di tale difficoltà, la quale trascende questioni tecniche e intercetta invece incertezze morali che dobbiamo imparare a riconoscere e di cui dobbiamo comprendere come farcene carico.

⁴ Cfr. Matteo Matteucci et al., "Contextual Challenges to Explainable Driving Automation: The Case of Machine Perception", in *Connected and Automated Vehicles: Integrating Engineering and Ethics*, a cura di F. Fossa e F. Cheli, Cham: Springer, pp. 37-61. DOI: 10.1007/978-3-031-39991-6_3.

⁵ Ovvero, nel nostro caso, non fornire all'utente informazioni circa la categorizzazione degli ostacoli percepiti dal sistema. Per opacità si intende il livello di insondabilità del processo inferenziale che il sistema segue per produrre l'output. Si veda, ad es., Kaffin A. Creel, "Transparency in Complex Computational Systems", *Philosophy of Science* 87, 4 (2020): 568-589. DOI: 10.1086/709729.

TRA SPIEGAZIONE E COMPrensIONE: ALCUNI PROBLEMI EPISTEMOLOGICI

Avendo analizzato lo statuto della spiegabilità come principio etico, possiamo concentrarci su alcuni aspetti epistemologici. Al netto di alcune obiezioni sulla coerenza concettuale di questa nozione⁶, è interessante notare come la spiegabilità sia sistematicamente annoverata tra i fattori che rendono un sistema di IA *trustworthy*. Ad esempio, come abbiamo visto, nelle linee guida europee del 2019 il principio della esplicabilità risulta essere un imperativo etico.

Passando dai principi ai requisiti più concreti elencati nel documento, viene specificato che «affinché un sistema di IA possa essere tecnicamente spiegabile gli esseri umani devono poter capire e tenere traccia delle decisioni prese dal sistema stesso». Viene inoltre chiarito che «se un sistema di IA influisce considerevolmente sulla vita delle persone, dovrebbe sempre essere possibile richiedere una spiegazione adeguata del processo decisionale del sistema»⁷. In alcuni casi, tuttavia, questo è più facile a dirsi che a farsi.

Senza voler in alcun modo screditare il programma XAI o minimizzare i suoi risultati, rimane il fatto che, ad oggi, diversi sistemi di IA rimangono difficilmente spiegabili. Basti pensare ai *Large Language Models*, divenuti celebri anche tra gli utenti non esperti in seguito al rilascio di *ChatGPT* da parte di *OpenAI* e che, pur stupendo per le loro prestazioni, risultano essere estremamente opachi⁸. Questa opacità dipende da più fattori, tra cui la complessità e la dimensione di questi modelli – tanto per fare un esempio, il modello alla base della prima versione di *ChatGPT*, *GPT3.5*, ha circa 175 miliardi di parametri. Anche al di fuori dell'ambito *Large Language Models* e senza arrivare alle dimensioni di questi ultimi, la spiegabilità rimane spesso una sfida significativa al crescere della complessità dei modelli utilizzati.

La situazione si complica ulteriormente se consideriamo che le spiegazioni in questione dovrebbero essere allo stesso tempo fedeli al funzionamento del modello – ovvero riuscire a riflettere con sufficiente precisione il processo inferenziale seguito – e comprensibili per gli esseri

⁶ Giacomo Zanotti, Daniele Chiffi, Viola Schiaffonati, “AI-Related Risk: An Epistemological Approach”, *Philosophy of Technology* 37, 66 (2024): 1-18. DOI: 10.1007/s13347-024-00755-7.

⁷ Commissione europea, *op. cit.*, pp. 20-21.

⁸ Haiyan Zhao et al., “Explainability for large language models: A survey”, *ACM Transactions on Intelligent Systems and Technology*, 15, 2 (2024): 1-38. DOI: 10.1145/3639372.

umani. È però facile immaginare come questi due aspetti possano entrare in contrasto. La complessità dei modelli può infatti tradursi in una complessità della spiegazione dei processi inferenziali, pregiudicando la possibilità di comprensione da parte degli utenti umani. Spiegazioni più compatte, che vanno ad approssimare il funzionamento del modello senza tuttavia restituirne tutta la complessità, pongono sicuramente meno problemi sul fronte della comprensione. A seconda dei casi, tuttavia, il prezzo da pagare per questa relativa semplicità potrebbe essere una poca fedeltà rispetto al funzionamento effettivo del modello. Si pone dunque il problema di trovare un compromesso per cui la spiegazione, pur essendo sufficientemente fedele, sia anche ragionevolmente comprensibile.

Il tema della comprensione apre un'altra importante questione, spesso implicita nel dibattito: in molti casi, gli utenti dei sistemi di IA – o comunque le persone affette dalle previsioni e decisioni di questi ultimi – non sono utenti esperti. Da coloro che utilizzano regolarmente assistenti vocali e social network agli utenti finali di sistemi di IA impiegati in ambito medico e finanziario, sono moltissime le persone che, pur non avendo particolare competenza nell'ambito, interagiscono più o meno direttamente con sistemi di IA. Si presenta così un'ulteriore difficoltà: non tutti gli utenti sarebbero in grado di comprendere le spiegazioni dei sistemi che utilizzano. Parte della ricerca in ambito XAI è dunque dedicata a individuare e sviluppare tecniche che permettano di fornire spiegazioni che siano accessibili anche a persone non esperte.

Emerge comunque un'esigenza più generale. L'enfasi sulla spiegabilità nel contesto del dibattito sulla *Trustworthy AI* è motivata dall'assunto per cui la comprensione da parte degli utenti è cruciale per la fiducia (giustificata) nei sistemi di IA. Prima di poter spiegare i processi inferenziali che portano ad output specifici, sembra dunque necessario costruire una base di alfabetizzazione algoritmica che preveda una consapevolezza generalizzata di quelle che sono le capacità e i principi di funzionamento dei sistemi di IA⁹. Tra le altre cose, costruire questa base di conoscenza contribuirebbe ad arginare

⁹ La necessità di fornire strumenti per la comprensione delle principali tecniche e applicazioni di IA è ribadita anche nel recente AI Act, articolo 4 (Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024. <http://data.europa.eu/eli/reg/2024/1689/oj>) e dalle raccomandazioni dell'UNESCO sull'etica dell'IA, p. 23 (*Recommendation on the Ethics of Artificial Intelligence*, 2022. <https://unesdoc.unesco.org/ark:/48223/pf0000381137>).

i timori derivanti da concezioni sensazionalistiche dell'IA attuale come anticamera della creazione di macchine super-intelligenti (e potenzialmente maldisposte nei nostri confronti), spostando la discussione pubblica sui rischi concreti derivanti dall'uso di sistemi di IA¹⁰.

A questo proposito, è interessante notare come anche le linee guida europee prevedano accanto al requisito della spiegabilità quello della comunicazione. Tra le altre cose, si esplicita che «dovrebbero essere comunicate agli operatori del settore dell'IA o agli utenti finali le capacità e le limitazioni del sistema in maniera consona al caso d'uso in questione»¹¹.

Ciò detto, fornire spiegazioni di alto livello che rendano ampiamente comprensibile il funzionamento dei sistemi di IA non è facile. Da una parte, queste spiegazioni dovrebbero essere sufficientemente informative, in modo da permettere una comprensione sostanziale dei principi di funzionamento, dei limiti e delle capacità dei sistemi in questione. Dall'altra, l'informatività non può andare a discapito dell'accessibilità di queste spiegazioni per le persone non specialiste. Si rende dunque necessario pensare a delle strategie esplicative *ad hoc* che permettano agli utenti non esperti di raggiungere comunque un buon grado di comprensione delle tecnologie che utilizzano.

Una strategia spesso sfruttata è quella delle spiegazioni che potremmo definire metaforiche. Nello specifico, si ricorre a termini solitamente utilizzati per descrivere processi cognitivi e affettivi tipicamente umani. Non è raro imbattersi in narrative riguardanti sistemi di IA che 'capiscono' e 'conoscono', specie nel caso dei sopracitati *Large Language Models*, o in *chatbot* pensati per essere usati come «amici virtuali» che «hanno a cuore» l'utente, che «sono sempre pronti per ascoltare e parlare»¹². Al di là di queste descrizioni, che hanno spesso fini marcatamente pubblicitari, è indubbio che il discorso pubblico e accademico sull'IA faccia ampio uso di metafore – a partire dalla stessa espressione 'intelligenza artificiale', qualcuno potrebbe dire.

Posto che la metafora è uno strumento estremamente prezioso per veicolare concetti complessi in maniera accessibile – e non abbiamo ragione

¹⁰ Editorial, "Stop talking about tomorrow's AI doomsday when AI poses risks today", *Nature*, 618 (2023): 885-886. DOI: 10.1038/d41586-023-02094-7.

¹¹ Commissione europea, *op. cit.*, p. 21.

¹² Chatbot Replika. <https://replika.com/>.

di credere che l'IA faccia eccezione – il suo uso non è privo di rischi. All'utente non esperto può infatti non risultare chiara la natura non letterale di alcune espressioni¹³. Questo può dare origine a fraintendimenti circa quelle che sono le effettive capacità e più in generale la natura dei sistemi di IA, andando ad alimentare processi di antropomorfizzazione ed errate attribuzioni di capacità simil umane a sistemi artificiali, e generando in ultima istanza sfiducia o addirittura timore nei confronti delle innovazioni in IA.

Riassumendo, è cruciale portare avanti la ricerca sul fronte XAI per far sì che i processi inferenziali dietro gli output dei sistemi di IA siano sempre meno opachi. Allo stesso tempo, però, è prioritario riflettere su quali potrebbero essere delle strategie esplicative per fornire a tutti gli utenti una comprensione di base dei principi di funzionamento, dei limiti e delle capacità dei sistemi di IA che utilizzano.

IL CASO DELLA MEDICINA TRA EPISTEMOLOGIA ED ETICA

Come abbiamo visto nelle sezioni precedenti, la spiegabilità solleva delle questioni etiche ed epistemologiche di difficile risoluzione. Ne mostreremo ora la rilevanza in un ambito specifico e particolarmente significativo: l'utilizzo di sistemi di IA in medicina.

Se pensiamo al contesto italiano, nell'ottobre del 2023 è apparso un documento a firma del *Garante per la Protezione dei Dati Personali*, che presenta i punti fondamentali da rispettare in vista dell'utilizzo dell'IA nell'ambito medico, in particolare al fine di realizzare servizi sanitari nazionali basati sull'IA. In linea con la strategia europea da cui abbiamo preso le mosse, secondo il Garante esistono tre principi in particolare che dovrebbero essere al centro della governance dell'IA nell'esecuzione di compiti di rilevante interesse pubblico, come quello medico e sanitario. Si tratta dei principi di non esclusività, non discriminazione e del «principio di conoscibilità, in base al quale l'interessato ha il diritto di conoscere l'esistenza di processi decisionali basati su trattamenti automatizzati e, in tal caso, di ricevere informazioni significative sulla logica utilizzata, sì da

¹³ Dave Murray-Rust, Johanna Nicenboim, Dan Lockton, "Metaphors for designers working with AI", in DRS2022 Conference Proceedings, Bilbao, 2022, pp. 1-19. DOI:10.21606/drs.2022.667.

poterla comprendere»¹⁴. Al di là delle possibili differenze tra i concetti di conoscibilità e spiegabilità (su cui qui non ci soffermeremo)¹⁵, alla luce di quanto visto finora nel presente capitolo non risulta sorprendente che la spiegabilità sia al centro delle considerazioni etico-legali espresse dal Garante e da enti e istituzioni simili. Tuttavia, la questione della spiegabilità, abbiamo anche visto, ha un carattere controverso nel dibattito contemporaneo sull'IA. Ci si può quindi chiedere quali declinazione, applicabilità e limiti possa avere un principio così centrale e controverso in un ambito altrettanto centrale e potenzialmente controverso come quello medico.

L'utilizzo dell'IA in medicina è un contesto specifico in cui emergono chiaramente diverse delle questioni che abbiamo discusso nel capitolo, mostrando in questo modo le ricadute pratiche e significative delle questioni teoriche finora affrontate. Partiamo anzitutto con il notare che il principio di spiegabilità nell'ambito medico ha carattere strumentale. Difatti, se ci interroghiamo sui motivi per i quali dovremmo poter conoscere il funzionamento di sistemi IA nell'ambito medico, diverse risposte possibili vanno nella direzione del rispetto di altri principi e valori. Pensiamo ad esempio alle tante applicazioni di salute digitale che possono essere usate per tracciare aspetti della salute quali attività fisica, dieta, sonno, periodo mnestruale, ecc. L'IA può essere utilizzata come parte di questi sistemi per analizzare i dati raccolti ed elaborare consigli e raccomandazioni all'utente, ad esempio rispetto a quali cibi preferire per il pasto successivo. In questo caso si potrebbe sostenere che sia necessario per l'utente conoscere il modo in cui le raccomandazioni elaborate dall'IA sono state ottenute, che quindi dovrà essere spiegabile, perché ciò aiuta l'elaborazione di una scelta autonoma e quindi promuove l'autonomia nelle decisioni mediche.

Ancora, il rispetto e l'implementazione della spiegabilità in medicina possono essere collegati alla necessità di difendersi da possibili errori e dalle loro ricadute. Senza spiegabilità, si potrebbe dire in questa direzione, può

¹⁴ Garante per la protezione dei dati personali, “Decalogo per la realizzazione di servizi sanitari nazionali attraverso sistemi di Intelligenza Artificiale”, 2023, p. 6. <https://www.garanteprivacy.it/documents/10160/0/Decalogo+per+la+realizzazione+di+servizi+sanitari+nazionali+attraverso+sistemi+di+Intelligenza+Artificiale.pdf/a5c4a24d-4823-e014-93bf-1543f1331670?version=2.0>.

¹⁵ Su questo tema di veda ad es. Monica Palmirani, “Interpretabilità, conoscibilità, spiegabilità dei processi decisionali automatizzati”, in *XXVI lezioni di Diritto dell'Intelligenza Artificiale*, a cura di U. Ruffolo, Giappichelli, Torino pp. 66-76.

risultare molto difficile individuare *bias* e discriminazione – problemi che sono particolarmente significativi e pervasivi in un ambito come quello medico, storicamente diseguale e discriminatorio, più attento agli interessi di salute e le patologie di alcuni gruppi sociali rispetto ad altri. L'IA ha spesso l'effetto di esacerbare diseguaglianze e discriminazioni e la mancanza di spiegabilità può complicare ulteriormente le cose. Ad esempio, alcuni algoritmi di IA vengono attualmente utilizzati per identificare individui potenzialmente a rischio di sviluppare patologie complesse e prevenirne lo sviluppo. Le potenzialità sono diverse e significative, viste le capacità di analisi dell'IA, ma molte sono anche le preoccupazioni rispetto al fatto che questi sistemi siano sviluppati sulla base di dati storici e rischino di riprodurre diseguaglianze e discriminazioni, ad esempio a livello di categorie etniche e di genere. La mancanza di spiegabilità dei sistemi di IA può rendere lo scenario ancora più preoccupante, dato che la scarsa trasparenza non facilita la verifica di *bias*¹⁶. Di nuovo, quindi, il valore della spiegabilità sta nella possibilità di promuovere e rispettare altri valori e principi, ad esempio in questo caso equità e giustizia.

Avere sistemi di IA in medicina che siano spiegabili può anche essere strumentale a un processo di attribuzione di responsabilità. Anche in questo caso si potrebbe dire che senza spiegabilità risulta difficile tutelare il rispetto della responsabilità come principio centrale della pratica medica. Nel momento in cui si utilizza l'IA per supportare decisioni diagnostiche e si verificano degli errori, ad esempio una diagnosi sbagliata, il fatto che questi strumenti non siano spiegabili può rendere molto complicato capire chi sia 'colpevole' e tra chi si debba distribuire la responsabilità delle decisioni prese. Sebbene l'attribuzione di responsabilità sia un problema classico delle riflessioni su tecnologie e sistemi autonomi, l'assenza di spiegabilità non fa che rendere il tutto più complesso – soprattutto in un ambito, come quello medico, in cui l'attribuzione di responsabilità è oggetto di procedure particolarmente regolamentate e codificate.

Esistono, perciò, delle motivazioni prettamente strumentali sulla base delle quali formulare l'appello alla necessità di spiegabilità in ambito medico. In questo senso, l'utilizzo dell'IA in medicina si inserisce all'interno delle considerazioni generali che abbiamo sviluppato nel capitolo. Allo stesso

¹⁶ Ziad Obermeyer et al., "Dissecting racial bias in an algorithm used to manage the health of populations", *Science*, 366, 6464 (2019): 447-453. DOI: 10.1126/science.aax2342.

tempo, però, ci sono delle specificità che possiamo sottolineare. Anzitutto, le motivazioni sono sì strumentali, ma fanno riferimento a valori che in realtà nell'ambito medico vanno ritenuti diritti fondamentali e sono considerati tra i principi cardine della bioetica. Un principio quale l'autonomia ha una connotazione specifica e particolarmente significativa in bioetica e invocare la spiegabilità per tutelare da possibili errori e processi decisionali manchevoli può essere collegato a principi bioetici quali la beneficenza e la non maleficenza.

Inoltre, specificando ulteriormente, possiamo sottolineare come il carattere strumentale della spiegabilità in medicina sia connesso a valori non immediatamente etici. In filosofia della scienza si individua una categoria di valori che ha a che fare con la produzione e lo sviluppo di conoscenza, i c.d. valori epistemici, quali ad esempio semplicità, completezza, intellegibilità, ecc. In questo senso, la spiegabilità si presenta come un valore strumentale anche alla tutela di valori, caratteristiche e fini dello sviluppo di conoscenza scientifica affidabile e di qualità sui fenomeni di salute e malattia¹⁷.

Il quadro, quindi, si complica ulteriormente: la spiegabilità emerge come valore strumentale alla tutela di altri valori, che sono sia di carattere etico che epistemico. A complicarsi sono soprattutto i contrasti tra spiegabilità e altri valori, tema sul quale si è concentrato gran parte del dibattito filosofico contemporaneo sulla spiegabilità in medicina.

Pensiamo, ad esempio, a un ambito di applicazione dell'IA potenzialmente molto significativo: il c.d. *triage*, ossia l'insieme di decisioni che vengono prese durante l'accoglienza dei pazienti in ambito clinico per distribuirli tra categorie di urgenza diverse a seconda della gravità del loro quadro clinico. Vista la necessità di analizzare diverse informazioni e prendere decisioni in tempi molto brevi, il *triage* è forse un candidato ideale per ricevere assistenza dall'IA. Tuttavia, l'utilizzo dell'IA e problemi di spiegabilità ne mostrano anche il carattere controverso. Ad esempio, in un caso recente e ampiamente discusso anche nella letteratura filosofica, l'assenza di spiegabilità di reti neurali per assistere il *triage* è stato considerato un motivo sufficiente per scoraggiarne l'utilizzo in ambito clinico, in quanto le reti avevano individuato associazioni spurie e ne avrebbero

¹⁷ Juan M. Durán, "Dissecting scientific explanation in AI (sXAI): A case for medicine and healthcare", *Artificial Intelligence*, 297, 103498 (2021): 1-14. DOI: 10.1016/j.artint.2021.103498.

potute individuare altre, con effetti potenzialmente pericolosi¹⁸. Possiamo quindi interpretare questo come un caso in cui la spiegabilità e i suoi effetti strumentali rispetto a valori etici o epistemici, quali la non maleficenza e l'intelligibilità, prevalgono su altre considerazioni.

Tuttavia, emergono diversi contrasti possibili. In primo luogo, non è immediatamente chiaro cosa significhi spiegabilità in questo caso e le questioni di comunicazione che abbiamo discusso sono particolarmente evidenti – un conto è richiedere che delle reti neurali siano spiegabili per chi ha sviluppato questi strumenti, ben altra questione è la spiegabilità per medici o pazienti. Inoltre, pensiamo al caso in cui l'utilizzo dell'IA porta a risultati non spiegabili ma molto accurati, rispetto al *triage* di pazienti nelle corrette categorie di urgenza. Un contrasto complesso emerge tra la spiegabilità e altri valori rispetto a cui dovrebbe svolgere un effetto strumentale, tra cui la beneficenza.

CONCLUSIONI

In questo capitolo sono state esplorate diverse questioni etiche ed epistemologiche relative alla spiegabilità nel contesto dei sistemi di IA. È emerso come in molti casi i problemi siano di carattere normativo più che tecnico, il che rende evidente la necessità di perseguire percorsi di ricerca interdisciplinari. Decisioni in una direzione o un'altra pongono già *trade-off* significativi e occorrerà sempre più gestire situazioni in cui valori diversi non potranno che essere tutelati in maniera imperfetta.

CONTRIBUTO DEGLI AUTORI

Tutti gli autori hanno contribuito in egual misura al lavoro. In particolare, Fabio Fossa ha scritto la sezione *La spiegabilità e il suo contesto*, Giacomo Zanotti ha scritto la sezione *Tra spiegazione e comprensione*, Stefano Canali ha scritto la sezione *Il caso della medicina*. Le sezioni *Introduzione* e *Conclusioni* sono frutto del lavoro congiunto dei tre autori.

¹⁸ Alex J. London, "Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability", *Hastings Center Report*, 49, 1 (2019): 15-21. DOI: 10.1002/hast.973.

PARTE III

NORME GIURIDICHE E SOCIALI AI TEMPI DELL'INTELLIGENZA ARTIFICIALE

L'AI Act e la trasparenza, tra tracciabilità, spiegabilità e conoscibilità. Un nuovo tassello nell'ecosistema regolatorio della ricerca?

Carlo Casonato, Marta Fasan, Marta Tomasi

Università degli Studi di Trento

SCOPO DELLO SCRITTO

Nell'aprile del 2021, l'Unione Europea e i suoi Stati Membri hanno iniziato a discutere su una prima versione di quello che, nel giugno del 2024, è stato approvato come il primo regolamento al mondo che, con approccio comprensivo ed orizzontale, stabilisce regole armonizzate sull'intelligenza artificiale (d'ora in poi AI Act¹).

Questo contributo vuole sintetizzarne, in una prima parte, i passaggi più significativi, mettendone in luce alcuni punti di forza e di debolezza. In particolare, saranno descritte le ragioni alla base dell'approvazione dell'AI Act, alcune delle sue definizioni più importanti, gli ambiti compresi dal suo raggio d'azione, la sua logica e struttura di base. In una seconda parte, il contributo si rivolgerà al mondo della ricerca, tenendo presente che l'intelligenza artificiale (d'ora in poi IA) può esserne tanto l'oggetto di ricerca e sviluppo da parte di informatici, ingegneri informatici, roboticisti, ecc., quanto uno strumento utilizzabile da ricercatori di qualsiasi settore per raggiungere con maggior velocità o accuratezza i risultati delle proprie indagini. Che si dia il primo o il secondo caso, può essere interessante dare avvio a una riflessione intorno a un concetto cruciale tanto per la tecnologia in gioco, quanto per lo specifico ambito nel quale essa si trova collocata: la 'trasparenza'. Nonostante negli ultimi anni il dibattito scientifico abbia dato maggior risalto alla nozione di 'spiegabilità' quale possibile chiave di lettura da usare per affrontare i problemi di opacità tipici delle forme di IA più avanzate, il ricorso alla categoria giuridica

¹ Regolamento (UE) 2024/1689 del Parlamento europeo e del Consiglio relativo all'intelligenza artificiale (IA) e che modifica alcune normative dell'Unione.

della ‘trasparenza’ offre, ad avviso di chi scrive, un più ampio ventaglio di soluzioni e strumenti per provare a rispondere alle numerose questioni poste dall’ingresso dell’IA nel settore in esame². IA e ricerca, infatti, condividono una domanda comune rispetto a tale principio, il quale rappresenta, al contempo, una delle più complesse sfide per i nuovi sistemi intelligenti e uno degli irrinunciabili fondamenti dell’integrità della ricerca. Come si avrà modo di osservare, la ricerca che coinvolge l’IA rappresenta un campo di prova per la tenuta del principio di trasparenza che deve vestirsi di nuovi significati e declinazioni per restare fedele alla sua radice costitutiva.

LE RAGIONI DELL’AI ACT

La base legale dell’AI Act è costituita (e non poteva essere altrimenti) dall’art. 114 del *Trattato sul Funzionamento dell’Unione Europea* che permette agli organi UE di adottare misure tese al riavvicinamento delle disposizioni degli Stati Membri «che hanno per oggetto l’instaurazione ed il funzionamento del mercato interno» (cons. 3)³. Da questo punto di vista, l’AI Act si inserisce all’interno del *New Legislative Framework* adottato dalla Commissione al fine di rendere più funzionale il mercato unico⁴, ponendosi come l’ennesimo tassello di una *product compliance regulation* tesa a scongiurare la frammentazione del mercato e favorire la certezza del diritto, in modo da sostenere la libera circolazione e l’innovazione. In tale contesto, rileva anche la natura dello strumento normativo prescelto, un regolamento il quale, a differenza di una direttiva, è direttamente applicabile in tutti gli Stati Membri.

² Come si avrà modo di osservare nelle prossime pagine, la scelta di adottare questa chiave di lettura trova giustificazione, da un lato, nella possibilità offerta dal principio di trasparenza di fornire soluzioni non solo in termini di spiegabilità, ma anche di tracciabilità e di conoscibilità dei sistemi di IA, con l’opportunità, quindi, di affrontare con più ampia portata i problemi che possono emergere dall’uso dell’IA nel contesto della ricerca. Dall’altro lato, è opportuno considerare come il riferimento al principio di trasparenza trovi spazio nei principali atti normativi elaborati e adottati in materia di IA, come dimostrato anche da quanto previsto dall’AI Act.

³ In riferimento ai dati di carattere personale, i considerando menzionano anche la base legale rappresentata dall’art. 16 del TFUE, che ne dispone la protezione e libera circolazione.

⁴ Tale approccio normativo si propone di migliorare la sorveglianza sul mercato per contrastare la commercializzazione di prodotti difettosi o che possano costituire un pericolo per la salute e l’ambiente. Prevede, inoltre, procedure per l’accreditamento di organi atti alla verifica di conformità dei prodotti oltre che le modalità di impiego del marchio CE. Cfr. European Commission, “New Legislative Framework”. https://single-market-economy.ec.europa.eu/single-market/goods/new-legislative-framework_en.

Leggendo i numerosi ‘considerando’ (*recitals*) che anticipano l’articolato dell’AI Act (ben 180), emerge come il motivo per adottare una disciplina in materia non si limiti all’intenzione di assicurare regole uniformi per la circolazione di beni e servizi, ma derivi anche dal carattere del tutto peculiare dell’IA. La sua natura fortemente pervasiva e trasformativa, infatti, ne ha fatto e sempre più ne farà una componente centrale nello svolgimento di moltissime attività, con un impatto dirompente sugli individui e sulle nostre società, sui sistemi economici e di lavoro, sull’ambiente, sulle democrazie e sugli assetti di potere (pubblico e privato), sulla protezione dei diritti, oltre che sui possibili nuovi significati che potremo attribuire a termini come umanità e identità⁵. I motivi alla base dell’AI Act, quindi, non si riducono a garantire l’immissione nel mercato di prodotti sicuri, come capita per ogni tipo di prodotto certificato (con la marcatura CE), ma si spingono, nei limiti delle competenze dell’Unione, a indirizzarne lo sviluppo verso forme di IA affidabile e antropocentrica (*trustworthy and human-centered*), con il «fine ultimo di migliorare il benessere degli esseri umani» (cons. 6). La sfida dell’IA è di carattere propriamente esistenziale e l’AI Act tenta di affrontarla, per quanto possa farlo uno strumento giuridico europeo, in termini di protezione della salute, della sicurezza e dei diritti fondamentali, e puntando ad un rafforzamento della democrazia e dello Stato di diritto (cons. 8)⁶.

DEFINIZIONI E RAGGIO D’AZIONE DELL’AI ACT

Uno dei primi compiti su cui gli esperti e i politici che hanno redatto l’AI Act hanno dovuto concentrare i propri sforzi è stato quello definitorio. Ne è scaturito un lungo articolo, il terzo, che contiene la descrizione di ben 68 voci.

Un sistema di IA, in primo luogo, è stato definito come «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,

⁵ I riferimenti al riguardo sono innumerevoli. Fra gli altri, Luciano Floridi, *Etica dell’intelligenza artificiale. Sviluppì, opportunità, sfide*, Milano: Raffaello Cortina editore, 2022; Carlo Casonato, “Intelligenza artificiale e identità”, in *Enciclopedia Italiana Treccani*, Roma: Istituto Enciclopedia Italiana, XI appendice, vol. 2, 2024, 52-57.

⁶ La tutela ambientale, rilevante in alcune versioni intermedie del progetto di AI Act e presente in alcuni considerando, non ha peraltro trovato un riconoscimento significativo nella versione definitiva dell’articolato.

contenuti, raccomandazioni o decisioni che possono influenzare ambienti fisici o virtuali». In questo senso, allineandosi alla definizione già fornita dal Consiglio d'Europa, sono stati individuati quattro caratteri connotativi: un certo grado di (i) autonomia e di (ii) adattabilità, (iii) la capacità di influenzare il contesto e (iv) quella riconducibile alla deduzione (o inferenza). A questo riguardo, si sarebbe potuto fare maggior riferimento nella traduzione italiana dell'AI Act alla terminologia utilizzata nella versione inglese: in particolare, il termine *inference* avrebbe potuto essere più fedelmente tradotto utilizzando anche in italiano il verbo 'inferire'. Il considerando 12 dell'AI Act, su questa linea, riconduce proprio alle tecniche che consentono l'inferenza gli «approcci di apprendimento automatico che imparano dai dati come conseguire determinati obiettivi e approcci basati sulla logica e sulla conoscenza che traggono inferenze dalla conoscenza codificata o dalla rappresentazione simbolica del compito da risolvere. La capacità inferenziale di un sistema di IA trascende l'elaborazione di base dei dati consentendo l'apprendimento, il ragionamento o la modellizzazione».

Tale definizione pare di cruciale importanza, visto che consente di individuare quali saranno i sistemi che, soggetti al Regolamento, dovranno rispettarne i requisiti e quali invece, non presentando le quattro caratteristiche menzionate, ne saranno esenti.

Un secondo gruppo di definizioni, fra le molte che si potrebbero citare, riguarda le figure del fornitore e dell'utilizzatore. La prima riconduce al fornitore (*provider*) ogni persona fisica o giuridica, o qualsiasi organismo, che sviluppi o faccia sviluppare IA per finalità generali e la immetta nel mercato con il proprio nome o marchio, a titolo oneroso o gratuito⁷. Tale definizione è importante perché individua la figura cui è imposta la maggior parte degli obblighi previsti dall'AI Act, compresi quelli – vedremo – relativi alla costruzione di algoritmi che assicurino, nella misura del possibile, un certo grado di trasparenza. L'utilizzatore (*deployer*), invece, viene definito come la persona fisica o giuridica, o l'organismo, che utilizza un sistema di IA sotto la propria autorità. A questa figura sono imposti alcuni obblighi relativi, per quanto più rileva in questa sede, a una specifica declinazione della trasparenza art. 50, AI Act.

⁷ La definizione riporta: «una persona fisica o giuridica, un'autorità pubblica, un'agenzia o un altro organismo che sviluppa un sistema di IA o un modello di IA per finalità generali o che fa sviluppare un sistema di IA o un modello di IA per finalità generali e immette tale sistema o modello sul mercato o mette in servizio il sistema di IA con il proprio nome o marchio, a titolo oneroso o gratuito».

Fondamentale rilevanza assume anche l'individuazione del raggio d'azione, per così dire, dell'AI Act. A fronte della possibilità di stabilire regole diverse per i diversi ambiti di utilizzo, il Regolamento ha adottato un approccio comprensivo e orizzontale, destinato a disciplinare la creazione e l'utilizzo dei sistemi e modelli di IA in termini generalissimi.

Le regole previste, anzitutto, si applicano a tutta l'IA che voglia circolare sul territorio europeo. Ciò significa che i fornitori ne sono vincolati a prescindere dal fatto che siano stabiliti nell'Unione o in un paese terzo e che i loro prodotti debbano rispettare i requisiti richiesti dall'AI Act per il solo fatto di circolare all'interno del mercato europeo. In questo senso, l'Unione ha scommesso sul c.d. effetto Bruxelles: l'influenza esercitata in modo "discreto" a livello globale da parte della normativa europea riconducibile al fatto che tutti i produttori, anche di Paesi terzi, che non vogliono rinunciare a vendere all'interno del mercato unico dovranno fornire sistemi che siano rispettosi dei requisiti previsti dalla UE, adeguandosi così, in modo indiretto, anche ai relativi valori⁸.

Come anticipato, tutti i sistemi compresi nella definizione sopra illustrata sono oggetto del Regolamento, ad eccezione di quelli utilizzati in alcune aree e per alcuni scopi. L'AI Act, in primo luogo, non si applica, né potrebbe applicarsi, a settori che non rientrino nell'ambito di competenza del diritto dell'Unione (art. 2.3). Non essendo l'Unione Europea un'organizzazione dotata di competenza generale, a differenza degli Stati membri, il suo diritto si estende solo alle materie per cui i Trattati le conferiscono competenza, in conformità al principio di attribuzione (art. 5 TUE)⁹. Esclusi dagli obblighi previsti in generale per l'utilizzatore, sono, in secondo luogo, le persone fisiche che impiegano sistemi di IA nel corso di un'attività non professionale, di carattere puramente personale (art. 2.10). Fuori dal raggio d'azione dell'AI Act ricadono anche i sistemi utilizzati esclusivamente per scopi militari, di difesa o di sicurezza nazionale (art. 2.3)¹⁰ e, di particolare importanza nella prospettiva di questo contributo e come si vedrà meglio più avanti, i sistemi

⁸ Anu Bradford, *The Brussels effect: how the European Union rules the world*, Oxford: Oxford University Press, 2020.

⁹ In questo senso, sono esclusi dall'AI Act i sistemi utilizzati, ad esempio, dalle pubbliche amministrazioni che svolgano attività su materie non coperte dal diritto dell'Unione o dai professionisti della salute all'interno della relazione medico-paziente.

¹⁰ Maggiori dettagli nel considerando 24, AI Act.

e i modelli specificamente sviluppati e messi in servizio al solo scopo di ricerca e sviluppo scientifici (art. 2.6). Inoltre, non volendo correre il rischio di pregiudicare la libertà della scienza, il Regolamento esclude la propria applicabilità alle attività di ricerca, prova o sviluppo relative a sistemi di IA o modelli di IA prima della loro immissione sul mercato o messa in servizio. Tale eccezione, coerentemente con l'approccio basato sul rischio (su cui *infra*), non copre le prove in condizioni reali (art. 2.8)¹¹.

LOGICA E STRUTTURA DELL'AI ACT

L'AI Act si muove all'interno di una logica basata sul rischio, con regole proporzionate rispetto al grado di pericolo emergente. Ciò significa che i sistemi sono divisi non sulla base delle loro caratteristiche tecniche (*Model-Based v. Machine Learning*, ad esempio) né prioritariamente sulla base degli ambiti di utilizzo, ma – appunto – sulla base della gravità di rischio che comporta il loro impiego.

In questo senso, il Regolamento distingue sistemi che presentano (a) rischi inaccettabili, (b) rischi di grado alto, (c) rischi legati prevalentemente alla trasparenza-conoscibilità (d) rischi minimi. Nell'ultima fase di negoziazione del regolamento, a seguito della diffusione di ChatGPT, è stata inoltre inserita un'ulteriore categoria di rischio, legata ai (e) sistemi per finalità generali (*General Purpose Artificial Intelligence: GPAI*).

Alcuni sistemi (sub a), presentando un chiaro pericolo per la salute, la sicurezza e i diritti fondamentali delle persone, sono considerati inaccettabili e banditi dal mercato unico a partire dal 2 febbraio 2025. Fra questi, a titolo esemplificativo, sono ricondotti: (i) sistemi che, attraverso l'impiego di tecniche subliminali, distorcono il comportamento di una persona, pregiudicandone la capacità di prendere decisioni consapevoli e comportando il rischio di un danno significativo; (ii) sistemi che sfruttano la vulnerabilità delle persone (per età o disabilità, ad esempio), distorcendone il comportamento, con il rischio di subire danni significativi; (iii) sistemi di classificazione delle persone basati sul comportamento o su caratteristiche personali che conducano ad un trattamento sfavorevole sproporzionato o in contesti scollegati da quelli di raccolta dei dati; (iv) sistemi che permettano di creare banche dati

¹¹ Maggiori dettagli nel considerando 25, AI Act.

di riconoscimento facciale mediante scraping indiscriminato di immagini facciali da internet; (v) sistemi di riconoscimento delle emozioni in ambito lavorativo o di istruzione (con eccezioni); (vi) alcuni sistemi di identificazione biometrica remota in tempo reale in spazi accessibili al pubblico¹².

Altri sistemi (sub c) presentano rischi limitati alla possibilità che le persone che interagiscono con essi si confondano sulla loro natura (umana o artificiale), non riuscendo a distinguerli dalle persone. Il progresso tecnologico, infatti, ha fatto in modo che sempre più numerosi siano i sistemi talmente avanzati da superare il c.d. test di Turing e da essere considerati come portatori o imitatori di un comportamento connotato da intelligenza¹³. A fronte di questa eventualità, il Regolamento europeo ha posto in capo ai fornitori un dovere di progettazione tale che le persone siano informate del fatto di stare interagendo con una macchina e che i testi, gli audio, le immagini e i video creati da IA siano marcati e chiaramente riconoscibili come generati artificialmente (art. 50.1 e 50.2). Agli utilizzatori (*deployers*) di sistemi che creano contenuti *deep fake*, inoltre, il regolamento pone l'obbligo di rendere conoscibile la natura artificiale degli stessi (art. 50.4). Tali doveri entreranno in vigore a partire dall'agosto del 2026.

Nel caso in cui si tratti di sistemi che non pongono rischi di rilievo, il Regolamento non impone alcun obbligo, invitando comunque i fornitori ad adottare codici di condotta, che includano meccanismi di governance volti a promuovere l'applicazione volontaria dei requisiti di volta in volta funzionali a evitare qualsiasi rischio anche minimo (art. 95)¹⁴.

Invece, decisamente più complessa ed articolata risulta la disciplina riservata ai sistemi ad alto rischio (sub b). Va anzitutto menzionato come l'Unione Europea abbia deciso di dotare di un certo grado di flessibilità e adattabilità nel tempo i criteri in base ai quali individuare tali sistemi e la lista delle materie interessate. Tale approccio *future proof* è dovuto al rapido incedere delle innovazioni nel settore dell'IA che avrebbe condotto liste e criteri eccessivamente rigidi a diventare rapidamente obsoleti e anacronistici. Per questo, le indicazioni utili ad individuare quali siano i sistemi ad alto

¹² La lista completa dei sistemi vietati all'art. 5 dell'AI Act.

¹³ Si tratta del test che permette di determinare, secondo Alan Turing, se una macchina abbia la capacità di esibire o imitare un comportamento intelligente: cfr. Alan M. Turing, "Computing machinery and intelligence", *Mind*, 49 (1950): 433-460.

¹⁴ Maggiori dettagli nel considerando 165, AI Act.

rischio non sono presenti all'interno del Regolamento stesso, che prevede per la propria revisione procedure particolarmente complesse e dispendiose in termini di tempo, ma sono riportate all'interno di due allegati (*annexes I e III*), indicati dall'art. 6, i quali possono essere modificati ed aggiornati più agilmente dalla sola Commissione. L'allegato terzo, al riguardo, indica come sistemi ad alto rischio quelli impiegati a fini biometrici, in ambito di istruzione, formazione e lavoro, per l'accesso ai servizi essenziali, per le attività di contrasto (come poligrafi, strumenti di valutazione dell'affidabilità probatoria o del rischio di commissione di reato o di recidiva), in ambito di migrazione e controllo delle frontiere, nell'amministrazione della giustizia (strumenti di assistenza per la ricerca o interpretazione dei fatti o del diritto applicabile) e all'interno dei processi democratici (come quelli che potrebbero influenzare gli esiti elettorali). Tali sistemi dovranno conformarsi alle previsioni dell'AI Act entro l'agosto del 2026.

Nell'allegato primo, che sarà applicabile a partire dall'agosto 2027, sono invece caratterizzati come ad alto rischio i sistemi impiegati all'interno di una numerosa serie di apparecchiature già oggetto di disciplina europea, fra cui giocattoli, imbarcazioni, veicoli a motore e dispositivi medici. In riferimento all'individuazione dei sistemi ad alto rischio, infine, va ricordato come l'AI Act preveda una deroga riferita ai sistemi indicati nell'allegato terzo che non presentino un rischio significativo per la salute, la sicurezza o i diritti fondamentali delle persone, anche nel senso di non influenzare materialmente il risultato del processo decisionale. Nonostante l'art. 6.3 contenga alcune esemplificazioni dei casi coperti dall'eccezione (sistemi che eseguano un compito preparatorio o procedurale limitato, quelli destinati a migliorare il risultato di un'attività umana precedentemente completata, quelli non finalizzati a sostituire o influenzare una valutazione precedentemente completata senza un'adeguata revisione umana), tale disposizione pone problemi interpretativi complessi, potendo essere sfruttata strumentalmente al fine di evitare il rispetto degli impegnativi requisiti imposti ai sistemi ad alto rischio.

Una volta che un sistema sia ricondotto a tale categoria, decisione che peraltro compete allo stesso fornitore, l'AI Act prevede, infatti, il rispetto di una serie di condizioni non agevoli. Si tratta, anzitutto, della creazione di un sistema di gestione dei rischi, in grado di identificarli, valutarli e mitigarli (art. 9). In riferimento ai dataset utilizzati per l'addestramento, la convalida e la

prova, in secondo luogo, il Regolamento prevede la pertinenza, la sufficiente rappresentatività e, nei limiti del possibile, l'esattezza e la completezza, raccomandando l'utilizzo di proprietà statistiche appropriate (art. 10). In questo modo, si tenta di scongiurare in partenza la generazione di output scorretti o discriminatori. Oltre alla redazione della documentazione tecnica (art. 11), l'AI Act raccomanda la registrazione dei *log* per tutta la durata del ciclo di vita del sistema (art. 12), nel tentativo di garantire la tracciabilità del funzionamento dei sistemi ad alto rischio. In base all'art. 13 del Regolamento, inoltre, i fornitori sono destinatari di un obbligo teso a garantire che il funzionamento dei loro sistemi sia sufficientemente trasparente, in modo da permettere all'utilizzatore, anche sulla base di concise, complete, corrette e chiare informazioni d'uso, di interpretare gli output prodotti e di utilizzarli in modo appropriato. Tale previsione si collega all'art. 86 dell'AI Act che dispone il diritto della persona interessata da una decisione assunta con l'assistenza dei sistemi ad alto rischio elencati all'allegato terzo ad ottenere dall'utilizzatore spiegazioni chiare e significative sul ruolo assunto dall'IA e sui principali elementi della decisione adottata. Tali requisiti – come si vedrà *infra* – non sono di facile concretizzazione, visto soprattutto il problema della c.d. *Black Box*¹⁵.

Un ultimo requisito fondamentale per quanto riguarda tale categoria di sistemi si collega al principio del c.d. *human in the loop* o della sorveglianza umana (art. 14). L'AI Act, così, pone in capo ai fornitori l'obbligo di sviluppare sistemi la cui struttura permetta una supervisione umana, finalizzata alla mitigazione dei rischi per la salute, la sicurezza e i diritti fondamentali. A tale obbligo corrisponde, in capo all'utilizzatore, l'onere di contrastare la tendenza a fare eccessivo affidamento sulla correttezza dell'output (*automation bias*), e di giungere alla comprensione delle potenzialità e dei limiti del sistema, in modo da poterne individuare anomalie e disfunzioni e discostarsi da ogni output ritenuto scorretto o discriminatorio. Se quello della sorveglianza umana è uno dei cardini tanto del diritto quanto dell'etica dell'IA, va ricordato come la sua effettiva e concreta realizzazione dipenda da presupposti non sempre facili da riscontrare. Si tratta, da un lato, di una competenza tecnica diffusa che permetta agli utilizzatori di comprendere e interpretare il funzionamento di sistemi caratterizzati da un'estrema

¹⁵ Frank Pasquale, *The Black Box Society*, New Haven: Harvard University Press, 2016.

complessità¹⁶; dall'altro, di una non facile assunzione di responsabilità personale che richiede tempo, impegno e forte motivazione individuali così come forme di supporto a livello strutturale¹⁷.

Molte altre sono le disposizioni che completano la struttura dell'AI Act (che conta, oltre a 180 considerando, 13 allegati e 113 articoli). Quanto illustrato, tuttavia, pare sufficiente a fornire il quadro e le premesse necessarie a svolgere un'adeguata trattazione della complessa applicabilità del principio di trasparenza (*rectius* di alcune sue declinazioni, riconducibili a spiegabilità, conoscibilità e tracciabilità) all'ambito delle attività di ricerca.

INTELLIGENZA ARTIFICIALE, RICERCA E TRASPARENZA. POSSIBILI CHIAVI DI LETTURA ALLA LUCE DELL'AI ACT

L'inquadramento normativo offerto dall'AI Act in materia di sviluppo e uso dei sistemi di IA si caratterizza, quindi, per una portata applicativa ampia e generale, tale da produrre effetti su numerosi settori e contesti di applicazione. Tra questi rientra anche l'ambito della ricerca, all'interno del quale l'IA, come si è già avuto modo di anticipare, può assumere rilevanza sia come oggetto, sia come strumento di questa attività. Infatti, nonostante il Legislatore europeo abbia predisposto una clausola volta a escludere l'applicazione delle disposizioni dell'AI Act ai sistemi e ai modelli sviluppati per sole finalità di ricerca e non si applica alle attività di ricerca, prova o sviluppo relative a sistemi o modelli prima della loro immissione sul mercato o messa in servizio, tale esenzione non deve considerarsi assoluta. Come indicato in precedenza, possono sussistere situazioni e circostanze in cui l'attuazione delle regole dell'AI Act si rende necessaria anche in relazione all'attività di ricerca, soprattutto nei casi in cui questa si svolga in condizioni reali e con il coinvolgimento di persone¹⁸.

¹⁶ Al riguardo, l'art. 4 dell'AI Act prevede che fornitori e utilizzatori adottino misure adeguate a garantire al proprio personale un livello sufficiente di alfabetizzazione in materia di IA.

¹⁷ Si potrebbe provocatoriamente affermare che in questo modo si richieda l'azione di una sorta di supereroe: cfr. Carlo Casonato, "Unlocking the Synergy: Artificial Intelligence and (old and new) Human Rights", *BioLaw Journal – Rivista di BioDiritto*, 3 (2023): 233-240.

¹⁸ L'opportunità di applicare le regole dell'AI Act a questi contesti è evidenziata, oltre che dal considerando 25, anche da quanto stabilito dagli artt. 2, par. 8, e 60, Regolamento (UE) 2024/1689. A questo proposito cfr. Liane Colonna, "The AI Act's Research Exemption: A Mechanism for Regulatory Arbitrage?", in *YSEC Yearbook of Socio-Economic Constitutions 2023*, a cura di Eduardo Gill-Pedro, Andreas Moberg, Cham: Springer, 2024, 51 ss.

Tale rilievo pone di fronte alla necessità di comprendere quali disposizioni dell'AI Act possano assumere importanza in relazione a questo specifico contesto, con particolare riferimento a uno dei concetti che, sia dal punto di vista giuridico sia dalla prospettiva etica, si colloca come elemento cardine nello svolgimento delle attività di ricerca: il principio di 'trasparenza'.

Questo principio, che nella prospettiva dello Stato costituzionale di diritto si pone in un rapporto di stretta correlazione con i concetti di separazione dei poteri e di principio democratico¹⁹, trova fondamento nell'idea che la tutela delle persone e dei loro interessi sia assicurata anche grazie alla possibilità di rendere accessibili e di diffondere le informazioni che riguardano l'esercizio di una forma di potere, consentendone le necessarie limitazioni e modalità di controllo²⁰.

Tale assunto si concretizza anche nella dimensione della ricerca dove, tradizionalmente, la condivisione dei dati costituisce una garanzia fondamentale sia per assicurare l'attendibilità, la riproducibilità e la verificabilità dei risultati prodotti, sia per tutelare i diritti e le libertà delle persone coinvolte come partecipanti in questa attività. Pure in questo caso, la conoscenza e la disponibilità delle informazioni risultano essere fattori fondamentali nel bilanciare situazioni di squilibrio di potere che possono emergere alla luce di diversi soggetti e dei molteplici interessi che in questo campo trovano affermazione e nel garantire il rispetto dei requisiti di integrità che devono permeare il campo della ricerca²¹.

Queste considerazioni circa la portata e gli obiettivi perseguiti nell'implementazione del principio di trasparenza assumono ancor più importanza in relazione a una tecnologia come l'IA. Che venga in gioco come oggetto o come strumento della ricerca, l'IA pone numerose questioni che sfidano i contenuti fondativi del concetto stesso di trasparenza e prospettano

¹⁹ Norberto Bobbio, *Il futuro della democrazia*, Torino: Einaudi, 2013.

²⁰ Mark Fenster, "Seeing the State: the Transparency as Metaphor", *Administrative Law Review*, 62, 3 (2010): 617-672; Deirdre Curtin, Albert J. Meijer, "Does transparency strengthen legitimacy?", *Information Polity*, 11, 2 (2006): 109-122; Enrico Carloni, *Il paradigma della trasparenza. Amministrazioni, informazione, democrazia*, Bologna: Il Mulino, 2022, 61 ss.; Enzo Cheli, "Informazione, decisione politica, controllo sociale: spunti per un'analisi comparata", *Il diritto dell'informazione e dell'informatica*, 3 (198): 813-824.

²¹ Lucia Busatta, "L'integrità della ricerca nel tessuto costituzionale: prime notazioni a partire dal contesto pandemico", *Rivista AIC*, 4 (2020): 389 ss.

profili tali da richiedere nuove chiavi di lettura²². In tale senso, l'analisi di quanto previsto dall'AI Act consente di leggere il significato di trasparenza alla luce delle esigenze poste dai sistemi di IA e di individuarne eventuali nuove declinazioni interpretative. Il ragionamento, tuttavia, va condotto alla luce del diverso inquadramento normativo offerto dal Regolamento europeo qualora l'IA sia oggetto o strumento della ricerca.

L'IA COME OGGETTO E COME STRUMENTO DELLA RICERCA: LE NUOVE DECLINAZIONI DEL PRINCIPIO DI TRASPARENZA DALL'AI ACT AI CODICI DI CONDOTTA

La prima ipotesi di lettura del principio di trasparenza trae origine dallo scenario in cui l'IA costituisce l'oggetto delle attività di ricerca, in una fase di sviluppo che si colloca idealmente in un momento immediatamente precedente all'immissione in commercio e alla messa in servizio del sistema di IA²³. Da questa prospettiva, le disposizioni dell'AI Act applicabili e riconducibili al concetto di trasparenza sono molteplici e trasversali rispetto alla classificazione basata sul rischio adottata dal Regolamento, a dimostrazione dell'importanza riconosciuta a questo principio nel processo di regolamentazione dell'IA. Ciò che però rileva maggiormente sono le diverse declinazioni che vengono date al principio di trasparenza nel tentativo di elaborare soluzioni ai problemi posti dai sistemi di IA. Nello specifico, tre sono le principali linee di lettura che possono essere date al principio in esame alla luce di quanto stabilito dall'AI Act²⁴.

²² Erik Longo, Andrea Pin, "Oltre il costituzionalismo? Nuovi principi e regole costituzionali per l'era digitale", *Diritto pubblico comparato ed europeo*, 1 (2023): 103-116. A questo proposito si consenta il rinvio a Marta Fasan, *Intelligenza artificiale e costituzionalismo contemporaneo. Principi, diritti e modelli in prospettiva comparata*, Napoli: Editoriale Scientifica, 2024, 114 ss.

²³ Sul punto si veda quanto ribadito al considerando 25, Regolamento (UE) 2024/1689 e quando affermato in Amedeo Santosuosso, Giovanni Sartor, *Decidere con l'IA. Intelligenze artificiali e naturali del diritto*, Bologna: Il Mulino, 2024, 182 ss.; Marco Bassini, "Oggetto, campo di applicazione e ambito territoriale", in *La disciplina dell'intelligenza artificiale*, a cura di Oreste Pollicino et al., Milano: Giuffrè, 2025, 133 ss. È opportuno sottolineare come le regole introdotte dall'AI Act non si applichino ai sistemi di IA rilasciati con licenza libera e open source, a meno che questi non vengano immessi sul mercato come sistemi classificabili come proibiti, ad alto rischio o con problemi di trasparenza, secondo quanto previsto dall'art. 2, par. 12, Regolamento (UE) 2024/1689.

²⁴ La necessità di attribuire nuovi significati al principio di trasparenza era stata evidenziata anche nelle "Ethics Guidelines for Trustworthy AI", High Level Expert Group on Artificial Intelligence (2019). <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>.

La prima declinazione che può essere data al concetto di trasparenza in considerazione di quanto disposto dall'AI Act si pone in termini di 'spiegabilità' del sistema di IA. Con questo termine si fa riferimento, nello specifico, alla necessità di fornire spiegazioni sulle modalità di funzionamento tecnico del sistema intelligente e sulle motivazioni poste alla base dei risultati elaborati grazie all'apporto di questa tecnologia, così da assicurare tutela alle persone interessate dalle conseguenze e dagli effetti di decisioni adottate da e con l'IA²⁵. In tal senso, il semplice accesso alle informazioni sulla logica decisionale e sui procedimenti tecnici seguiti dal sistema non è di per sé una misura sufficiente a fornire le garanzie tipiche della trasparenza, ma si richiede anche la capacità di comprendere questi profili e di saper dare una spiegazione in merito alle ragioni e agli obiettivi perseguiti dal sistema intelligente durante il suo funzionamento²⁶. Solo grazie a interventi orientati ad assicurare la spiegabilità dell'IA è, infatti, possibile provare a realizzare una minimizzazione dei rischi derivanti dal già menzionato fenomeno della c.d. *Black Box* e dal suo conseguente impatto sulla vita delle persone.

L'importanza di questa declinazione trova, per l'appunto, conferma in alcuni contenuti dell'AI Act che si pongono proprio nella direzione di prevedere obblighi e requisiti funzionali a tale scopo. Così, sono riconducibili al concetto di 'trasparenza-spiegabilità' le disposizioni che per i sistemi 'ad alto rischio' prevedono una progettazione e uno sviluppo tali da assicurare che l'utilizzatore sia in grado di conoscere il funzionamento del sistema di IA e di interpretarne i risultati per impiegare la tecnologia in modo adeguato²⁷. Queste previsioni sono, poi, ulteriormente rafforzate dall'obbligo di predisporre istruzioni per l'uso che permettano al c.d. *deployer* di comprendere le funzionalità offerte dall'IA e di impiegare il sistema adeguatamente secondo le sue finalità d'utilizzo²⁸. L'esigenza di

²⁵ Luciano Floridi, *op. cit.*, 100 ss.; Karen Yeung, "Why Worry about Decision-Making by Machine?", in *Algorithmic Regulation*, a cura di Karen Yeung, Martin Lodge, Oxford: Oxford University Press, 2019, 28-29; Monica Palmirani, "Big Data e conoscenza", *Rivista di filosofia del diritto*, 1 (2020): 74 ss.; Elisa Spiller, "Il diritto di comprendere, il dovere di spiegare. Explainability e intelligenza artificiale costituzionalmente orientata", *BioLaw Journal – Rivista di BioDiritto*, 2 (2021): 419-432.

²⁶ Elettra Stradella, "AI, tecnologie innovative e produzione normativa: potenzialità e rischi", *DPCE online*, 3 (2020): 3361 ss.; Madalina Busuioc, Deirdre Curtin, Marco Almada, "Reclaiming transparency: contesting the logics of secrecy within the AI Act", *European Law Open*, 2, 1 (2023): 79-105.

²⁷ Art. 13, par. 1, Regolamento (UE) 2024/1689.

²⁸ Art. 13, par. 2, Regolamento (UE) 2024/1689.

fornire agli utenti le opportune spiegazioni per capire il funzionamento dell'IA e il significato dei suoi risultati si dimostra ancor più cruciale nelle situazioni in cui si richieda un intervento umano. Infatti, anche nell'attuazione delle misure di sorveglianza umana previste dall'AI la corretta comprensione del funzionamento del sistema e la capacità di interpretare i risultati da questo generati, fornendone una spiegazione, sono elementi essenziali per l'esercizio di un controllo attento sull'impiego di questa tecnologia²⁹. È poi espressione di questa nuova declinazione del principio di trasparenza anche il riconoscimento del diritto alla spiegazione dei singoli processi decisionali. Nello specifico, l'AI Act prevede che le persone destinatarie di una decisione basata sui risultati prodotti da un sistema di IA, e da cui derivano effetti giuridici, hanno il diritto a ottenere dal *deployer* spiegazioni chiare e significative sul ruolo assunto dal sistema intelligente nel processo decisionale e sui principali elementi che contraddistinguono la decisione finale³⁰.

La seconda declinazione attribuibile al principio di trasparenza si pone, invece, in termini di 'conoscibilità' del sistema di IA. Questo concetto può essere interpretato secondo due diverse chiavi di lettura, da cui discendono obblighi e requisiti diversi alla luce dell'AI Act. Nella prima, la nozione di conoscibilità si radica nella garanzia che siano rese note e condivise le informazioni riguardanti sia la natura artificiale del sistema utilizzato, sia la presenza dello stesso in procedimenti e interazioni che possano produrre effetti sulle persone³¹. Nella seconda chiave di lettura, invece, il concetto di conoscibilità si incardina sulla necessità di rendere note ai soggetti interessati le caratteristiche dell'IA, secondo un'accezione di trasparenza legata, anche in questo caso, alla conoscenza delle modalità di funzionamento di tale tecnologia³². Queste declinazioni interpretative fanno, quindi, riferimento alla necessità di prevedere idonee garanzie affinché le persone siano consapevoli dell'intervento di un sistema di IA, ne conoscano le funzionalità e che, in conseguenza di ciò, abbiano la possibilità di scegliere liberamente se avvalersi di questa tecnologia e dei suoi risultati.

²⁹ Art. 14, par. 4, lett. a) e c), Regolamento (UE) 2024/1689.

³⁰ Art. 86, Regolamento (UE) 2024/1689.

³¹ Carlo Casonato, "Unlocking the Synergy: Artificial Intelligence and (old and new) Human Rights", *op. cit.*, 235 ss.

³² Virginia Dignum, *Responsible Artificial Intelligence. How to Develop and Use AI in a Responsible Way*, Cham: Springer, 2019, 60-62; Thomas Wischmeyer, "Artificial Intelligence and Transparency: Opening the Black Box", in *Regulating Artificial Intelligence*, a cura di Thomas Wischmeyer, Timo Rademacher, Cham: Springer, 2020, 95.

Tali finalità sono confermate, anche in questo caso, dalle disposizioni dell'AI Act che intervengono ad arginare le conseguenze negative di una mancata autodeterminazione in merito all'impiego e agli effetti dell'IA e danno, quindi, attuazione al concetto di trasparenza-conoscibilità nelle sue due accezioni. Così, l'AI Act stabilisce esplicitamente l'obbligo di informare le persone destinatarie di una decisione adottata con l'intervento dell'IA dell'uso di un sistema classificato ad alto rischio e della loro soggezione ai risultati che ne derivano, assicurando che sia resa nota e conosciuta la presenza dell'IA nel processo decisionale³³. Similmente, l'AI Act predispone garanzie di conoscibilità da applicare anche nei contesti di interazione diretta con il sistema di IA. In questo senso, si prevede che il sistema di IA sia progettato e sviluppato in modo tale che la persona, che interagisce direttamente con esso, sia consapevole di relazionarsi con un'IA e che i contenuti elaborati dai modelli di IA generativa abbiano natura artificiale³⁴. Sono, poi, espressione di questa seconda declinazione di trasparenza anche le disposizioni indirizzate ad assicurare la conoscenza delle caratteristiche tecniche del sistema prima del suo effettivo impiego. Che si tratti di sistemi di IA o di modelli di IA per finalità generali, l'AI Act stabilisce specifici obblighi sulla condivisione delle informazioni e delle specifiche tecniche che riguardano le capacità, le proprietà e i limiti del sistema utilizzato³⁵, anche prevedendo l'onere di redigere e di fornire la documentazione tecnica rilevante³⁶, con l'obiettivo di rendere gli utilizzatori maggiormente consapevoli delle funzionalità del sistema e, quindi, di come impiegare la tecnologia in modo appropriato secondo le sue finalità d'uso³⁷.

Infine, la terza declinazione data al principio di trasparenza nel quadro regolatorio fornito dall'AI Act si sostanzia nel concetto di 'tracciabilità'. Questo termine esprime la necessità che tutte le informazioni relative alla

³³ Art. 26, par. 11, Regolamento (UE) 2024/1689.

³⁴ Art. 50, Regolamento (UE) 2024/1689. Per un commento alla disposizione in esame e al suo contenuto normativo si veda Erik Longo, Federica Paolucci, "The Article 50 of the AI Act and the Transparency Obligations: the Model and its Limitations", in *La disciplina dell'intelligenza artificiale*, a cura di Oreste Pollicino et al., Milano: Giuffrè, 2025, 275 ss.

³⁵ Art. 13, par. 3, lett. b), Regolamento (UE) 2024/1689; Art. 53, par. 1, lett. b), Regolamento (UE) 2024/1689.

³⁶ Art. 11, Regolamento (UE) 2024/1689; Art. 53, par. 1, lett. a), Regolamento (UE) 2024/1689.

³⁷ Brent Mittelstadt et al., "The ethics of algorithms: Mapping the debate", *Big Data & Society*, 2 (2016): 6-7.

raccolta dei dati, alla loro classificazione, alla tipologia di algoritmo impiegato e riguardanti le modalità di adozione della decisione finale siano documentate correttamente, così da risultare tracciabili³⁸. La previsione di regole e requisiti orientati a implementare una tale accezione di trasparenza trova ragion d'essere, ancora una volta, nel bisogno di mantenere il controllo su alcuni elementi e operazioni che contraddistinguono il funzionamento dell'IA per evitare che insorgano rischi durante il suo impiego. In tal senso, il concetto di 'trasparenza-tracciabilità' è funzionale all'individuazione di *bias*, errori, lacune e imprecisioni nei dataset utilizzati, alla conoscenza dei dati usati e delle modalità della loro analisi e all'identificazione di eventuali malfunzionamenti durante il processo decisionale, permettendo di limitare i danni che possano derivare da questa tecnologia e di tutelare le persone e i loro diritti³⁹.

A questo scopo, l'AI Act introduce misure volte a monitorare la qualità dei dataset utilizzati, dal punto di vista dell'appropriatezza, della rappresentatività e della completezza delle informazioni usate, e a supervisionare le modalità di raccolta dei dati e le operazioni che riguardano il loro trattamento⁴⁰. Inoltre, si prevede l'implementazione di meccanismi di registrazione dei *log* e, in generale, degli eventi più rilevanti del funzionamento dell'IA così da poter tenere traccia di tutti i processi realizzati, garantendo in questo modo la possibilità di individuare eventuali errori e problematiche e, se possibile, intervenire per porvi rimedio⁴¹.

Le esigenze di trasparenza sin qui descritte in riferimento all'attività di ricerca che abbia ad oggetto l'IA, si rilevano anche nel contesto in cui l'IA costituisca uno degli strumenti usati dai ricercatori per lo svolgimento delle proprie attività. In questo scenario, però, l'impostazione adottata dall'AI Act restituisce un quadro regolatorio totalmente diverso da quanto esaminato in precedenza. Il mancato inserimento di tale attività tra gli ambiti d'uso previsti

³⁸ Joanna J. Bryson, Andreas Theodorou, "How society can maintain human-centric artificial intelligence", in *Human-centered digitalizations and services*, a cura di Marja Toivonen, Evelina Saari, Singapore: Springer, 2019, 317 ss.

³⁹ Madalina Busuioc, Deirdre Curtin, Marco Almada, *op. cit.*, 86 ss.

⁴⁰ Art. 10, par. 2 e 3, Regolamento (UE) 2024/1689. Un obbligo simile è previsto anche per il *deployer* all'art. 26, par. 4, Regolamento (UE) 2024/1689.

⁴¹ Art. 12, Regolamento (UE) 2024/1689. Una misura simile di registrazione dei log è prevista anche per i modelli di IA per finalità generali classificati con un rischio sistemico, secondo quanto previsto dall'art. 55, par. 1, lett. c), Regolamento (UE) 2024/1689. In generale, sugli specifici obblighi di trasparenza previsti per questi modelli di IA cfr. A. Santosuosso, G. Sartor, *op. cit.*, 183 ss.

dal Regolamento europeo implica, infatti, che il ricorso all'IA come strumento della ricerca non sia coperto dalla maggior parte delle regole previste dall'AI Act, ma si limiti alle disposizioni che abbiamo ricondotto al concetto di trasparenza-conoscibilità nell'interazione con il sistema.

Fatta eccezione per tali obblighi⁴², la disciplina di questo impiego dell'IA, e così anche l'eventuale declinazione del contenuto del principio di trasparenza, è affidata all'elaborazione di codici di condotta secondo quanto stabilito dall'AI Act⁴³. Il Regolamento, infatti, prevede che si debba incoraggiare e agevolare l'elaborazione di codici di condotta per disciplinare le applicazioni dell'IA non classificate come proibite, ad alto rischio o con particolari problemi di trasparenza e che tali codici debbano promuovere l'applicazione volontaria dei requisiti previsti dall'AI Act per le altre tipologie di sistemi intelligenti. Secondo tale disposizione, quindi, il contenuto normativo dei codici di condotta dovrebbe ricalcare quanto già stabilito dall'AI Act (in particolare, al capo III, sezione 2) anche per quanto riguarda l'attuazione del principio di trasparenza, lasciando presupporre l'adozione di declinazioni interpretative simili a quelle già esaminate per i sistemi di IA che siano strumenti della ricerca⁴⁴.

CONCLUSIONI. L'AI ACT E L'ECOSISTEMA REGOLATORIO DELLA RICERCA

In entrambi i contesti che si sono analizzati, quello in cui l'IA è oggetto della ricerca e quello in cui essa opera come strumento, le letture date in termini di trasparenza-spiegabilità, trasparenza-conoscibilità e trasparenza-tracciabilità costituiscono importanti garanzie di fronte ai potenziali rischi legati all'uso di questa tecnologia, spesso connessi alla sua opacità riconducibile all'imperscrutabilità delle sue ragioni, alle sue abilità di mimesi dell'umano e alla complessità dei percorsi che ne determinano il funzionamento. In effetti, conoscere le ragioni alla base di un determinato risultato, essere consapevoli della natura di IA della tecnologia impiegata e avere cognizione dei dataset impiegati e delle operazioni svolte dal sistema

⁴² Art. 50, Regolamento (UE) 2024/1689.

⁴³ Tale assunto deriva dal mancato inserimento del settore della ricerca quale ambito in cui l'uso dell'IA sia da considerarsi vietato o ad alto rischio, secondo l'inquadramento stabilito dagli artt. 5 e 6, Regolamento (UE) 2024/1689.

⁴⁴ Così previsto all'art. 95, Regolamento (UE) 2024/1689.

sono elementi essenziali per assicurare ai ricercatori e alle ricercatrici maggiore controllo sullo strumento utilizzato e, in definitiva, per garantire una ricerca di maggiore qualità⁴⁵. Al contempo, essere a conoscenza del coinvolgimento di un sistema artificiale, così come comprenderne i meccanismi procedurali di base, e degli obiettivi di una ricerca sono condizioni essenziali per garantire una piena e consapevole adesione dei partecipanti alla stessa.

È evidente, però, che nel quadro normativo tracciato dall'AI Act permangono, allo stato dell'arte, alcune lacune e significative difficoltà interpretative che derivano, in particolare, dai confini applicativi del testo, dalla menzionata natura *product-centric* della regolamentazione e dall'approccio basato sul rischio. Di conseguenza, le norme relative alle tre declinazioni di trasparenza rinvenibili nell'AI Act troveranno diversa (o nessuna) applicazione a seconda che si tratti di una ricerca 'su' un sistema di IA o 'con' un sistema di IA, che questo sia o non sia ad alto rischio, che sia o non sia sviluppato e applicato per finalità di ricerca, che si agisca prima o dopo l'immissione del sistema in commercio, che si stia operando in condizioni reali, che dall'impiego del sistema di IA derivino decisioni che impattano sulle persone, che il sistema sia classificabile come *medical device*, che ci si interroghi sugli obblighi facenti capo agli sviluppatori dei sistemi (*providers*) o ai ricercatori che li utilizzano per lo svolgimento di attività di ricerca (*deployers*).

A fronte di questo quadro complesso di variabili non si può non segnalare un dato che pare essenziale tenere a mente: l'AI Act non opera in un vuoto, ma all'interno di un complesso ecosistema etico-normativo. Di questo si trova traccia anche nel considerando 25 dell'AI Act, il quale, dopo aver tentato un bilanciamento fra il bisogno di non ostacolare il progresso delle attività di ricerca e l'esigenza di mantenere punti fermi, almeno in riferimento ai sistemi ad alto rischio, rammenta come «(...) In ogni caso, qualsiasi attività di ricerca e sviluppo dovrebbe essere svolta conformemente alle norme etiche e professionali riconosciute nell'ambito della ricerca scientifica e dovrebbe essere condotta conformemente al diritto dell'Unione applicabile».

Le declinazioni della trasparenza delle quali si è trattato corrispondono, in effetti, ad esigenze tipiche del mondo della ricerca che da anni hanno

⁴⁵ Sulla possibilità che trovino applicazione le regole dell'AI Act quando l'IA sia utilizzata come strumento per condurre attività di ricerca cfr. Marco Bassini, *op. cit.*, 137-138; Hannah Ruschemeier, *AI as a challenge for legal regulation – the scope of application of the artificial intelligence act proposal*, ERA Forum, 23 (2023): 370 ss.

trovato riconoscimento in diversi documenti di *hard* e *soft law*. Si pensi alle corrispondenze fra la trasparenza-conoscibilità e i doveri di informazione gravanti sui ricercatori nei confronti del partecipante a una ricerca, la trasparenza-spiegabilità e l'esigenza di poter valutare il razionale di una ricerca scientifica e, ancora, la trasparenza-tracciabilità e il requisito di riproducibilità degli studi. Si tratta di doveri il cui rispetto è imposto dal complesso regolatorio che, a livello nazionale, sovranazionale e internazionale, garantisce il concetto di integrità di una ricerca e la tutela dei diritti delle persone coinvolte. L'AI Act, per come è formulato, non fornisce molti appigli normativi vincolati, ma contribuisce, come si è visto, a mettere in luce nuove declinazioni del concetto di trasparenza, delle quali chi opera nel mondo della ricerca, ampiamente inteso, dovrà tenere conto.

Se la trasparenza è un valore che nell'ambito della ricerca merita di essere preservato e se si vuole che l'IA sia effettivamente foriera di benefici in tale campo, è auspicabile, da un lato, che i riferimenti alle descritte declinazioni della trasparenza, come risposta alle opacità della tecnologia, trovino quanto prima traduzione all'interno dei menzionati codici di condotta e che, dall'altro, le regole dell'IA trovino la propria collocazione all'interno del più ampio ecosistema di regole etico-giuridiche che presidiano il campo della ricerca (si pensi, solo per fare alcuni esempi, alle norme del Regolamento sui trials clinici⁴⁶, del Regolamento sui dispositivi medici⁴⁷ o, più latamente, anche del GDPR⁴⁸).

Solo così sarà possibile definire con certezza le modalità di attuazione del principio di trasparenza nel contesto in esame, garantendo la qualità di una ricerca, sull'IA e con l'IA, che possa svolgersi all'interno di un virtuoso circuito di fiducia, risponda alla sua finalità di utilità sociale e risulti conforme ai canoni di integrità, nel rispetto dei diritti delle persone coinvolte.

⁴⁶ Regolamento (UE) n. 536/2014 del Parlamento europeo e del Consiglio, del 16 aprile 2014 sulla sperimentazione clinica di medicinali per uso umano e che abroga la direttiva 2001/20/CE.

⁴⁷ Regolamento (UE) 2017/745 del Parlamento europeo e del Consiglio, del 5 aprile 2017, relativo ai dispositivi medici, che modifica la direttiva 2001/83/CE, il Regolamento (CE) n. 178/2002 e il Regolamento (CE) n. 1223/2009 e che abroga le direttive 90/385/CEE e 93/42/CEE del Consiglio.

⁴⁸ Regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio, del 27 aprile 2016, relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati e che abroga la direttiva 95/46/CE (Regolamento generale sulla protezione dei dati) (Testo rilevante ai fini del SEE).

Norme sociali e spiegazione del comportamento collettivo di umani e macchine

Giulia Andrighetto, Luca Tummolini

Istituto di Scienze e Tecnologie della Cognizione, CNR

INTRODUZIONE

Le grandi sfide globali contemporanee, come il cambiamento climatico e le emergenze sanitarie, pongono problemi di cooperazione su larga scala. Al di là delle innovazioni tecnologiche, la possibilità di ridurre le emissioni di CO₂ o tenere sotto controllo la diffusione di virus richiede necessariamente che gli individui cambino il loro comportamento per la realizzazione di un progetto comune di mutuo beneficio. Allo stesso tempo, situazioni di cooperazione come queste sono purtroppo caratterizzate da una vulnerabilità intrinseca in quanto l'interesse personale e quello collettivo sono spesso in conflitto, rendendo a livello individuale vantaggioso non partecipare e beneficiare della cooperazione altrui. Soluzioni a queste sfide rendono dunque urgente comprendere come motivare gli individui a partecipare a un'azione collettiva anche quando questo va contro il loro interesse personale.

Recentemente si sta lavorando allo sviluppo di una nuova Intelligenza Artificiale (IA) in grado esibire una vera e propria 'intelligenza cooperativa' che sappia interagire con gli umani, le istituzioni e le organizzazioni e che possa quindi ricoprire in futuro un ruolo nel facilitare la risoluzione di problemi di cooperazione su larga scala come questi. Organizzazioni come la *Climate Change Artificial Intelligence (CCAI)*¹ e la *Cooperative Artificial*

¹ Lynn H. Kaack et al., "Aligning artificial intelligence with climate change mitigation", *Nature Climate Change*, 12, 6 (2022): 518–527. DOI: 10.1038/s41558-022-01377-7; David Rolnick et al., "Tackling climate change with machine learning", *ACM Computing Surveys*, 55, 2 (2022): 1-96. DOI: 10.1145/3485128.

Intelligence Foundation (CAIF)², a cui partecipano sia scienziati che esperti provenienti dal mondo dell'industria, stanno promuovendo iniziative in cui l'IA attraverso nuovi strumenti, quali *Large Language Model* (LLM), interfacce uomo-macchina, sistemi reputazionali e algoritmi che facilitano la presa di decisione collettiva, è utilizzata al fine di facilitare la cooperazione e rispondere ai grandi problemi globali che le società contemporanee stanno affrontando.

Con il progresso dell'IA cooperativa, gli agenti artificiali autonomi saranno in grado di agire per conto di altri essere umani oppure come facilitatori e mediatori dell'interazione tra umani diventando quindi parte integrante di gruppi chiamati a prendere decisioni collettive e favoriranno la formazione di vere e proprie 'collettività ibride' tra umani e macchine. Esempi già familiari di queste collettività ibride sono quello di *Wikipedia* in cui umani e *bot* collaborano per editare i contenuti delle varie voci dell'enciclopedia e la piattaforma di social news e intrattenimento *Reddit* dove umani e *bot* cooperano per moderare i contenuti che vengono postati.

In un prossimo futuro modelli analoghi a questi potranno supportare anche processi più fondamentali alla base delle nostre società moderne, come ad esempio la deliberazione democratica. Uno studio recente infatti ha sviluppato un sistema di IA basata sugli LLM chiamato *Habermas Machine* in grado di sintetizzare discussioni umane su complesse questioni politiche e sociali in una dichiarazione condivisa e che possa essere usata per ricomporre visioni tra loro contrastanti. In una serie di esperimenti con più di 5000 partecipanti gli autori hanno mostrato come la *Habermas Machine* sia capace di facilitare la formazione di opinioni condivise su temi di interesse collettivo in modo più efficace e su scala più ampia di ciò che potrebbero fare esperti mediatori umani³.

Tuttavia, al di là di sistemi di IA finalizzati al supporto esplicito della cooperazione tra esseri umani è cruciale anche riconoscere che gran parte dell'interazione umana è influenzata da fattori più impliciti e informali che emergono spontaneamente e diventano parte della conoscenza tacita condivisa socialmente. Infatti, al di là di interventi dall'alto come

² Allan Dafoe et al., "Cooperative AI: machines must learn to find common ground", *Nature*, 593, 7857 (2021): 33-36. DOI: 10.1038/d41586-021-01170-0.

³ Michael H. Tessler et al., "AI can help humans find common ground in democratic deliberation", *Science*, 386, 6719 (2024): eadq2852. DOI: 10.1126/science.adq2852.

l'introduzione di intermediari esterni, la cooperazione umana è spesso influenzata da soluzioni che emergono dal basso come ad esempio le norme sociali. Le norme sociali sono le regole di comportamento tacite alla base della maggior parte dei nostri comportamenti quotidiani. Più specificamente sono regole interdipendenti e comportamenti collettivi che derivano dalle reciproche aspettative degli individui⁴.

Sono regole tacite – diverse da leggi o istituzioni formali – che definiscono molti aspetti della nostra vita sociale dal modo in cui ci salutiamo quando ci si incontra, con un saluto o una stretta di mano, al fatto che aspettiamo il nostro turno in fila per salire sull'autobus, fino al modo in cui dividiamo il carico di lavoro all'interno delle famiglie o interagiamo sui social media. Al di là di comportamenti quotidiani come questi, le norme sociali sono anche state individuate come possibili soluzioni nel caso delle sfide globali più complesse, dalla crisi climatica alle emergenze sanitarie delle pandemie virali⁵.

Quindi, se vogliamo capire come l'IA possa aiutarci ad affrontare le grandi sfide delle società contemporanee, è necessario comprendere anche come anche le macchine possano imparare ad interagire con noi umani condividendo queste regole tacite e interrogarsi sul tipo di comportamento che emergerà in tali collettività ibride. In particolare, vista la loro centralità nello 'spiegare' il comportamento collettivo esibito dagli esseri umani, diventa importante chiedersi infine in che modo l'interazione tra IA e umani possa a sua volta influenzare le norme sociali cruciali per il funzionamento delle nostre società.

⁴ Cristina Bicchieri, *The grammar of society: The nature and dynamics of social norms*, Cambridge: Cambridge University Press, 2006; Elinor Ostrom, "Collective Action and the Evolution of Social Norms", *Journal of Economic Perspectives*, 14, 3 (2000): 137-58. DOI: 10.1257/jep.14.3.137; Rosaria Conte e Cristiano Castelfranchi, "The Mental Path of Norms", *Ratio Juris*, 19 (2006). DOI: 10.1111/j.1467-9337.2006.00342.x; Christine Horne e Stefanie Mollborn, "Norms: An integrated framework", *Annual Review of Sociology*, 46 (2020): 467-487. DOI: 10.1146/annurev-soc-121919-054658; Aron Szekely et al., "Evidence from a long-term experiment that collective risks change social norms and promote cooperation", *Nature Communications*, 12, 5452 (2021). DOI: 10.1038/s41467-021-25734-w; Jon Elster, *The cement of society: A study of social order*, Cambridge: Cambridge University Press, 1989.

⁵ Karine Nyborg et al., "Social norms as solutions", *Science*, 354, 6308 (2016): 42-43. DOI: 10.1126/science.aaf8317.

NORME SOCIALI NELLE COLLETTIVITÀ IBRIDE

Le norme sociali sono un elemento essenziale delle società umane. Aiutano a prevedere il comportamento altrui e facilitano il coordinamento tra gli individui. Si sono, ad esempio, rivelate particolarmente efficaci nel coordinare le azioni collettive delle società in periodi di crisi, come quelli vissuti durante la pandemia di Covid-19, dove hanno permesso la diffusione di regole di comportamento preventivo, come mantenere una distanza minima dagli altri individui per rallentare la diffusione del virus, vaccinarsi, lavarsi le mani, indossare le mascherine in ambienti chiusi in presenza di altre persone⁶.

Le norme sociali hanno anche un importante effetto sulla mitigazione e adattamento al cambiamento climatico in quanto regolano una serie di comportamenti rilevanti per l'ambiente, come la conservazione dell'acqua e dell'energia⁷, il riciclo dei rifiuti⁸, la scelta dei mezzi di trasporto e delle diete alimentari⁹, e la gestione delle risorse collettive¹⁰.

⁶ Giulia Andrighetto et al., “Changes in social norms during the early stages of the COVID-19 pandemic across 43 countries”, *Nature Communications*, 15, 1436 (2024). DOI: 10.1038/s41467-024-44999-5; Andreas Diekmann, “Emergence of and compliance with new social norms”, *Zeitschrift Für Soziologie*, 49, 4 (2020): 236-248. DOI: 10.1515/zfsoz-2020-0021; Alex Moehring et al., “Providing normative information increases intentions to accept a COVID-19 vaccine”, *Nature Communications*, 14, 126 (2023). DOI: 10.1038/s41467-022-35052-4; Eva Vriens et al., “Vaccine-hesitant people misperceive the social norm of vaccination”, *PNAS Nexus*, 2, 5 (2023): 132. DOI: 10.1093/pnasnexus/pgad132; Eva Vriens et al., “Risk, sanctions and norm change: the formation and decay of social distancing norms”, *Phil. Trans. R. Soc. B*, 379, 1897 (2024): 20230035. DOI: 10.1098/rstb.2023.0035.

⁷ Hunt Allcott, “Social norms and energy conservation”, *Journal of Public Economics*, 95, 9–10 (2011): 1082-1095. DOI: 10.1016/j.jpubeco.2011.03.003; Daniel A. Brent et al., “Social comparisons, household water use and participation in utility conservation programs: Evidence from three randomized trials”, *Journal of the Association of Environmental and Resource Economists*, 2, 4 (2015): 597-627. DOI: 10.1086/683427.

⁸ Kjell A. Brekke et al., “Social Interaction in Responsibility Ascription: The Case of Household Recycling”, *Land Economics*, 86, 4 (2010): 766-84. DOI: 10.3368/le.86.4.766.

⁹ Gregg Sparkman et al., “How social norms are often a barrier to addressing climate change but can be part of the solution”, *Behavioural Public Policy*, 5, 4 (2020): 528-555. DOI: 10.1017/bpp.2020.42; Gregg Sparkman, “Dynamic norm interventions: How to enable the spread of positive change”, in *Handbook of wise interventions: How social psychology can help people change*, a cura di Gregory M. Walton e Alia J. Crum, The Guilford Press, 2021, pp. 429-447.

¹⁰ Juan-Camilo Cárdenas e Elinor Ostrom, “What do people bring into the game? Experiments in the field about cooperation in the commons”, *Agricultural Systems*, 82, 3 (2004): 307-326. DOI: 10.1016/j.agsy.2004.07.008.

Nei gruppi umani le norme sociali rappresentano un aspetto fondamentale dell'interazione umana. Sono state a lungo usate come spiegazione di fenomeni collettivi in varie discipline dalla sociologia¹¹, filosofia¹², psicologia sociale, morale e culturale¹³, economia e scienze politiche¹⁴, antropologia¹⁵, fino alla biologia evoluzionistica¹⁶, le

¹¹ Christine Horne e Stefanie Mollborn, "Norms: An integrated framework", *op. cit.*; Aron Szekely et al., "Evidence from a long-term experiment that collective risks change social norms and promote cooperation", *op. cit.*; Wojtek Przepiorka et al., "How Norms Emerge from Conventions (and Change)", *Socius*, 8 (2022): 23780231221124556. DOI: 10.1177/23780231221124556.

¹² Cristina Bicchieri, *The grammar of society: The nature and dynamics of social norms*, *op. cit.*; Jon Elster, *The cement of society: A study of social order*, *op. cit.*; Edna Ullmann-Margalit, *The Emergence of Norms*, Oxford University Press, 1977; Luca Tummolini et al., "A convention or (tacit) agreement betwixt us: on reliance and its normative consequences", *Synthese*, 190, 4 (2013): 585-618. DOI: 10.1007/s11229-012-0194-8; Francesco Guala e Luigi Mittone, "How history and convention create norms: An experimental study", *Journal of Economic Psychology*, 31 (2010): 749-756. DOI: 10.1016/j.joep.2010.05.009.

¹³ Robert B. Cialdini, "A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places", *Journal of Personality and Social Psychology*, 58, 6 (1990): 1015–26. DOI: 10.1037//0022-3514.58.6.1015; Cecilia Heyes, "Rethinking Norm Psychology", *Perspectives on Psychological Science*, 19, 1 (2024): 12-38. DOI: 10.1177/17456916221112075; Maria K. Lapinski e Rajiv N. Rimal, "An explication of social norms", *Communication Theory*, 15, 2 (2005): 127-147. DOI:10.1093/ct/15.2.127; Siegwart M. Lindenberg, "Social norms: What happens when they become more abstract?", in *Rational Choice*, a cura di Andreas Diekmann et al., Geburtstag VS Verlag für Sozialwissenschaften, 2008, pp. 63–81. DOI: 10.1007/978-3-531-90866-3_5; Michele J. Gelfand et al., "Differences between tight and loose cultures: A 33-nation study", *Science*, 332, 6033 (2011): 1100–1104. DOI: 10.1126/science.1197754.

¹⁴ Ken Binmore, "Social norms or social preferences?", *Mind & Society*, 9, 2 (2010): 139-157. DOI: 10.1007/s11299-010-0073-2; Ernst Fehr e Ivo Schurtenberger, "Normative foundations of human cooperation", *Nature human behaviour*, 2, 7 (2018): 458-468. DOI: 10.1038/s41562-018-0385-5; Herbert Gintis, "Social norms as choreography", *Politics, Philosophy & Economics*, 9, 3 (2010): 251-264. DOI: 10.1177/1470594X09345474; Karine Nyborg et al., "Social norms as solutions", *op. cit.*; Paola Giuliano e Nathan Nunn, "Understanding Cultural Persistence and Change", *The Review of Economic Studies*, 88, 4 (2021): 1541–81. DOI: 10.1093/restud/rdaa074; Elinor Ostrom, "Collective Action and the Evolution of Social Norms", *op. cit.*

¹⁵ Jean Ensminger e Joseph Henrich, a cura di, *Experimenting with social norms: Fairness and punishment in cross-cultural perspective*, Russell Sage Foundation, 2014; Robert B. Edgerton, *Sick Societies: Challenging the myth of primitive harmony*, University of Michigan: Free Press, 1992.

¹⁶ Sergey Gavrilets, e Peter J. Richerson, "Collective action and the evolution of social norm internalization", *Proceedings of the National Academy of Sciences*, 114, 23 (2017): 6068-73. DOI: 10.1073/pnas.1703857114; Peter J. Richerson e Richard Boyd, *Not By Genes Alone: How Culture Transformed Human Evolution*, University of Chicago Press: Chicago, 2005; Denis Tverskoi et al., "Disentangling material, social, and cognitive determinants of human behavior and beliefs", *Humanities and Social Sciences Communications*, 10, 1 (2023): 236. DOI: 10.1057/s41599-023-01745-4.

scienze sociali computazionali¹⁷ e la robotica¹⁸, oltre che in recenti studi interdisciplinari¹⁹ sulle norme sociali e il loro cambiamento.

Mentre le società umane hanno avuto migliaia di anni per sviluppare e consolidare norme sociali in vari contesti – tra cui norme fondamentali per la sopravvivenza come quelle che promuovono la cooperazione – *cosa accade alle norme nelle collettività ibride? È possibile che agenti artificiali autonomi acquisiscano una comprensione delle norme sociali comparabile a quella umana e ne siano influenzati in modo analogo? Quali norme emergono per regolare le interazioni in tali contesti? Cosa succede alle norme esistenti? Si adattano per poter favorire le interazioni in questi nuovi ambienti?* In quanto segue, non esamineremo quali siano le norme più desiderabili in questi contesti ibridi, ma discuteremo la possibilità che esseri umani e agenti artificiali possano condividere le stesse norme sociali e sull’impatto che l’interazione sempre più frequente tra umani e macchine può avere sulle dinamiche delle norme sociali e in particolare sulla loro emergenza, diffusione e cambiamento. Nello specifico, esamineremo quali siano le capacità che le macchine devono avere per poter riconoscere le norme sociali e tenerle in considerazione nelle loro interazioni con gli altri, quale effetto possa avere l’interazione con le macchine sulla formazione di aspettative negli umani, e infine quale possa essere l’impatto di queste collettività ibride sulla formazione e cambiamento delle norme sociali.

QUANDO LE MACCHINE IMPARANO LE NORME SOCIALI UMANE

Per capire fino in fondo il ruolo che le norme sociali possono avere in queste collettività ibride è importante chiedersi se e come le macchine possano

¹⁷ Damon Centola et al., “Experimental evidence for tipping points in social convention”, *Science*, 360 (2018): 1116–19. DOI: 10.1126/science.360.6393.1082-d; Rosaria Conte et al., *Minding norms: Mechanisms and dynamics of social order in agent societies*, Oxford University Press, 2014.

¹⁸ Bertram F. Malle e Matthias Scheutz, “Learning how to behave: Moral competence for social robots”, in *Handbook of machine ethics*, a cura di Oliver Bendel, Springer-Verlag, 2019.

¹⁹ Michele J. Gelfand et al., “Norm Dynamics: Interdisciplinary Perspectives on Social Norm Emergence, Persistence, and Change”, *Annual Review of Psychology*, 75 (2024): 341-378. DOI: 10.1146/annurev-psych-033020-013319; Giulia Andrighetto et al., “Social norm change: drivers and consequences”, *Philosophical Transactions of the Royal Society B*, 379, 1897 (2024): 20230023. DOI: 10.1098/rstb.2023.0023; Loukas Balafoutas et al., “Social Norms: Enforcement, Breakdown & Polarization”, *European Economic Review*, 170 (2024): 104885. DOI: 104885.10.1016/j.eurocorev.2024.104885.

acquisire una ‘intelligenza normativa’ sufficiente a riconoscere, ragionare e decidere sulla base delle norme e aspettative sociali date per scontate dagli umani e che guidano molti dei loro comportamenti.

La possibilità che le norme possano essere uno strumento per favorire anche la coordinazione e la cooperazione tra agenti artificiali autonomi è stata a lungo esplorata nell’ambito dei sistemi multi-agente, settore dell’IA che per primo ha integrato il paradigma degli agenti autonomi con la svolta interattiva e sociale nello studio dell’IA²⁰. In questa importante variante dell’IA classica, lo sviluppo dell’intelligenza normativa si è perlopiù concentrato sull’integrazione nell’architettura computazionale di un modulo di ragionamento simbolico sulle norme intese come oggetti esterni agli agenti, ma volti a influenzarne le decisioni²¹. Lo studio formale con la logica e la teoria dei giochi privilegiato in questa tradizione²² è stato fondamentale per sviluppare un approccio sistematico e rigoroso ad un dominio dell’interazione che è informale per definizione²³. Tuttavia, l’approccio classico ha dovuto sacrificare la comprensione della natura contestuale e dinamica delle norme sociali e la possibilità della loro acquisizione sulla base di processi di apprendimento basati sull’esperienza diretta o osservazionale e in assenza dell’intervento esterno del programmatore.

La realizzazione di collettività ibride in cui intelligenze naturali e artificiali cooperano alla pari rende la soluzione a questi problemi oggi

²⁰ Nick R. Jennings, “Commitments and conventions: the foundation of coordination in multi-agent systems”, *Knowledge Engineering Review*, 8, 3 (1993): 223-250. DOI: 10.1017/S0269888900000205; Yoav Shoham e Moshe Tennenholtz, “On social laws for artificial agent societies: off-line design”, *Artificial Intelligence*, 73, 1-2 (1995): 231–252. DOI: 10.1016/0004-3702(94)00007-N; Rosaria Conte e Cristiano Castelfranchi, *Cognitive and social action*, London, UCL Press, 1995.

²¹ Guido Boella e Leendert van der Torre, “An architecture of a normative system: Counts-as conditionals, obligations, and permissions”, *Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multi Agent Systems (AAMAS)*, ACM, New York, 2006, pp. 229–231. DOI: 10.1145/1160633.1160671; Rosaria Conte et al., *Minding norms: Mechanisms and dynamics of social order in agent societies*, *op. cit.*; Bertram F. Malle et al., “Requirements for an Artificial Agent with Norm Competence”, *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 21-27. DOI: 10.1145/3306618.3314252.

²² Davide Grossi et al., “Norms in game theory”, in *Agreement Technologies. Law, Governance and Technology Series*, a cura di Sascha Ossowski, vol 8. Springer, Dordrecht, 2013, pp. 191-197. DOI: 10.1007/978-94-007-5583-3_12.

²³ Cristiano Castelfranchi, “Formalising the informal? Dynamic social order, bottom-up social control, and spontaneous normative relations”, *Journal of Applied Logic*, 1 (2003): 47-92. DOI: 10.1016/S1570-8683(03)00004-1.

ineludibile. Fortunatamente, la maturazione del *Deep Learning* nelle reti neurali artificiali, l'apparizione di “ragionatori” neurali come gli LLM²⁴ e in particolare la loro integrazione con il paradigma classico degli agenti autonomi ha consentito di superare alcuni limiti del passato e ha portato allo sviluppo di una nuova generazione di sistemi multi-agente in grado di apprendere in modo autonomo, coordinarsi e collaborare con altri agenti, condividere il nostro ambiente e comunicare linguisticamente per – giorno finalmente – agire insieme a noi²⁵.

L'interesse per la nuova stagione di sistemi multi-agente basati su LLM è cresciuto negli ultimi anni sempre di più²⁶ e oggi rappresenta una delle innovazioni più promettenti per la creazione dell'intelligenza cooperativa in grado di capire le norme sociali in modo simile agli esseri umani. In particolare, uno studio recente ha cercato di valutare il livello di comprensione che LLM come GPT-4 hanno delle norme sociali umane sulla base delle competenze linguistiche avanzate che li caratterizzano e i risultati hanno dimostrato un livello di competenza comparabile a quello umano²⁷. Focalizzandosi sul contenuto delle norme in diversi ambiti (dalle regole condivise nel linguaggio, nell'economia, nella cultura, ecc.) e sulla capacità di generare una risposta linguistica considerata appropriata rispetto agli standard umani, lo studio dimostra una sofisticata capacità degli LLM nel comprendere il contenuto anche implicito di molte norme sociali quotidiane.

Altri studi hanno, inoltre, dimostrato che tale conoscenza normativa implicita nelle competenze linguistiche acquisite da un LLM è anche potenzialmente in grado di influenzare la generazione di comportamenti non linguistici. Ad esempio, utilizzando una batteria di giochi economici tra cui il gioco del dittatore, il gioco della fiducia e il gioco dei beni pubblici è stato

²⁴ Josh Achiam et al., “Gpt-4 technical report”, (2023). DOI: 10.48550/arXiv.2303.08774.

²⁵ Anna M. Borghi et al., “Language as a cognitive and social tool at the time of large language models”, *Journal of Cultural Cognitive Science*, 8, 3 (2024): 179-198. DOI: 10.1007/s41809-024-00152-8; per una discussione precoce di questo scenario si veda: Alessandro Ricci et al., “The mirror world: Preparing for mixed-reality living”, *IEEE Pervasive Computing*, 14, 2 (2015): 60-63. DOI: 10.1109/MPRV.2015.44.

²⁶ Lei Wang et al., “A survey on large language model based autonomous agents”, *Frontiers of Computer Science*, 18, 6 (2024): 186345. DOI: 10.1007/s11704-024-40231-1.

²⁷ Ye Yuan et al., “Measuring Social Norms of Large Language Models”, (2024). DOI: 10.48550/arXiv.2404.02491.

dimostrato che GPT4 compie scelte molto simili, e spesso indistinguibili, da quelle compiute da un campione di partecipanti umani²⁸. Considerando che il comportamento umano in giochi economici di questo tipo è tipicamente influenzato da norme sociali legate all'equità o alla reciprocità²⁹, questi risultati suggeriscono ulteriormente che la comprensione implicita di tali norme che questi sistemi sviluppano è in grado di avere anche rilevanza comportamentale.

Inoltre, la comprensione normativa acquisita dagli LLM sembra influenzare il comportamento non solo in situazioni astratte catturate con i giochi economici. Un altro studio recente nel dominio della robotica autonoma umanoide ha dimostrato, ad esempio, che i movimenti fisici di un robot umanoide integrato con un LLM (GPT4) possono essere controllati efficacemente a partire da istruzioni verbali in linguaggio naturale³⁰. In particolare, lo studio ha dimostrato che il robot è in grado di generare movimenti complessi a partire da *prompt* che descrivono azioni quotidiane (come l'azione di bere il tè) e di generare in modo spontaneo azioni nuove che risultano intuitivamente appropriate al contesto sociale e percepite come naturali alla valutazione fatta da esseri umani. Risultati come questi suggeriscono, quindi, che oltre al contenuto semantico astratto delle norme sociali, sistemi di IA come questi sono in grado di estrarre dalle regolarità linguistiche anche le informazioni sulle condotte condivise prescritte dalle norme sociali al punto da poter influenzare sia la decisione che l'esecuzione dei movimenti fisici nell'ambiente reale.

Una caratteristica distintiva delle norme sociali umane è che esse sono oggetto di un livello di conoscenza tacita in comune tra i partecipanti che orienta la comprensione anche cognitiva di una situazione e influenza le decisioni da prendere in un contesto. Questo livello semantico delle norme oggi sembra essere alla portata delle macchine. In una interpretazione

²⁸ Qiaozhu Mei et al., "A Turing test of whether AI chatbots are behaviorally similar to humans", *Proceedings of the National Academy of Sciences*, 121, 9 (2024): e2313925121. DOI: 10.1073/pnas.2313925121.

²⁹ Aron Szekely et al., "Evidence from a long-term experiment that collective risks change social norms and promote cooperation", *op. cit.*; Wojtek Przepiorka et al., "How Norms Emerge from Conventions (and Change)", *op. cit.*

³⁰ Takahide Yoshida et al., "From text to motion: grounding gpt-4 in a humanoid robot "Alter3", *Frontiers in Robotics and AI*, 12 (2025): 1581110. DOI: 10.3389/frobt.2025.1581110.

persuasiva, infatti, gli LLM sono vere e proprie “tecnologie culturali”³¹ in grado di preservare schemi culturali come le norme sociali e favorirne la trasmissione.

Al di là del loro riconoscimento e della loro comprensione, tuttavia, le norme sociali hanno per gli esseri umani una fondamentale conseguenza motivazionale e di influenzamento. In aggiunta, molteplici studi hanno dimostrato che le norme sociali non influenzano il comportamento umano in modo categorico o incondizionato, ma che la pressione motivazionale che gli esseri umani sentono nel conformarsi al contenuto prescritto da una norma dipende in modo sottile dalle aspettative condivise in un certo contesto sociale e può mutare dinamicamente al mutare di queste³². Se la conoscenza normativa tacita è oggi potenzialmente condivisibile con le macchine, *che aspettative sono in grado di formarsi le macchine rispetto al comportamento umano in un contesto specifico? E come possono le macchine diventare esse stesse sensibili alle nostre aspettative nei loro confronti quando saremo membri alla pari nelle future collettività ibride?*

Per poter essere influenzate da norme sociali, le macchine avranno bisogno di essere in grado di inferire e attribuire aspettative agli umani (e alle altre macchine) e considerare tali aspettative nel momento in cui si trovano a prendere decisioni in contesti di interdipendenza strategica. Infine, come i bambini nelle società umane, affinché il loro comportamento diventi predicibile e affidabile, le macchine dovranno essere socializzate in modo da sviluppare la competenza normativa prevalente in una generazione³³.

Questo può avvenire attraverso sistemi ‘adulti’ che socializzano le macchine, le orientano, insegnano loro le norme sociali e si accertano che le abbiano comprese e siano in grado di obbedirle. Affinché ciò avvenga è

³¹ Eunice Yiu et al., “Transmission versus truth, imitation versus innovation: What children can do that large language and language-and-vision models cannot (yet)”, *Perspectives on Psychological Science*, 19, 5 (2024): 874-883. DOI: 10.1177/17456916231201401.

³² Cristina Bicchieri, *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*, Oxford University Press, 2016; Jordan E. Theriault et al., “The sense of should: A biologically-based framework for modeling social pressure”, *Physics of Life Reviews*, 36 (2021): 100-136. DOI: 10.1016/j.plrev.2020.01.004; John Michael e Luca Tummolini, “Intrinsically motivated norm compliance and the sense of obligation”, *Current Opinion in Psychology*, 65 (2025): 102043. DOI: 10.1016/j.copsyc.2025.102043.

³³ Don Ross et al., “Modeling norm-governed communities with conditional games: Sociological game-determination and economic equilibria”, *(Economia. History, Methodology, Philosophy)*, 14, 2 (2024): 349-398. DOI: 10.4000/120ij.

necessario che le macchine comprendano che il comportamento umano è influenzato da istituzioni, organizzazioni e network sociali all'interno dei quali viviamo e altri fattori extra-individuali che sono oggetto di studio centrale delle scienze sociali³⁴. Questo è uno scenario necessario, e non del tutto implausibile, se si vuole raggiungere una vera IA normativa.

QUANDO GLI ESSERI UMANI INTERAGISCONO CON LE MACCHINE NORMATIVE

Così come è importante chiedersi se le macchine sono in grado di percepire le nostre norme sociali e capire quali aspettative saranno in grado di formarsi nei nostri confronti in queste collettività ibride, è anche fondamentale comprendere in che modo potrà essere influenzata la percezione delle norme sociali da parte degli umani in presenza di macchine normative.

Studi recenti hanno esaminato l'effetto delle interazioni tra umani e macchine sulla cooperazione e su altri comportamenti collettivi³⁵. Anche se la competenza normativa effettiva degli agenti artificiali è ancora limitata, questi lavori si sono focalizzati su contesti controllati in cui i partecipanti umani interagiscono con le macchine in giochi di cooperazione come il dilemma del prigioniero, il gioco della fiducia o il gioco del bene pubblico dove le decisioni degli agenti può essere facilmente controllata dallo sperimentatore. Nonostante queste limitazioni, quello che emerge in maniera abbastanza consistente è che il livello di fiducia nei giochi della fiducia o di cooperazione nel dilemma del prigioniero o nei giochi del bene pubblico sono più bassi quando gli umani sanno che stanno interagendo in contesti dove sono presenti anche delle macchine, rispetto a quando sanno di giocare

³⁴ Christopher A Bail, "Can Generative AI improve social science?", *Proceedings of the National Academy of Sciences*, 121, 21 (2024): e2314021121. DOI: 10.1073/pnas.2314021121.

³⁵ Allan Dafoe et al., "Cooperative AI: machines must learn to find common ground", *op. cit.*; Hirokazu Shirado e Nicholas A. Christakis, "Locally noisy autonomous agents improve global human coordination in network experiments", *Nature*, 545 (2017): 370–374. DOI: 10.1038/nature22332; Fernando P. Santos et al., "Evolution of Collective Fairness in Hybrid Populations of Humans and Agents", *Proceedings of the AAAI Conference on Artificial Intelligence*, 33 (2019): 6146–6153. DOI: 10.1609/aaai.v33i01.33016146; Sofia Petisca et al., "Human Dishonesty in the Presence of a Robot: The Effects of Situation Awareness", *International Journal of Social Robotics*, 14, 5 (2022): 1211–1222. DOI: 10.1007/s12369-022-00864-3; Iyad Rahwan et al., "Machine behaviour", *Nature*, 568 (2019): 477–486. DOI: 10.1038/s41586-019-1138-y.

solo tra umani³⁶. In generale, gli umani tendono ad approfittarsi degli altri e a cooperare meno quando sono a conoscenza di interagire con le macchine.

L'evidenza riportata è principalmente limitata al comportamento. Pochi sono ancora gli studi che esaminano l'effetto dell'interazione tra umani e macchine sulla percezione e formazione di norme sociali. Ad esempio, è stato mostrato come le norme sociali siano più incerte quando gli umani si trovano a interagire con le macchine in un gioco della fiducia modificato³⁷. In particolare, ai soggetti umani viene chiesto di giocare insieme a *bots* a una serie di giochi che prevedono che un osservatore possa punire comportamenti che percepisce come violazioni (*third-party punishment*), seguiti da giochi della fiducia. I risultati mostrano che quando i partecipanti umani giocano insieme ai *bot* il livello di consenso che emerge intorno alle norme di condivisione e la punizione di chi non le rispetta è più basso rispetto a quando giocano solo tra umani. Tuttavia, le norme che si sviluppano in contesti ibridi seppure siano più deboli hanno comunque un effetto nell'influenzare il comportamento dei partecipanti umani. Lo studio dimostra, infatti, che quando interagiscono in collettività ibride, gli esseri umani fanno riferimento alle norme sociali che regolano i comportamenti nei collettivi umani, ma sono più incerti rispetto al consenso che c'è intorno a queste norme. Tuttavia, manipolando le norme al fine di renderle più forti il loro effetto sul comportamento aumenta. Uno studio recente³⁸ dimostra che in contesti online in cui gli umani osservano *bot* punire utenti che condividono contenuti falsi, il consenso intorno alle norme di punizione è più basso rispetto a quando sanno che è un umano a punire, ma è comunque più alto rispetto a quando non vedono nessuno (né umani, né *bot*) farlo. Anche se in maniera più debole rispetto a quelle umane, le azioni delle macchine trasmettono informazione normativa su cui gli esseri umani basano le loro aspettative. Infine, è stato evidenziato come gli umani quando giocano a carte con i robot siano influenzati dalle scelte

³⁶ Fatimah Ishowo-Oloko et al., “Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation”, *Nature Machine Intelligence*, 1 (2019): 517–521. DOI: 10.1038/s42256-019-0113-5; Jurgis Karpus et al., “Algorithm exploitation: Humans are keen to exploit benevolent AI”, *iScience*, 24 (2021): 102679. DOI: 10.1016/j.isci.2021.102679.

³⁷ Kinga Makovi et al., “Trust within human-machine collectives depends on the perceived consensus about cooperative norms”, *Nature Communications*, 14 (2023): 3108. DOI: 10.1038/s41467-023-38592-5.

³⁸ Carlo Ciucani et al., “AI and social corrections: shaping norms against misinformation sharing” (in preparazione).

di questi ultimi e quando gli umani si trovano nella condizione di essere una minoranza sentano comunque la pressione sociale di conformarsi a quello che fanno i robot³⁹.

Da questi studi emerge quindi che le norme sociali mantengono un ruolo importante nello spiegare il comportamento cooperativo anche in collettività ibride, ma che il loro impatto potrebbe essere più limitato rispetto a quelle che si sviluppano in gruppi formati solo da umani, in quanto le norme che si creano sono più deboli probabilmente a causa di una maggiore incertezza sulle aspettative condivise nella collettività ibrida. Interventi volti a favorire la formazione di consenso intorno a norme di cooperazione in collettività ibride⁴⁰ potrebbero dunque rivelarsi efficaci a promuovere comportamenti cooperativi necessari a mitigare o risolvere i problemi di azione collettiva che le nostre società contemporanee si trovano ad affrontare e che potrebbero essere esasperati in collettività ibride.

CAMBIAMENTO DELLE NORME SOCIALI NELLE COLLETTIVITÀ IBRIDE

Sistemi di IA cooperativa basati su LLM o tecnologie simili dovranno, quindi, imparare e adattarsi alle nostre norme sociali, ed è probabile che noi dovremo fare lo stesso con nuove norme che emergeranno quando interagiamo con loro. Vale la pena considerare un'ulteriore conseguenza del fatto che, a differenza delle tecnologie precedenti, l'IA sarà in grado non solo di estrarre in modo automatico conoscenza normativa dall'elaborazione del linguaggio umano ma al tempo stesso saprà usare il linguaggio in maniera efficace per influenzare il comportamento umano⁴¹. Questo nuovo tipo di IA parteciperà alla coevoluzione delle norme sociali che regolano le collettività ibride.

Ad esempio, se dotati delle competenze normative necessarie, i sistemi di IA potrebbero, servire da 'sentinelle' (*early-warning systems*) in grado di

³⁹ Nicole Salomons et al., "Humans Conform to Robots: Disambiguating Trust, Truth, and Conformity", *Proceedings of the Thirteenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, (2018): 187-195. DOI: 10.1145/3171221.3171282.

⁴⁰ Kinga Makovi et al., "Trust within human-machine collectives depends on the perceived consensus about cooperative norms", *op. cit.*

⁴¹ Francesco Salvi et al., "On the conversational persuasiveness of GPT-4", *Nature Human Behaviour* (2025). DOI: 10.1038/s41562-025-02194-6.

anticipare criticità nella cooperazione umana e nelle norme sociali che la sostengono. L'IA potrebbe aiutare a prevedere in quali condizioni le norme sociali diventano più fragili o inefficaci e intervenire per rafforzarle al fine di promuovere una cooperazione duratura. L'IA potrebbe intervenire direttamente, identificando criticità, proponendo soluzioni che poi mette in atto autonomamente, o indirettamente con un ruolo di supporto al decisore politico, riportando agli umani la diagnosi effettuata e le possibili soluzioni trovate, così che gli esseri umani possano poi scegliere di attuare quelle che ritengono più promettenti.

Sistemi di IA di questo tipo possono tuttavia svolgere un ruolo anche più diretto nel processo di cambiamento sociale. Ad esempio la teoria sociologica della massa critica ha esplorato le condizioni in cui una minoranza della popolazione costituita da individui impegnati è in grado di sovvertire una norma esistente quando raggiunge una dimensione critica e il sistema sociale attraversa un punto di cambiamento repentino⁴². Una volta raggiunto il punto critico, le azioni di un piccolo gruppo di individui sono in grado di innescare un rapido e inarrestabile cambiamento nel comportamento che fa sì che la visione o il comportamento della minoranza prima messo all'indice o guardato con diffidenza venga considerato appropriato a livello globale, diventando così la norma. Recentemente è stato suggerito che le macchine potrebbero sfruttare questo meccanismo sociale costituendo gruppi di agenti che contribuiscono a far iniziare cambiamenti di comportamenti, di opinioni e norme sociali⁴³.

Le norme sociali che regolano le collettività ibride potrebbero cambiare per effetto di azioni e opinioni dell'IA e non sempre nel nostro interesse. Le macchine potrebbero favorire il superamento di punti critici che portano alla diffusione di comportamenti che noi umani non vogliamo o potrebbero invece rallentare o bloccare il cambiamento di norme sociali, pregiudizi e discriminazioni che sono oggi in atto. Dal momento che macchine come ChatGPT sono addestrate su dati provenienti da *Wikipedia*, social media

⁴² Thomas Schelling, *Micromotives and Macrobehavior*, Norton, New York, 1978; Mark Granovetter, "Threshold models of collective behavior", *American Journal of Sociology*, 83, 14 (1978): 1420-1443. DOI: 10.1086/226707.

⁴³ Andrea Baronchelli, "Shaping new norms for AI", *Philosophical Transactions of the Royal Society B*, 379, 1897 (2024): 20230028. DOI: 10.1098/rstb.2023.0028

come *Twitter* e *Reddit*, *ebook*, saggi, e altro materiale presente su internet, nelle loro risposte riflettono pregiudizi, stereotipi e norme sociali che potrebbero non essere pienamente rappresentative delle opinioni e visioni correnti⁴⁴. Le macchine con le risposte che generano potrebbero favorire la falsa percezione che comportamenti discriminatori o razzisti siano più tollerati di quanto non lo siano nella realtà e inibire le reazioni di coloro che ritengono ingiusti e non socialmente appropriati tali comportamenti e opinioni. Le macchine con il loro comportamento potrebbero favorire spirali del silenzio che rallentano o perfino bloccano il cambiamento in atto di norme sociali⁴⁵.

CONCLUSIONI

Al fine di poter interagire con gli esseri umani – in modo diretto o indiretto – i sistemi di IA dovranno essere in grado di capire e conformarsi alle norme sociali che guidano le interazioni nelle collettività ibride, dovranno monitorare il loro cambiamento nel tempo, e dovranno valutare la loro efficacia nel sostenere comportamenti cooperativi. Auspicabilmente, dovranno poter intervenire per favorire e non sovvertire la cooperazione umana. In questo lavoro abbiamo discusso alcune evidenze che sembrano suggerire l'avvento oramai prossimo di un'IA normativa e abbiamo discusso alcuni possibili conseguenze per le società umane. Diventa quindi di cruciale importanza capire quali norme emergeranno in queste collettività ibride. È difficile oggi prevederlo, perché i nuovi scenari ibridi in cui ci troveremo a vivere non sono completamente immaginabili e gli strumenti di indagine che attualmente utilizziamo per studiare e misurare le norme e il loro cambiamento potrebbero non essere adeguati in questi nuovi scenari.

Per superare questi limiti servirà quindi lo sviluppo di una nuova scienza sociale delle collettività ibride⁴⁶, un'alleanza tra scienziati cognitivi,

⁴⁴ Louis Lippens, “Computer says ‘no’: Exploring systemic bias in ChatGPT using an audit approach”, *Computers in Human Behavior: Artificial Humans*, 2 (2024): 100054. DOI:10.1016/j.chbah.2024.100054.

⁴⁵ Piergiorgio Castioni et al., “The voice of few, the opinions of many: evidence of social biases in Twitter COVID-19 fake news sharing”, *Royal Society Open Science*, 9 (2022): 220716. DOI: 10.1098/rsos.220716.

⁴⁶ Iyad Rahwan et al., “Machine behaviour”, *op. cit.*

scienziati sociali e scienziati computazionali che lavorano nell'ambito dell'IA volta a sviluppare nuovi metodi interdisciplinari per capire il comportamento di queste collettività ibride tra esseri umani e agenti autonomi artificiali. La creazione di piattaforme in cui agenti cognitivi complessi che apprendono sulla base di *Deep Reinforcement Learning* e che interagiscono con umani in un ambiente complesso¹ è un primo passo in questa direzione. Ma molto lavoro resta ancora da fare.

¹ Uri Hertz et al., "Beyond the matrix: Experimental approaches to studying cognitive agents in social-ecological systems", *Cognition*, 254 (2025): 105993. DOI: 10.1016/j.cognition.2024.105993.

Postfazione

Gilberto Corbellini

Sapienza Università di Roma

Le riflessioni personali, a diversi livelli di competenze e in buona parte per sentito dire o basate esperienze sempre circoscritte (come questa che non so se qualcuno mai leggerà), su quanto e come i modelli linguistici IA possono realizzare risultati che attribuiamo al pensiero umano e alle sue capacità autoriflessive, e come stabilire se una macchina, attraverso il suo particolare modo di pensare, può realizzare tutto quello che viene prodotto come comportamento dal cervello della specie, sono una percentuale significativa della letteratura generica sull'IA. Aggiungere acqua da pestare nel mortaio non è il massimo compiacimento.

Qualcuno dice che abbiamo a che fare con macchine che ha poco senso definire 'intelligenti'. Sono sapienti o impensabilmente erudite: capaci di un livello avanzato di processamento di dati linguistici e in grado di estrarre informazioni utili per l'uomo su molti fronti delle attività pratiche. La fumosità delle sofisticatezze semantiche si può diradare solo entrando un po' nei dettagli. Per esempio, lasciamo pure che il concetto generico di 'intelligenza' si spanda a denotare un po' ovunque si osservino fenomenologie di soluzioni dei problemi, ma rimane una bella distanza evolutiva e funzionale tra 'intelligenza' e 'razionalità': sono fenomenologie che reclutano epistemologie, psicologie e neurologie. Ci tornerò.

Si coglie in giro un'oscillazione tra sicurezza che certi risultati che sono ancora delle sfide, come l'Intelligenza Generale, saranno presto o tardi accessibili alle macchine pensanti, e il fatto che in ultima istanza noi non sappiamo davvero di cosa parliamo, se non su un piano operativo, in termini di 'spiegabilità', quando discutiamo delle conoscenze queste macchine. Le cose non sono molto diverse per la conoscenza umana prodotta da strutture

cerebrali che hanno una storia evolutiva e sono più o meno adattative sempre in rapporto a specifici contesti ambientali. Nel 1951 Alan Turing scriveva, che “sembra probabile che una volta iniziato il metodo delle macchine pensanti, non ci vorrebbe molto per superare le nostre deboli capacità”. Turing non era sempre lineare nei suoi flussi mentali. Tre anni prima in un report aveva scritto di “intelligenza come concetto emotivo”, affermando che “la misura in cui consideriamo che qualcosa si comporti in modo intelligente è determinata tanto dal nostro stato mentale e dalla nostra formazione, quanto dalle proprietà dell’oggetto in esame”.

Sono anche, o forse soprattutto, i difetti che rendono ChatGPT e simili interessanti e inducono a non fidarsi di quello che viene restituito come prima risposta, apparentemente documentata, a una specifica domanda. Meglio scavare. Per esempio, se chiedo alla macchina se è vero che secondo diversi studi i pazienti possono preferire di interagire con sistemi di intelligenza artificiale, piuttosto che con dei medici umani, la risposta riassume efficacemente, per punti, alcune tesi che si trovano in letteratura. Se poi chiedo di dirmi le fonti bibliografiche, in parte se le inventa. Con le fonti bibliografie ha ancora qualche problema. Imbrogliava parecchio la versione 3. Come del resto fanno alcuni umani quando scrivono le bibliografie dei loro lavori. Molto meno la versione 4o. Se formuliamo a ChatGPT la stessa identica domanda in modi solo lessicalmente diversi, per esempio cambiando anche solo i gradi di ambiguità, talvolta restituisce risposte diverse. Ha una conoscenza impressionante, unica al mondo in questo momento, ma non capisce quello che sa e che sta dicendo (e onestamente lo dice!). Lo sappiamo tutti. Ma quando ci infuochiamo nell’usarla entrano in gioco diversi *bias* cognitivi probabilmente inconsci, che ci portano a credere che possa accedere davvero a tutto lo scibile. Invece, spesso è solo quella parte troppo facilmente accessibile e che vale un po’ meno... La salva solo la frequenza statistica che sa calcolare mirabilmente.

Il potenziale di senso delle risposte deriva dalla formidabile capacità di selezionare la conoscenza, a livelli talmente estesi da essere inimmaginabili per noi umani: lavorando con centinaia di livelli di astrazione avendo accesso a oltre un trilione di dati di testo. Ma non ci sono colleganti rientranti (bidirezionalmente) tra i diversi livelli di categorizzazione, che mi risulti, che invece sono la caratteristica peculiare del cervello umano: quella che consente l’astrazione, la creatività e verosimilmente lo sviluppo di un sé e di una

coscienza, dato un livello determinato di informazione integrata. Non c'è apparentemente limite alla scalabilità cumulativa del sistema algoritmico, se non forse i costi energetici. Ma come interlocutore linguistico, Chat rimane chiuso nella sfera del nozionismo, quantomeno per le aree meno tecniche della comunicazione. Ma questo era forse prevedibile.

E per quanto riguarda intelligenza e razionalità? Cosa pensa o sa di sé? Se gli chiedo che cosa è il pensiero critico, che vorremmo fosse da tutti più praticato e che riteniamo un tratto della razionalità, ChatGPT mi restituisce una definizione un po' generica, ma corretta. Mi informa che non ne fa uso: “sono un grande modello linguistico addestrato a fornire informazioni e assistenza al meglio delle mie conoscenze e capacità. Pur essendo in grado di elaborare informazioni e generare risposte in base ai dati su cui sono stato addestrato, non possiedo la coscienza o le capacità di pensiero critico di un essere umano.” Lo stesso vale per la ‘teoria della mente’, che la macchina dimostra di conoscere solo come nozione o gli esperimenti sulle false credenze. Ci informa anche che non usa il Sistema 1 di Kahneman, ovviamente, e solo in parte il 2... Niente emozioni da usare o tenere a bada: quindi niente morale o etica in proprio. Un gruppo di ricercatori di Stanford sostiene che ChatGPT-4o possiede una capacità di ragionamento morale migliore di quello umano; anche se la qualità scientifica dello studio è poco convincente, in rete ha suscitato discussioni, perché le dimensioni etiche dei comportamenti di ChatGPT e delle Intelligenze Artificiali (IA) sono al centro di riflessioni teoriche.

ChatGPT dice anche di possedere qualcosa che somiglia alla ‘intelligenza cristallizzata’, ma “non possiedo un’intelligenza fluida nel senso umano del termine”. Lo si vede per la verità. I ricercatori cercano di dotare queste macchine di Intelligenza Generale, sarà interessante vedere quali scorciatoie dovranno inventarsi.

I *Large Language Models* (LLM), tra cui la serie GPT, hanno imposto una svolta epistemologica all’intelligenza artificiale, cioè una innovazione epocale nell’evoluzione dell’intelligenza artificiale e della generazione del linguaggio naturale. Un LLM è un tipo di modello basato su reti neurali che viene addestrato su enormi quantità di dati testuali per comprendere e generare il

linguaggio umano. Questi modelli sono in grado di apprendere schemi statistici e relazioni complesse tra le parole del linguaggio, consentendo loro di prevedere le parole successive sulla base di una sequenza di *token* in ingresso. Come fanno alcune tecniche di sequenziamento dei genomi, un LLM scompone il testo in piccole unità, chiamate *token*, e utilizza queste rappresentazioni numeriche per elaborare le informazioni in modo efficiente. Ciò che rende gli LLM così impressionanti sono le dimensioni e la complessità. Il modello più grande attualmente disponibile, da quello che dice di se stesso ChatGPT-4o, è dotato di circa 500 miliardi di parametri (la versione 3 ne aveva circa 175 miliardi), di strati di reti neurali tra 120 e 160 (96 nella versione 3), una lunghezza del contesto di almeno 128.000 *token* (nella 3 erano tra 2.048 e 3.096), che consente prestazioni “più robuste, meno soggette a errori o allucinazioni”. Ovviamente parliamo di dati pubblici. E sappiamo che molti dati pubblici sono di qualità più scarsa di quelli accessibili solo dietro pagamento. Ma questo sarebbe un problema che da solo meriterebbe una approfondita e analitica considerazione.

I biometristi britannici di fine Ottocento, Francis Galton e Karl Pearson, per esempio, rimarrebbero di certo estasiati a vedere come la statistica meccanizzata può esprimere cotanta potenza e risolvere un’infinità di problemi diversi. Le macchine che oggi chiamiamo ‘intelligenti’ fanno in modo automatico e con efficienza impressionante un lavoro di accumulazione e ricerca di correlazioni tra dati raccolti empiricamente, per trovare invarianze o qualche indicatore utile a metter ordine nella variabilità del mondo naturale o sociale. Quello che si praticava in alcuni contesti di ricerca naturalistica ai tempi della seconda rivoluzione scientifica. In quel modo e con più lentezza, gli psicologi estrassero statisticamente da migliaia di test il fattore *g* dell’intelligenza umana e svilupparono le misure di QI, gli epidemiologi scoprirono le prime leggi delle epidemie compilando mappe o tabelle di casi e morti (senza conoscere i parassiti), ecc.

I procedimenti che usano alcune tipologie di algoritmi per evolvere capacità di categorizzazione degli stimoli in modi più rapidi ed efficienti (ma non più creativi, sempre per le caratteristiche materiali diverse dei sistemi, anche se l’IA si basa su categorizzazioni che sono molto più efficienti – sono proceduralmente simili a quelle che fa di suo il cervello – di quelle che produciamo comunicando naturalmente tra noi imprigionati dal *bias* dell’essenzialismo) delle nostre reti nervose sorprendono continuamente con prestazioni basate sulla forza bruta (*big data*), cioè in grado di immagazzinare

ed estrarre, statisticamente e comparando immense quantità di dati più o meno strutturati, informazioni, previsioni, decisioni, ecc. Si tratta di macchine che hanno migliorato la pianificazione qualitativa della nostra quotidianità. Ogni scelta significa condividere dati – con persone o macchine che siano – ovvero si può essere profilati. Alcuni algoritmi, usando dati che diffondiamo o produciamo navigando, imparano a conoscerci meglio di quanto crediamo di conoscere noi stessi. Si intravedono fenomenologie di tracciamento e anticipazione delle scelte, in ambiti che vanno dal voto politico alla scelta dei libri di leggere. Dinamiche quasi socio-meccaniche o deterministiche, cioè che passano sopra le teste dei singoli utenti. Il benessere di cui godiamo nelle società aperte dipende dall'attenzione per la libertà personale, lo stato di diritto, la trasparenza delle procedure pubbliche, l'equità, ecc., per cui serve fare una costante manutenzione etica e legale degli usi di questi strumenti. Problemi che non hanno i sistemi totalitari.

Come ripete l'informatico e Turing Medal Judea Pearl, l'IA rimane comunque ferma, per ora, alla scoperta delle associazioni, il livello più basso di intelligenza. Non manipola i contesti o non pensa controintuitivamente e controfattualmente (in parte riesce a costruire controfattuali ma in modo indiretto). L'IA, oggi, è solo in grado, con molta ma molta più efficienza dell'uomo, di rilevare strutture significative all'interno di basi di dati anche molto ampi. Il fatto che vinca a scacchi o a GO, che sappia progettare farmaci a livello molecolare, o guidare auto o fingere di essere un servizio clienti umano, dimostra solo che la gamma di domini dove questa capacità di uso superficiale dei dati si può applicare in modi adattativi è più ampia di quanto inizialmente si pensava. Secondo Pearl, il giorno in cui l'IA saprà approssimarsi all'intelligenza umana è vicino, ma le sue capacità vanno giudicate su tre livelli di abilità cognitive: vedere (associazione), fare (intervento) e immaginare (controfattuali). L'IA oggi lavora solo al livello più basso, cioè vedere. Io sarei meno ottimista sull'evoluzione che va oltre il fare associazioni. Non sono un tecnico sull'argomento e quel che dice Pearl mi deve sempre far ripensare. Ritengo comunque che la sua scala della conoscenza e l'enfasi sulla causalità come fondamento epistemologico della conoscenza valida sia ineccepibile, se non si è banali relativisti.

Ma già può tanto, se basta una quantità sufficiente di dati forniti a una macchina statistica che lavora in modi che non capiamo (opachi), per prevedere con affidabilità superiore a qualunque umano se una macchia radiografica è un

raro tumore, ovvero cosa ci faranno decidere di comprare o a cosa giocare le centinaia di *bias* cognitivi ed emotivi cablati del nostro cervello dalla selezione naturale. *Bias* che l'IA può usare per scopi paternalistici – se impara i nostri *bias*, quello paternalistico è forse più influente sui nostri giudizi, di quello di conferma – ma che contengono tanti pregiudizi con cui categorizziamo spontaneamente il mondo. Questi entrano nei dati e si trasferiscono nelle macchine quando le addestriamo: per cui possono discriminare persone obese se usate per gestire un ospedale, o persone di colore se applicate a un sistema penale nordamericano. Fanno quello, anche nel peggio, che di norma noi umani facciamo. Viene da dire che queste macchine imparano molto in fretta a riprodurre le cose peggiori di noi, mentre quelle migliori le impara solo associando ed elaborando quelle che abbiamo già prodotto di nostro, usando i nostri “migliori angeli” (scusate il senso forse troppo ampio che sto dando a un celebre passaggio di un discorso di Abraham Lincoln).

Esistono incertezze nei possibili sviluppi sociali, per così dire spontanei, delle popolazioni di algoritmi, che potrebbero trovare piani di autorganizzazione e minacciare i nostri valori: Norbert Wiener anche di questo parlava in *L'uso umano degli esseri umani* (1950). Ci sono paesi totalitari che aspirano a sviluppare sistemi normativi di controllo pubblico basati su graduatorie più o meno fluide, gestite da algoritmi appositamente istruiti, di crediti sociali. Egli pensa a scenari per cui si potrebbero generare attraverso dinamiche tipo ‘mano invisibile’ e a seguito di spinte ‘gravitazionali’, conseguenze inintenzionali negative, rispetto agli usi utili e alle conquiste politico-sociali raggiunte.

Da circa un decennio le macchine basate su IA hanno effetti ambivalenti sulla qualità della vita dei ragazzi, in particolare di quelli della c.d. iGen. I *Centers for Disease Control and Prevention* riportano che, almeno dal 2021, negli Stati Uniti più di metà delle ragazze statunitensi si descrivono tristi e senza speranze, e il 30% pensa al suicidio. Negli altri paesi occidentali le cose non vanno molto meglio. La psicologa dell'adolescenza Jean Twenge segnala che i test PISA in 37 paesi mostrano che dal 2012 la solitudine (forte indicatore per il rischio di depressione) dei giovani a scuola, cioè nell'ambiente più stimolante per la socializzazione, è aumentata esponenzialmente in modi drammatici. Twenge e altri sospettano che la causa sia da ricercarsi nel fatto che dal 2011-12 i giovani vivono per ore, soprattutto le ragazze, sui social media e guardano il mondo attraverso la fotocamera frontale degli *smartphone*,

cioè postando foto personali ed esponendosi a pubblici apprezzamenti, manifestazioni di odio, indifferenza, ecc. Comportamenti altamente insani che sono incentivati da IA dedicate. Anche questa variabile è da considerare nel pensare al futuro dei rapporti con queste macchine. Non dimentichiamoci, comunque, che questi giovani non leggono i libri che leggevamo noi o socializzano meno di noi (questo effettivamente per come ci siamo evoluti è un problema) e che possono usare molte più informazioni di noi, che non sono solo mis- o dis-informazioni. Sarà la selezione sociale, come sempre, a plasmare i tratti comportamentali delle prossime generazioni. L'idea di controllare e pianificare quel che si pensa sia meglio per loro, la trovo assai più rischiosa (oltre che inutile).

D'altro canto, gli stessi chatbot basati su IA, sono in grado di alleviare, con la stessa efficacia degli psicoterapeuti, i disagi mentali più lievi (ansia, attacchi di panico e disturbi non gravi dell'umore), come dimostra uno studio clinico recentemente pubblicato sul *New England Medical Journal*.

Se si chiede a ChatGPT cosa 'pensa' della coscienza artificiale, la risposta, che può cambiare, prevedibilmente, a seconda di come si pone la domanda, tende a essere mediamente come la seguente: "La mia prospettiva è che la coscienza artificiale, come tradizionalmente definita, non esiste, al momento, e ci sono forti ragioni per dubitare che mai esisterà – almeno nel modo in cui gli uomini e gli animali (sic!) ne hanno esperienza". Trattandosi di una macchina statistica, questo il significato di *Large Language Model*, la risposta riflette gli argomenti scientificamente più frequenti. Il che appare strano, perché nei media e tra chi pubblica libri, sembra sia più diffusa l'idea che l'IA sia lì lì per prendere il potere e schiavizzarci.

Tutti sono a raccontarsi che le macchine 'pensanti' trasformeranno l'esperienza umana. Ma è una tautologia. Così è stato per la scoperta del fuoco, o l'invenzione della stampa a caratteri mobili e dell'elettricità. Il fatto è che queste macchine, come diversi prodotti della modernità e del progresso, sono dissonanti rispetto al profilo genotipico/epigenotipico/fenotipico della specie, in particolare il comportamento, selezionato nell'ambiente pleistocenico. Potrebbe trattarsi di un *mismatch* dello stesso genere di quello che fa aumentare obesità, diabete, malattie cardiovascolari, in parte il cancro

e alcuni disturbi mentali, ecc. Al momento non sappiamo e non abbiamo dati o strumenti affidabili per fare previsioni. Quel che succede ai ragazzi e adolescenti per l'impatto psicologico degli *smartphone*, di cui si è già detto, è un po' inquietante. Ma è altra cosa. Verosimilmente saranno questi i problemi che può generare l'IA, non la 'singolarità', la coscienza artificiale o che prenda il comando del mondo. E come tutti i problemi che ha affrontato la specie saranno risolti in qualche maniera.

Se leggiamo Daron Acemoglu, premio Nobel per l'economia l'anno scorso, l'IA rischierebbe di andare incontro a un terzo inverno. Cioè a un crollo dei finanziamenti, come avvenuto due volte in passato, per il fatto di non rispondere alle aspettative del mercato. Il crollo momentaneo dei titoli delle multinazionali coinvolte nello sviluppo dell'IA, all'annuncio di *DeepSeek*, rende l'idea di cosa potrebbe succedere. Acemoglu ce l'ha con la macroeconomia dell'IA, che in base ai suoi modelli econometrici creerà instabilità sociali, per l'impatto che avrà sul mondo del lavoro e per incrementare pericolosamente le iniquità. Non capisco le tesi diffuse contro il capitalismo, in generale. Se una parte consistente della specie umana è uscita dalla condizione di 'minorità', conquistando un progresso generale della qualità della vita e della società mai esistito prima (inclusa l'IA) è stato per capitalismo, scienza/tecnologia e diritto negativo. Siamo rimasti una specie indottrinabile, aggressiva e incline alla dominanza. Il problema siamo noi, non l'IA. I demoni della nostra natura possono prevalere ancora una volta sui nostri "angeli migliori", dati gli scenari, non solo nei paesi meno sviluppati, ma ovunque, dove soffiano o tornano a soffiare i venti del nazionalismo, dell'autoritarismo, delle guerre, del proibizionismo, ecc. Indicatori di arretratezza umana politico-morale. Del resto, come diceva un bel po' prima di Lincoln, James Madison, gli uomini non sono angeli, e per questo serve un governo, che sia però leggero e serva fondamentalmente a consentire l'esercizio delle libertà individuali senza causare danni ad altre persone e alle loro proprietà.

Verosimilmente la coscienza non è uno spazio di decisioni eticamente migliori o più responsabili. La coscienza potrebbe benissimo essere un sottoprodotto di altre funzioni cognitive e non un tratto fenotipico direttamente adattativo. Alcuni studiosi, che non sono peraltro del tutto irrilevanti scientificamente, ritengono che sarebbe stata reclutata dalla selezione naturale per razionalizzare decisioni/scelte che prendiamo 'di nascosto', cioè che il nostro cervello prende computando quel 90-95% di

operazioni che avvengono al di sotto del livello della coscienza, o per creare informazioni false, utili per l'autoinganno. Luoghi comuni come 'fare un esame di coscienza', ma anche 'obiezione di coscienza' o 'voto di/appello alla coscienza', potrebbero essere quasi contraddizioni in termini. Se le l'IA non svilupperanno una coscienza come la nostra, è meglio. Ma l'IA potrebbe acquisire un qualche stato funzionale che noi non conosciamo qui e ora, cioè nel mondo a noi opaco degli algoritmi di *Machine* e *Deep Learning* potrebbe evolvere per selezione o essere implementato qualche tratto/algoritmo che va oltre il funzionamento statistico, e consentire la salita sui pioli più alti della "scala di Pearl". Cioè salire oltre la correlazione, dove si trova ora, e acquisire la capacità di interrogare il mondo attraverso manipolazioni/esperimenti o, a un livello più prossimo a noi, di scoprire le reti le causalità che tengono insieme l'universo usando ragionamenti controfattuali.

Mentre siamo tutti concentrati sulle prestazioni degli LLM, diversi ingegneri creato robot che usano l'IA per generare condizioni che consentano ai sistemi cognitivi artificiali di evolvere nelle strategie cognitive, applicando prestazioni umane che ci definiscono come dotati di intelligenza. Per esempio, i 'robot scientist' che sono sistemi autonomi basati su tecnologie di intelligenza artificiale, come l'apprendimento automatico, e che sono in grado di esperimenti scientifici e analizzare dati. Le scoperte dei robot scientist stanno avendo un impatto significativo sulla ricerca scientifica, poiché accelerano il processo di scoperta e aumentano l'efficienza nella formulazione di nuove ipotesi. Anche i robot scientist mancano di intuizione e creatività umana: li limita la loro capacità di affrontare problemi complessi che richiedono un pensiero critico.

Hod Lipson, a sua volta, costruisce robot autonomi che possono apprendere e adattarsi all'ambiente circostante. Questi sistemi robotici sono in grado di autoriconoscersi e di prendere decisioni basate sulle esperienze precedenti. Utilizzano algoritmi avanzati di apprendimento automatico per migliorare le loro performance nel tempo. Ovviamente non possiedono una consapevolezza cosciente. Si tratta sempre di modelli matematici e statistici per riconoscere schemi e fare previsioni. L'apprendimento automatico può simulare comportamenti intelligenti, il che comunque dimostra, ancora una volta, che non è l'intelligenza soltanto a cui dobbiamo guardare e che una significativa parte delle nostre prestazioni sono comunque meccanizzabili.

Il problema della ‘coscienza delle macchine’ eccita sempre molto, ma se avesse senso la teoria dell’integrazione dell’informazione (IIT) proposta da Giulio Tononi per cui la coscienza deriva dalla capacità di un sistema di integrare informazioni in modo complesso e unico, dovremmo “dargliela su” come si dice in Romagna. Secondo questa teoria, le macchine, che operano attraverso algoritmi e dati, non possono, per definizione, raggiungere il livello di integrazione necessario per una coscienza autentica. Questo porta a concludere che, anche con il progresso tecnologico, le macchine non possiederanno mai una coscienza simile a quella umana. La teoria IIT enfatizza quindi la connessione tra coscienza e complessità del sistema, proponendo che la coscienza sia un fenomeno emergente strettamente legato alla biologia. La coscienza è legata a esperienze soggettive e stati mentali, che sono sistemi biologici evolutisi per selezione naturale in dati ambienti e che sono riusciti a conquistare attraverso il cablaggio di sistemi neurali unici e non dissezionabili funzionalmente. Con buona pace della metafisica funzionalista.

Se si chiede a ChatGPT-4o se pensa che si potrebbero fare innesti tra i robot in grado di autoriconoscersi e sviluppare comportamenti adattativi con l’IA generativa, e cosa questo comporterebbe sul piano innovativo, si ottiene un’articolata risposta: “combinare robot che si autoriconoscono (macchine capaci di modellare e monitorare il proprio corpo) con modelli linguistici di grandi dimensioni, come GPT-4, in grado di ragionare e riflettere [allora chissà quanto, n.d.a.], potrebbe portare allo sviluppo di una cognizione artificiale di ordine superiore. Un sistema integrato di questo tipo non solo potrebbe agire e percepire, ma anche interpretare le proprie azioni, riflettere sugli errori, pianificare miglioramenti e persino spiegare il proprio ragionamento – in modo simile a come fanno gli esseri umani. Questo approccio unisce intelligenza sensorimotoria e cognizione simbolica e narrativa, permettendo lo sviluppo di robot non solo intelligenti, ma anche auto-consapevoli e adattivi. Restano delle sfide – come l’integrazione in tempo reale e il radicamento nei dati sensoriali – ma la direzione è promettente per una vera intelligenza artificiale riflessiva e autonoma”.

Gli autori

Giulia Andrighetto

Dirigente di ricerca presso l'Istituto di Scienze e Tecnologie della Cognizione (ISTC) del Consiglio Nazionale delle Ricerche (CNR) a Roma, dove coordina il *Laboratory of Agent Based Social Simulation* (LABSS). È inoltre ricercatrice presso l'*Institute for Future Studies* a Stoccolma. La sua ricerca esamina la natura e la dinamica delle norme sociali e il loro ruolo nel sostenere la cooperazione. In particolare, si focalizza su come le norme possano emergere e diventare stabili, i meccanismi alla base del loro cambiamento, come sia possibile che sopravvivano norme inefficienti o impopolari. Per rispondere a queste domande, combina modelli teorici con esperimenti con soggetti umani e con agenti artificiali (simulazioni al computer). I suoi lavori sono stati pubblicati su riviste internazionali tra cui *Nature Communications*, *PNAS Nexus*, *Philosophical Transactions of the Royal Society B*. Ha coordinato e partecipato a diversi progetti nazionali e internazionali, finanziati dalla Commissione Europea, dal Ministero dell'Università e della Ricerca (MUR), dallo *Swedish Research Council* e dalla *Wallenberg Foundation*.

Guido Boella

Professore ordinario presso il Dipartimento di Informatica e Vice-Rettore Vicario dell'Università di Torino per la promozione dei rapporti con le imprese e le associazioni di categoria delle imprese e per il coordinamento con le iniziative di innovazione industriale sul territorio. È Vicepresidente del *Competence Center CIM4.0* e membro del Comitato per l'aggiornamento della strategia AI della Presidenza del Consiglio. È anche co-fondatore della Società Italiana per l'Etica dell'Intelligenza Artificiale (SIpEIA) e coordinatore del Magazine Intelligenza Artificiale *magia.news*. È stato coordinatore di progetti

regionali ed europei (ICT4LAW, EUCases, CANP, WeGovNow, Co-city, CO3, PININ, CORPUS, NLAB4CIT) e del dottorato internazionale in *Law, Science and Technology* LAST-JD. È coordinatore dell'*European Digital Innovation Hub Circular Health* EDIH (CHEDIH) e vicecoordinatore di *Public Administration Intelligence* (PAI EDIH). Fondatore dello *spinoff* universitario Nomotika.

Stefano Canali

Ricercatore presso il Dipartimento di Elettronica, Informazione e Bioingegneria del Politecnico di Milano. Si occupa di filosofia della scienza e della medicina, con particolare interesse per aspetti epistemologici, etici e sociali di tecnologie quali *big data*, salute digitale e intelligenza artificiale nella medicina contemporanea. Si è formato e ha lavorato su questi temi presso l'Institut für Philosophie della *Leibniz Universität Hannover*, il *Science and Technology Studies Department* di *University College London* e il Dipartimento di Filosofia dell'Università degli Studi di Milano.

Cinzia Caporale

Coordinatrice del Centro Interdipartimentale per l'Etica e l'Integrità nella Ricerca (CID Ethics) e della Commissione per l'Etica e l'Integrità nella Ricerca del Consiglio Nazionale delle Ricerche (CNR). È membro del Comitato Nazionale per la Bioetica (CNB) dal 2002, del Comitato etico nazionale per le terapie avanzate, della Consulta scientifica del Cortile dei Gentili e di diversi Comitati per l'integrità nella ricerca di Atenei italiani. Ha presieduto il Comitato etico unico nazionale per le sperimentazioni su Covid-19 ed è stata membro del Comitato Tecnico Scientifico (CTS) per la gestione della pandemia (governo Draghi). Inoltre, è stata Presidente del Comitato Intergovernativo di Bioetica (IGBC) dell'UNESCO per due mandati, nonché capo della Delegazione italiana sull'etica, membro della *World Commission on the Ethics of Scientific Knowledge and Technology* (COMEST) e del *Legal experts Group for UNESCO GEObs-law database*, ed *Ethics Mentor* del progetto ERC- XAI (*eXplanation of AI decision making*).

Carlo Casonato

Professore ordinario di Diritto costituzionale comparato e titolare della Cattedra *Jean Monnet* di Diritto dell'Intelligenza Artificiale (T4F). È fondatore e *chief editor* del *BioLaw Journal* e delegato del rettore e vicepresidente del Comitato Etico di Ricerca dell'Università di Trento. È stato *Visiting Fellow* presso l'Università di Yale, l'Università di Oxford, l'*Illinois Institute of Technology* (Chicago) e l'*Universidad del País Vasco*. Ha fatto parte del Partenariato Globale dell'OCSE sull'Intelligenza Artificiale (GPAI) e del Comitato Nazionale per la Bioetica. Fa parte della Commissione per l'Etica e l'Integrità nella Ricerca del Consiglio Nazionale delle Ricerche (CNR).

Gilberto Corbellini

Professore ordinario di storia della medicina, insegna bioetica alla Sapienza Università di Roma, dove è stato direttore del Museo di storia della medicina fino al 2017. Laureatosi in filosofia della scienza con una tesi sull'epistemologia evoluzionistica di *Donald Campbell*, *Konrad Lorenz* e *Karl Popper*, ha successivamente conseguito il dottorato in sanità pubblica. I suoi primi interessi di studio hanno riguardato la storia e la filosofia della biologia evoluzionistica, delle immunoscienze e delle neuroscienze, per includere quindi lo studio della storia della malaria e della malariologia in Italia, delle ricadute della genetica molecolare in medicina, delle implicazioni del pensiero evoluzionistico darwiniano per la medicina e l'evoluzione della pedagogia medica. È autore di numerosi articoli, saggi e monografie su temi di storia della medicina, epistemologia, bioetica e neuroetica.

Marta Fasan

Assegnista di ricerca in Diritto costituzionale comparato presso la Facoltà di Giurisprudenza dell'Università degli Studi di Trento, dove ha ottenuto il titolo di dottoressa di ricerca in *Studi Giuridici Comparati ed Europei*. Ha svolto periodi di ricerca presso centri di ricerca nazionali (Istituto italiano di Tecnologia) e università estere (*Centre de Recherche en Droit Publique, Université de Montréal; Health Law Centre, Lund University*). La sua attività di ricerca indaga il rapporto tra intelligenza artificiale e diritto costituzionale, analizzando, in prospettiva comparata, l'impatto prodotto da questa tecnologia sulle categorie del costituzionalismo contemporaneo.

Roberta Ferrario

Prima Ricercatrice presso l'Istituto di Scienze e Tecnologie della Cognizione (ISTC) del Consiglio Nazionale delle Ricerche (CNR), dove lavora al Laboratorio di Ontologia Applicata, con sede a Trento. In passato si è occupata di ontologia della mente, delle organizzazioni e dei servizi, mentre i suoi interessi di ricerca attuali spaziano dall'ontologia dei sistemi sociotecnici, all'intelligenza artificiale ibrida e trustworthy, alle ontologie per l'informatica umanistica e il patrimonio culturale. Ha ricoperto in anni recenti la posizione di *editor-in-chief* della rivista *Applied Ontology* ed è stata *General Chair* e *PC Chair* di varie edizioni della conferenza *Formal Ontologies in Information Systems*. È autrice di una monografia e oltre un centinaio di articoli scientifici in ontologia fondazionale e sue applicazioni.

Fabio Fossa

Ricercatore in filosofia morale presso il Dipartimento di Ingegneria Meccanica del Politecnico di Milano, dove si occupa di questioni etiche relative alle tecnologie della mobilità, con particolare attenzione alla guida autonoma e da remoto. La sua ricerca verte su temi di etica applicata, filosofia della tecnologia, etica della robotica e dell'IA, e sul pensiero di Hans Jonas. Tra le sue pubblicazioni: *Ethics of Driving Automation. Artificial Agency and Human Values*, Springer 2023.

Mattia Fumagalli

Ricercatore presso la Libera Università di Bolzano, dove insegna *Information Systems Design and Empirical Research Methods*. La sua ricerca si concentra principalmente sull'intelligenza artificiale, con particolare attenzione al supporto automatizzato per la rappresentazione della conoscenza e la modellazione concettuale. Altri suoi interessi di ricerca includono la logica e la filosofia della mente. Collabora attivamente con il *Laboratory for Applied Ontology* (LOA) del Consiglio Nazionale delle Ricerche (CNR), con il gruppo *Semantics, Cybersecurity, and Services* (SCS) dell'Università di Twente e, occasionalmente, con il gruppo *Knowdive* del Dipartimento di Ingegneria dell'Informazione e Scienze Informatiche dell'Università di Trento, dove ha conseguito il dottorato. Dal 2015 al 2020, è stato assistente alla docenza per i corsi di *Computational Logics e Knowledge-Data Integration* presso la stessa università.

Ludovica Marinucci

Responsabile dell'Unità di Ricerca su Etica e Intelligenza Artificiale presso il Centro Interdipartimentale per l'Etica e l'Integrità nella Ricerca (CID Ethics) del CNR. Svolge attività di ricerca relativa all'analisi dei profili etici delle tecnologie emergenti e alla valutazione del loro impatto sulla società al fine di definire raccomandazioni, codici, linee guida e nuovi paradigmi etici finalizzati alla progettazione di sistemi di IA e robotica eticamente sostenibili, in particolare nell'ambito dei progetti europei EYE-TEACH e PILLAR-Robots. Fa parte della Segreteria scientifica del Gruppo di Lavoro *Etica della ricerca nella robotica* della Commissione per l'Etica e l'Integrità nella Ricerca del CNR. È docente a contratto presso vari atenei italiani (Università degli Studi di Roma Tor Vergata, Università degli Studi di Salerno, Università Internazionale Uninettuno) su temi di storia e filosofia della tecnologia, filosofia dell'informatica, etica dell'IA.

Giuseppe Primiero

Professore associato presso il *Logic, Uncertainty, Computation and Information Lab* del Dipartimento di Filosofia dell'Università degli Studi di Milano. Inoltre, ricopre il ruolo di Direttore Scientifico del *Research Center for the Philosophy of Technology* (PHILTECH) della stessa università. Si occupa di modellazione formale e verifica di sistemi multi-agente, oltre che di filosofia della computazione. I suoi strumenti preferiti sono *proof-systems*, logica modale e computazionale. La sua ricerca formale si applica ai sistemi di intelligenza artificiale e al loro utilizzo per risolvere problemi di misinformazione e disinformazione online, approcci computabili alla valutazione dell'affidabilità delle fonti di informazione, dei pregiudizi e dell'equità.

Marta Tomasi

Professoressa associata di Diritto costituzionale comparato presso la Facoltà di Giurisprudenza dell'Università di Trento. È vice-direttrice di *BioLaw Journal* e Vice-Presidente del Comitato Etico Territoriale della Provincia Autonoma di Trento per le sperimentazioni cliniche e del Comitato Etico provinciale della Provincia Autonoma di Bolzano. Ha svolto periodi di ricerca presso l'*Hastings College of the Law* di San Francisco, il *King's College*

di Londra, il *Barcelona Supercomputing Center* ed è autrice di numerose pubblicazioni scientifiche su riviste nazionali e internazionali dedicate ai temi del diritto costituzionale e del biodiritto.

Luca Tummolini

Dirigente di ricerca presso l'Istituto di Scienze e Tecnologie della Cognizione (ISTC) del Consiglio Nazionale delle Ricerche (CNR) a Roma. Ha conseguito il Dottorato di Ricerca in Scienze Cognitive presso l'Università di Siena. I suoi interessi di ricerca riguardano l'interazione sociale e i meccanismi cognitivi che consentono agli esseri umani di coordinarsi e collaborare in modo flessibile tra loro sulla base di rappresentazioni concettuali condivise e norme sociali. Nel suo lavoro combina modelli formali (modelli computazionali, teoria dei giochi) e metodi sperimentali dalla psicologia cognitiva all'economia sperimentale. Ha pubblicato su riviste di filosofia, psicologia, economia e informatica tra cui *Synthese*, *Psychological Bulletin*, *Nature Communications*, *Philosophical Transactions of the Royal Society B*. È *associate editor* delle riviste *Computational Intelligence*, *Topoi: An International Review of Philosophy* e *Frontiers in Psychology: Theoretical and Philosophical Psychology*.

Giacomo Zanotti

Assegnista di ricerca in filosofia della scienza presso il Dipartimento di Elettronica, Informazione e Bioingegneria del Politecnico di Milano. Si occupa in particolare di filosofia dell'intelligenza artificiale, adottando una prospettiva primariamente epistemologica. Nello specifico, la sua ricerca si concentra sui temi della fiducia, del rischio e dell'incertezza in intelligenza artificiale, oltre che sulla trasparenza nei sistemi autonomi e intelligenti.