CRISTIANO CALÌ

COME CI CAMBIA LA TECNOLOGIA.

L'AGENCY DELLE AI E LA CAPACITÀ COGNITIVA

DI PRENDERE DECISIONI RAZIONALI

ABSTRACT: HOW TECHNOLOGY CHANGES US. *THE AGENCY OF AI AND THE COGNITIVE ABILITY TO MAKE RATIONAL DECISIONS*
The article explores the concept of agency in the context of artificial intelligence, analyzing the approach of Stuart Russell and Peter Norvig. After examining three possible levels of machine agency, the contribution reflects on the challenge of attributing agency to algorithms by proposing two interaction paradigms: "substitutivist" and "complementarist". The latter suggests collaboration between humans and AI, emphasizing a shared decision-making process. The ultimate aim of the contribution is to raise ethical considerations regarding autonomous AI decisions on one hand and the human-machine interaction on the other, especially when such usage involves the cognitive ability to make free decisions.

1. *Introduzione*[1]

A partire dagli anni Cinquanta del secolo scorso, gli algoritmi in grado di compiere operazioni un tempo appannaggio esclusivo degli esseri umani sono state definite

[1] Per la prima parte di questo contributo mi sono rifatto a una presentazione orale tenuta durante la conferenza annuale della *SISEC - Società Italiana di Sociologia Economica,* svoltasi a Brescia nel febbraio 2023. Il testo di quella presentazione è rinvenibile in C. Calì, *Razionalità degli agenti e "razionalità" delle macchine. Un tentativo di confronto*, in *Internet nel nuovo. Millennio. Diritto, società e tecnologia*, a cura di A. Alù, A. Ciccarello, *Internet nel nuovo. Millennio. Diritto, società e tecnologia*, Editoriale Scientifica, Napoli 2023, in corso di stampa. Una precedente versione inedita del presente saggio è stata premiata col premio *Myllennium Award 2023*, per la sezione *MyBook*.

intelligenze artificiali (AI); una nozione talmente controversa che Stuart Russell (Docente a Berkeley) e Peter Norvig (Direttore della ricerca in *Google*), due tra i più importanti studiosi di AI, nella prefazione alla quarta edizione italiana del loro manuale – utilizzato nelle maggiori accademie per introdurre all'intelligenza artificiale – dichiarano come al centro del loro testo vi sia la nozione di *agente*. Questa definizione, a detta loro, dovrebbe prendere il posto di quella molto più comune di *intelligenza artificiale*, dal momento che l'AI sarebbe soltanto «lo studio degli agenti che ricevono percezioni dall'ambiente ed eseguono azioni»[2]. Tale disciplina, tuttavia, non si limiterebbe allo studio degli agenti ma, *in primis*, permetterebbe di *costruire* agenti[3]. Risulta fondamentale, allora, cogliere cosa i due autori – veri e propri alfieri dell'AI – intendano con *agente*, soprattutto in ragione del fatto che essi assumono questo concetto nella sua accezione più pregnante.

La mia riflessione sarà volta, pertanto, a intercettare se e come sia possibile predicare l'*agency* di alcuni algoritmi di AI, e, quindi, a comprendere le modalità di interazione di quest'*agency* con l'*agency* umana.

A tal fine suggerirò due modalità d'interazione che, facendo ricorso a una sorta di neologismo, chiamerò paradigma *sostituivista* e *complementarista*. Dall'analisi di alcuni casi concreti legati al mondo della giurisprudenza, s'indicherà, infine, quale ritengo essere la sfida nel momento in cui gli algoritmi assumono un ruolo determinante proprio su quella capacità che è comunamente

[2] S. Russell, P. Norvig, *Intelligenza artificiale. Un approccio moderno* (2016), tr. it. Pearson, Milano-Torino 2021, 1, p. XXIV. Assunta in questa accezione l'intelligenza artificiale potrebbe essere una disciplina che studia anche l'essere umano, al pari dell'antropologia, e sono gli stessi autori del manuale a favorire una tale lettura nel momento in cui si afferma che «ogni agente implementa una funzione che mette in corrispondenza sequenze percettive e azioni», *ibid*.

[3] Credo che sia un caso più unico che raro, un vero e proprio *hapax*, il fatto che si parli di *costruzione* dell'agente e non di *costituzione* dell'agente o di semplice *analisi* dell'agente (espressioni sicuramente molto più comuni in diverse branche del sapere scientifico).

riconosciuta come "la cosa più propria che ciascuno di noi ha", ovvero la capacità di decidere liberamente dei nostri atti.

2. *Agenti o intelligenti?*

2.1. *Tre diversi gradi di "agency"*

Russell e Norvig sviluppano implicitamente (e forse anche inconsapevolmente) una riflessione "scalare" sulla nozione di agente, nel senso che a ogni grado di un particolare tipo di agente corrisponde un grado di sofisticatezza dell'AI intesa solo ed esclusivamente come disciplina.

Ad un primo livello (1) un'agente «è semplicemente qualcosa che agisce, che fa qualcosa»[4]. A questa nozione di agente corrisponde un'AI dedita a sviluppare macchine che emettano un *output* a partire da determinati *input* (per esempio un robot che pulisce il pavimento quando è sporco e che si spegne quando invece è pulito, o anche un qualsiasi altro *software* installato sul nostro *personal computer*). Questa dimensione più basica lascia spazio, tuttavia, ad un secondo livello, per il quale l'agente non è compreso soltanto come "colui che detiene l'*agere*" (nel senso di colui che *fa delle cose*) ma come colui che è "originatore delle proprie azioni", nel senso di colui che delibera sul da farsi. A questo secondo livello (2) – che è occupato da quello che può essere definito *agente artificiale* – corrisponde anche il secondo livello dell'AI, la quale non avrà più come obiettivo quello di costruire macchine-che-fanno-cose ma di elaborare «macchine in grado di *calcolare come agire* in modo efficace e sicuro in un'ampia varietà di situazioni nuove»[5].

Si badi bene che in questa affermazione, per quanto ambiziosa, si cela già un presupposto cruciale, per il quale la scelta è essa stessa azione, anche qualora essa non abbia un corrispettivo fenomenico osservabile esternamente (in altre parole: il semplice deliberare è esso stesso una modalità di agire).

[4] *Ibid.*, p. 6.
[5] *Ibid.*, p. 3 [corsivo mio].

Il "calcolare come agire", tuttavia, non è sufficiente, ed è qui che appare il terzo livello (3), quello dell'*agente* (artificiale) *razionale*, ovvero quello di un agente che «agisce in modo da ottenere il miglior risultato o, in condizioni di incertezza, il miglior risultato atteso»[6]. A questo livello dovrà corrispondere una disciplina in grado di progettare macchine che agiscano in vista di un fine (ottimizzandolo).

Si noti però come il parlare di *agenti razionali* non sia di certo più semplice che parlare di *intelligenze artificiali*. Quest'ulteriore definizione, che fa riferimento alla razionalità e non più all'intelligenza, infatti, non è scevra di complicazioni, dal momento che anche la razionalità, così come l'intelligenza, è ben lontana dall'essere raggiunta da parte delle AI.

Se però l'attributo di *razionale*, usato dai due ricercatori, è controverso, il sostantivo, di contro, è ormai pacificamente acquisito nel dibattitto. Gli algoritmi sono veri e propri *agenti*, e questo dato – benché possa sembrare inusuale – è molto più evidente di quanto non riconsegnino le altre definizioni.

Gli algoritmi oggigiorno, infatti, *prendono decisioni* e, dal momento che – stando a una tradizione filosofica e giuridica vastissima – «decidere è anche agire e tale verbo va assegnato ed esteso a chiunque agisca, si tratti dell'uomo o dell'intelligenza artificiale»[7], gli algoritmi non sono soltanto dei semplici agenti ma degli *agenti decisionali*.

Nel momento, allora, in cui si ha questa ambizione: progettare e costruire agenti artificiali razionali – ovvero agenti che, in vista di un fine/risultato (3), deliberano/calcolano (2) e quindi agiscono/riconsegnano *output* (1) – è necessario comprendere come tutto questo avvenga nella macchina-uomo per poi permettere ad una macchina, questa volta di silicio e non di carbonio, di

[6] *Ibid.*, p. 6.

[7] C. Canullo, *Chi decide? Intelligenza artificiale e trasformazioni del soggetto nella riflessione filosofica*, in E. Calzolaio (a cura di), *La decisione nel prisma dell'intelligenza artificiale*, Cedam, Milano 2020, p. 35.

implementarlo. Se questa volontà riproduttiva, infatti, può venire bypassata nel caso dell'intelligenza[8] non sembra potersi dire lo stesso per quanto concerne la deliberazione.

Essendo impossibile in questa sede soffermarsi su cosa comporti, ingegneristicamente parlando, la realizzazione di questo sogno (e se un approccio produttivo o riproduttivo possa essere più congeniale), è mia intenzione soffermarmi sulla nozione di *agente* che – a prescindere da come la intendono i due autori statunitensi – è divenuta un termine chiave nel dibattito odierno, dal momento che stante il fatto che al giorno d'oggi non abbiamo ancora *intelligenze* artificiali, è fuor di dubbio che abbiamo *agenti* artificiali.

2.2. *Artificial agency (AA)*

Tale suggerimento ci giunge da Luciano Floridi, per il quale sarebbe più opportuno non esprimersi più nei termini di *artificial intelligence* (AI) ma di *artificial agency* (AA): «Solo se si capisce che l'AI è una nuova forma di capacità di agire, o di *agency* […], e non una nuova forma di intelligenza, si può poi capire veramente la sua sfida etica e quindi affrontarla con successo»[9].

Questa capacità di agire in vista di uno scopo, allora, richiede di essere ben compresa, soprattutto in raffronto alla medesima capacità detenuta dall'essere umano, dal momento che, per la prima volta

[8] La modalità di approccio definita *riproduttiva* – dal momento che mira a riprodurre le capacità umane su artifici della tecnica – è stata percorso sin dagli albori delle AI ma si è dimostrata del tutto fallimentare per quanto riguarda l'intelligenza, tant'è che i maggiori progressi si sono riscontrati quando si è capito che le macchine non dovevano copiare i *Sapiens* o, per dirla con un'espressione ormai famosa: "L''uomo è riuscito a volare quando ha smesso di imitare gli uccelli". Quel paradigma, pertanto, è stato ormai abbandonato per ciò che attiene all'intelligenza ma non è possibile fare lo stesso per quanto concerne la capacità di deliberazione. Capire come gli esseri umani decidono risulta, infatti, cruciale, dal momento che, sino a non poco tempo fa, essi erano gli unici sul pianeta Terra capaci di decidere razionalmente ma, soprattutto, in grado di considerare, soppesare, attribuire valore a diversi corsi d'azione.

[9] L. Floridi, *Agere sine intelligere. L'intelligenza artificiale come nuova forma di agire e i suoi problemi etici*, in L. Floridi, F. Cabitza (a cura di), *Intelligenza artificiale. L'uso delle nuove macchine*, in Bompiani, Milano 2021, p. 145.

nella storia del nostro pianeta, l'*agency* viene "scollata" dall'intelletto.

Secondo il docente di Oxford, infatti, «oggi l'AI non è il matrimonio tra ingegneria (artefatti) e biologia (intelligenza almeno animale, se non umana), ma il divorzio dell'agire (*agency*) dalla necessità di essere intelligenti per aver successo»[10].

La cosiddetta AI, infatti, ha operato un divorzio tra *agere* e *intelligere*, tra *agency* artificiale e intelligenza naturale, il che è rivoluzionario:

> Non è mai stato possibile prima d'ora avere successo nello svolgimento di un compito senza essere intelligenti almeno come un cane pastore. Abbiamo modificato – prosegue Floridi – una delle equazioni fondamentali su cui si è sempre basata la storia umana e la valutazione morale, quella che identifica l'agire con l'agire naturale, almeno biologico se non umano[11].

Desidererei che questo fosse il punto di partenza. Al giorno d'oggi algoritmi in grado di esercitare un certo qual grado di *agency* sono di dominio comune. Cito soltanto i più comuni. I navigatori satellitari – come *Waze*, *Google Maps* o *Mappe* di Apple – non sono altro che algoritmi che selezionano per l'utente due/tre possibili percorsi alternativi tra una miriade di altre opzioni, incrociando dati come il flusso del traffico, la distanza in metri dal luogo d'interesse, le eventuali chiusure stradali.

I *software* in uso presso alcuni istituti di credito per l'assegnazione di mutui sono anch'essi algoritmi "decisionali", dal momento che – sulla base di vari parametri fiscali e non – *decidono* se a quel determinato utente può essere concesso o meno il mutuo.

Alcuni dei programmi utilizzati da diversi studi legali e società di consulenza, infine, sono algoritmi che sulla base di ripetizioni statiche di una parola o di tag preimpostati, sottopongono all'utente-umano alcuni dati e non altri.

Tutti questi casi ci debbono far riflettere sul dato che nel momento in cui ci avvaliamo di algoritmi per agevolare alcuni nostri processi decisionali, questi processi sono avvenuti in modo

[10] *Ibid.*, pp. 150-151.
[11] *Ibid.*, p. 151.

diametralmente differente da come li avrebbe condotti un essere umano. Questo però – lungi dal destare allarmismo – deve farci interrogare, perché, di fatto, stiamo realizzando una nuova modalità d'interazione che sino al giorno d'oggi non si era mai verificata. Infatti, se al tavolo di una riunione di *team*, sino a qualche tempo fa avrebbero avuto posto a sedere un gruppo di delegati delle varie arie di un'azienda – ciascuno con la possibilità di dire le motivazioni per le quali ha preso una determinata decisione (decisioni che si suppone essere state intelligentemente ponderate in vista di un fine) – oggi a quel medesimo tavolo non sediamo soltanto noi ma anche algoritmi i quali – pur senza avere intelligenza – decidono.

3. *Processi decisionali "a confronto"*

Per capire come le AI prendono decisioni è necessario volgersi brevemente ai metodi utilizzati per addestrarle.

3.1. *Come decidono le AI*

Uno dei modi più comuni per addestrare un'AI è quello di utilizzare l'apprendimento automatico, un processo che consente ai sistemi di imparare da dati-esempio senza essere esplicitamente programmati. Ci sono due tipi principali di apprendimento automatico: l'apprendimento supervisionato e quello non supervisionato[12].

L'apprendimento supervisionato utilizza dati etichettati, in cui ogni esempio è associato a una etichetta. Ad esempio, un'AI può essere addestrata a riconoscere immagini di gatti utilizzando un *set* di immagini di gatti etichettate come tali. Durante il processo di addestramento, l'AI "impara" a riconoscere le caratteristiche comuni delle immagini di gatti e le utilizza per classificare nuove immagini.

[12] Per un'introduzione, S. Russell, P. Norvig, *op. cit.*, 2, capp. 19-20.

L'apprendimento non supervisionato, invece, utilizza dati non etichettati e l'AI cerca di scoprire relazioni e *pattern* nel *dataset*[13].

Una intelligenza artificiale decide utilizzando un insieme di algoritmi e di modelli matematici che sono stati addestrati su dati specifici. Per esempio, un'AI addestrata per riconoscere immagini utilizza un algoritmo di apprendimento automatico, come una rete neurale, per analizzare le caratteristiche di un'immagine e classificarla in una categoria predefinita. La decisione di classificare un'immagine in una determinata categoria viene quindi presa in base a una serie di calcoli matematici volti a confrontare i dati provenienti dall'immagine e quelli forniti in fase di addestramento[14].

In generale, le decisioni prese da un'AI sono il risultato dell'elaborazione di grandi quantità di dati e dell'applicazione di algoritmi matematici sofisticati. Tuttavia, è importante notare che le decisioni prese da un'AI possono essere influenzate dai *bias* presenti nei dati di addestramento e dagli errori nell'implementazione dell'algoritmo.

Nel 2016, ad esempio, Microsoft ha introdotto su Twitter un'AI conversazionale di nome *Tay*. Gli utenti hanno iniziato a inviare a

[13] Un esempio di questo tipo di apprendimento è l'utilizzo di un algoritmo di *clustering* per raggruppare i dati in base alle loro similitudini. Le AI che utilizzano l'apprendimento supervisionato sono spesso utilizzate per compiti di classificazione, in cui la risposta desiderata è una categoria predefinita. Ad esempio, un'AI può essere addestrata a riconoscere se un'immagine contiene un gatto o un cane. Le AI che utilizzano l'apprendimento non supervisionato sono utilizzate per compiti di clustering e rilevamento di anomalie. Ad esempio, un'AI può essere utilizzata per rilevare transazioni fraudolente in un dataset di transazioni finanziarie. Oltre all'apprendimento automatico, ci sono anche altri metodi utilizzati per addestrare le AI, come l'apprendimento per rinforzo e l'apprendimento evolutivo. L'apprendimento per rinforzo utilizza un sistema di premi e punizioni per incoraggiare l'AI a prendere decisioni corrette, mentre l'apprendimento evolutivo utilizza algoritmi di ottimizzazione per far evolvere una soluzione iniziale verso una soluzione migliore.

[14] In altri casi, un'AI potrebbe utilizzare un algoritmo di apprendimento per rinforzo per prendere decisioni in un ambiente di gioco o di simulazione. In questo caso, l'AI esplora diverse azioni e riceve un premio o una punizione in base alla loro efficacia nel raggiungere un obiettivo predefinito. In base a queste informazioni, l'AI adatta la sua strategia per prendere decisioni più efficaci in futuro.

Tay ogni sorta di commenti misogini, razzisti e "Donald Trumpisti" e nel giro di ventiquattro ore *Tay* è diventato un teorico della cospirazione, misogino, razzista e xenofobo, e questo ha indotto Microsoft a staccare la spina[15].

Un rapporto del 2019 ha rilevato, inoltre, che gli algoritmi di riconoscimento facciale, come quelli che sbloccano i telefoni cellulari o identificano le persone nelle foto caricate sui *social media*, sono intrinsecamente razzisti e sessisti. Questo avverrebbe perché la maggior parte dei programmatori sono maschi e bianchi e, di conseguenza, il *set* di dati usato per addestrare questi algoritmi era fortemente influenzato[16]. Il problema dei *bias* esiste anche in medicina. Schemi di valutazione medica come il SOFA (*Sequential Organ Failure Assess*), ad esempio, denotano forti pregiudizi razziali[17]

Questi esempi del mondo reale dimostrano una realtà delle AI: esse tendono a riflettere i pregiudizi dei dati che vengono loro forniti. È cruciale, pertanto, valutare attentamente la validità e l'affidabilità delle decisioni prese da un'AI prima di utilizzarla in contesti critici.

Tuttavia, prescindendo dai *bias*, alla base dell'approccio sinora descritto vi è un pensiero che è molto più antico dell'AI, un pensiero per il quale pensare equivale a calcolare. E a tal proposito già Thomas Hobbes si era espresso in questi termini:

> Quando uno ragiona non fa altro che ottenere una somma totale attraverso una addizione di parti, o un resto sottraendo una somma da un'altra. […] E infatti come l'aritmetica ci insegna a sommare e a sottrarre in termini di numeri, così […] i logici insegnano le stesse cose nel campo della connessione fra le parole: sommando insieme due nomi si ha un'affermazione, sommando due affermazioni si ha un sillogismo, sommando alcuni sillogismi si ha una dimostrazione; e dalla somma, o conclusione di un sillogismo sottraggono una

[15] Cfr. C.M. Klugman, *Black Boxes and Bias in AI Challenge Autonomy*, in «*The American Journal of Bioethics*», XXI, 2021, p. 34.
[16] Cfr. P. Grother, M. Ngan, K. Hanaoka, *Face Recognition Vendor Test (FRVT). 3: Demographic effects. National Institute of Standards and Technology*, U.S. Department of Commerce, Washington D.C. 2019.
[17] W.D. Miller, M.E. Peek, W.F. Parker, *Scarce Resource Allocation Scores Threaten to Exacerbate Racial Disparities in Health Care*, in «Chest», 158, 2020, pp. 1332-1334; D.A. Vyas, L.G. Eisenstein, D.S. Jones, *Hidden in Plain Sight. Reconsidering the Use of Race Correction in Clinical Algorithms*, in «The New England Journal of Medicine», 383, 2021, pp. 874-882.

> proposizione per trovarne un'altra. […] Insomma in qualunque campo in cui c'è posto per l'addizione e la sottrazione c'è anche posto per la ragione; dove queste cose mancano la ragione non ha niente da fare. […] Poiché ragione in questo senso significa nient'altro che calcolo, cioè addizione e sottrazione[18].

In effetti, un primo modo per definire come agisce un agente razionale è quello di ricorrere alle inferenze, questo però non basta perché molte altre azioni razionali vengono compiute senza che prima vi sia stata un'inferenza di qualche tipo. In un primo tempo l'AI – partendo dal presupposto che il pensiero si potesse ridurre a calcolo – si è concentrata sul trasporre algoritmicamente le regole della logica del primo ordine.

In quei casi l'*agere* delle AI era circoscritto alla forma *if-then*. Ci si rese ben presto conto però che la logica era poco utile, se non altro per il fatto che la stragrande maggioranza della popolazione mondiale quando agisce non ricorre di certo al sillogismo aristotelico.

Trovo una certa difficoltà a comprendere come le parole di Hobbes possano essere state così superficiali, dal momento che se è vero che ben si addicono quando parliamo seguendo la logica aristotelica (per la quale da premesse vere si ottengono conclusioni vere), di contro non si attagliano al linguaggio naturale. Questo palese iato appare non appena ci volgiamo a considerare la deliberazione umana. È proprio a quest'ultima che si rifanno i nuovi algoritmi decisionali basati su modello.

3.2. *Algoritmi basati su modello e deliberazione umana*

Gli agenti basati su modello, infatti, prima di scegliere (e quindi di agire) devono «prendere in considerazione il futuro sotto due aspetti: "cosa accadrà se faccio così e cosà?" e anche "se faccio questo sarò soddisfatto?". La razionalità si esplicherà ulteriormente non già nell'avere obiettivi ma nel massimizzare l'utilità per raggiungere quel determinato obiettivo». Alla base

[18] T. Hobbes, *Il Leviatano* (1651), tr. it. Utet, Torino 1955, pp. 75-76.

del processo decisionale di un'AI vi è quindi la teoria dell'utilità e quella della probabilità, in base alle quali «un agente […] può scegliere razionalmente l'azione da intraprendere in base a ciò che crede e desidera»[19]. La combinazione delle due teorie permette all'agente di muoversi nel momento in cui le inferenze della logica del primo ordine (un agente puramente logico) non bastano. Infatti, quando l'agente artificiale si trova a dover agire senza che vi sia un obiettivo certamente utile, e quindi si trova in una condizione di incertezza se fare *a*, *non a* o *b*, allora viene introdotta la cosiddetta *funzione di utilità* che assegna un singolo numero per esprimere la desiderabilità di uno stato (in altre parole: viene assegnato un valore di utilità a ciascuna delle azioni che possono essere compiute). In tal caso l'agente realizzerà il principio della massima utilità attesa, ovvero sceglierà quell'azione che massimizza l'utilità attesa dell'agente. In questo *modus operandi*, secondo gli autori del noto manuale, può risiedere l'intelligenza, dal momento che «tutto ciò che un agente intelligente deve fare è calcolare le varie quantità, massimizzare l'utilità sulle sue azioni ed ecco tutto»[20].

Nelle parole di Russell e Norvig, però, si nota un ritorno a quel presupposto che essi stessi avevano bandito: il pensiero non è riducibile al calcolo. Per il padre del pragmatismo Charles Sender Peirce, ad esempio, «pensare non è calcolare. Si tratta piuttosto di compiere un balzo, formulando congetture»[21]; o, per dirla con un'immagine: se il calcolo porta a congiungere i puntini noti – come potrebbe avvenire in uno di quei giochini sulla *Settimana enigmistica*, l'analisi, invece, conduce a dare un senso a quei puntini[22].

[19] S. Russell, P. Norvig, *op. cit.*, p. 539.
[20] *Ibid.*, p. 540.
[21] Qui in E.J. Larson, *The Myth of Artificial Intelligence*, Harvard University Press, Cambridge/MA 2022, p. 97.
[22] Cfr. *Ibid.*, p. 96.

Cito, per far un esempio cogente, una pagina straordinaria di Edgar Allan Poe:

> Le facoltà mentali che si sogliono chiamare analitiche sono, di per se stesse, poco suscettibili di analisi. Le conosciamo soltanto negli effetti. […] Come l'uomo forte gode della sua potenza fisica e si compiace degli esercizi che mettono in azione i suoi muscoli, così l'analista si gloria di quella attività spirituale che serve a «risolvere». E trova piacere anche nelle occupazioni più comuni purché diano gioco al suo talento. […] E i risultati, abilmente dedotti dalla stessa essenza e anima del suo metodo, hanno veramente tutta l'aria dell'intuito. […] Il massimo potere della riflessione è più decisamente e utilmente provato dal modesto gioco della dama che non dalla complicata futilità degli scacchi. In quest'ultimo essendo i pezzi dotati di movimenti diversi e bizzarri e di valori diversi e variabili, quello che è soltanto complessità vien preso (errore abbastanza comune) per profondità[23].

Si badi però che le AI, a volte, possono "prendere decisioni" in modo del tutto randomico (si pensi all'aspirapolvere automatico o al GPS).

Una volta compreso, quindi, seppur per grandi linee, come decidono le AI, non ci rimane che comprendere cosa avviene qualora esse si sostituiscano a noi nell'esercizio della decisione o qualora esse interagiscano con noi in tale processo e per far questo indicherò, rispettivamente, questi possibili paradigmi con due neologismi: il paradigma *sostituivista* e quello *complementarista*.

4. *Il paradigma della sostituibilità*

Per comprendere il primo paradigma che qui suggerisco mi rifaccio alle parole del professor Paolo Moro, per il quale la domanda che oggi dovremmo porci consiste nel chiederci se un robot «dotato di intelligenza e di volontà artificiale del tutto analoghe, se non superiori, a quelle umane, possa essere titolare di libertà, diventando un centro di autodeterminazione e d'imputazione soggettiva di comportamenti interattivi»[24]. Perché questo scenario si possa realizzare non è necessario raggiungere la cosiddetta *Artificiale General Intelligence*, basterebbe richiamare alla

[23] E.A. Poe, *Racconti*, traduzione di D. Cinelli e E. Vittorini, Mondadori, Milano 1961, pp. 4-5.

[24] P. Moro, *Libertà del robot? Sull'etica delle macchine intelligenti*, in *Filosofia del diritto e nuove tecnologie. Prospettive di ricerca tra teoria e pratica*, a cura di R. Brighi e S. Zullo, Aracne, Roma 2015, p. 526.

memoria alcuni casi ormai famosi in cui l'intelligenza artificiale ha scelto in base a dei dati in ingresso da un lato, e a un autoapprendimento (del quale essa stessa è capace) dall'altro, senza che si rendesse necessario il concorso dell'essere umano. Cito alcuni casi famosi.

Nel 2013 il tribunale del Wisconsin (episodio sul quale avrò modo di soffermarmi a breve) ha negato la cauzione all'afroamericano Eric Loomis sulla base di un report dell'algoritmo COMPAS; nel 2014 le decisioni prese dall'algoritmo in uso presso il settore *human resources* di Amazon si sono dimostrate palesemente influenzate da *bias* di genere nell'attribuzione di cariche apicali nel settore ingegneristico dell'azienda; nel 2019 l'algoritmo *Apple Credit Card* da poco brevettato dalla nota casa americana riconsegnava significative differenze nell'attribuzione del massimale di credito tra uomini e donne; nel 2021, infine, il tribunale di Bologna ha sancito la condanna dell'algoritmo di Glovo dal momento che discriminava le *performance* dei *rider* basandosi su fattori che non dipendevano dai *rider* stessi, attribuendo, di fatto, responsabilità decisionale all'algoritmo.

Questi sono tutti esempi in cui le AI, in modo del tutto autonomo, hanno operato delle decisioni, condizionate però da palesi *bias* cognitivi. Non sarebbe possibile addentrarmi in questa sede sulle ripercussioni che l'utilizzo generalizzato di questi algoritmi comporterebbe (e sui più recenti sviluppi dell'*algorithm fairness* o dell'*explainable artificial intelligence*); mi preme far notare, tuttavia, come questo primo paradigma richieda imprescindibili approfondimenti dal momento che le decisioni prese dalla AI hanno un impatto sulla vita dei singoli e della società. Adduco soltanto un esempio per segnalare un'indifferibile riflessione etica in tale ambito.

Si pensi alle conseguenze provocate da un algoritmo che sulla base di alcuni *bias* non concedesse il mutuo ad alcune famiglie in necessità. In tanti potrebbero obiettare che un tale errore potrebbe

essere compiuto anche da un consulente di banca; la differenza, tuttavia, risiede nel fatto che se a un a consulente "umano" è possibile analizzare dalle tre alle dieci richieste in una giornata, un'AI, in pochi minuti, è in grado di effettuare un processo decisionale – sulla base di milioni di dati posseduti – riconsegnando un esito negativo a milioni di famiglie, che verrebbero gettate nella totale disperazione.

La soluzione più immediata, allora, anche in considerazione del fatto che siamo ancora lontani da un'intelligenza artificiale in grado di prendere decisioni cogenti, sarebbe quella di, cito il sito di IBM: «creare e offrire […] una tecnologia affidabile, in grado di aumentare (non sostituire) il processo decisionale manuale»[25]. Lo scenario prospettato è similare a quello che già si attua all'interno degli aeromobili, i quali potrebbero benissimo essere guidati esclusivamente dai sofisticatissimi computer di bordo installati ma, nondimeno, sono presenti due ufficiali in cabina a controllare[26].

Dato questo primo scenario, allora, è possibile abbandonare il paradigma della sostituibilità e rivolgersi al secondo modello che qui suggerisco come occasione di riflessione: quello della complementarità.

5. *La complementarità tra umani e macchine nel processo decisionale*

A questo paradigma è possibile ricondurre quella che è stata definita come *replacement technology*, ovvero quella tecnologia che *sostituisce* le fasi più pesanti o più tecniche di un lavoro (analisi della giurisprudenza, udienze *online*, piattaforme per la risoluzione delle controversie, soltanto per citare alcune

[25] *XAI (eXplainable AI - AI spiegabile)*, in https://www.ibm.com/it-it/watson/explainable-ai (ultimo accesso: 3-3-2023).

[26] Sui recenti tentativi di modificazione della prassi a bordo degli aerei, che fanno leva proprio sull'efficienza dei sistemi di AI, cfr. L. Berberi, *Aerei, le compagnie chiedono un solo pilota a bordo per tagliare i costi*, in https://www.corriere.it/economia/aziende/22_novembre_25/aerei-compagnie-chiedono-solo-pilota-bordo-tagliare-costi-fe5571d4-6ca0-11ed-a41d-76ead3b90d6e.shtml (ultimo accesso: 3-3-2023).

applicazioni in ambito legale). Questi strumenti si collocano "a metà strada": essendo né semplici supporti né sostitutivi dell'impegno umano ma, appunto, complementari.

Nel processo decisionale, nello specifico, essi «sostituiscono almeno un segmento del processo decisionale»[27] compiendo quella raccolta di dati previa alla decisione. In questo scenario «per sfruttare l'incredibile potenziale dell'Intelligenza Artificiale occorre una sapiente collaborazione tra persone e macchine, in cui – nella gran parte dei casi – *saranno sempre e comunque gli esseri umani ad avere l'ultima parola*»[28].

Questo approccio, che si rifà alla cosiddetta *AI debole* o *Narrow AI*, non sarebbe poi così preoccupante dal momento che esisterebbe sempre la presenza di un supervisore umano. Osservata da questa prospettiva la macchina non sarebbe nient'altro che una nuova e sofisticatissima leva d'Archimede alla quale l'essere umano demanda la parte più pesante e svilente del lavoro (al pari di un montacarichi o di uno strumento di catalogazione di testi), avocando a sé il controllo o comunque la parte più delicata del processo. Assumendo questo approccio *complementarista*, pertanto, molti errori potrebbero essere evitati. Per tornare agli esempi precedentemente addotti, infatti: il giudice si baserebbe sul *report* dell'algoritmo ma non ne sarebbe determinato; il *recruiter* potrebbe prendere in mano il *report* maschilista dell'algoritmo e reintrodurre comunque le quote rosa; un capo attento potrebbe valutare di caso in caso le *performance* dei *rider*, e così via.

Per rappresentarci meglio il paradigma *complementarista* nella relazione essere umano-AI, all'interno del processo decisionale, recupero le svariate applicazioni di geolocalizzazione già citate, e delle quali si fa ormai uso generalizzato attraverso gli

[27] E. Calzolaio, *Intelligenza artificiale ed autonomia della decisione: problemi e sfide*, *La decisione nel prisma dell'intelligenza artificiale*, a cura di E. Calzolaio, cit., p. 2.

[28] S. Maggioni, *AI & ETHICS – La sottile linea tra decisioni giuste e decisioni*, in https://www.sas.com/it_it/news/leading-art-innovation/whats-hot/ai-ethics-decisioni.html (ultimo accesso: 3-3-2023) [corsivo mio].

smartphone. *Google Maps*, *Maps* della Apple o *Waze* – solo per citare i più noti – sono tutti algoritmi che, nel momento in cui selezionano la strada per farci raggiungere la nostra meta tra decine o centinai di percorsi possibili, "decidono per noi". Proprio questo esempio così pregnante permette di far emergere quello che è l'approccio maggioritario nell'utilizzo dell'intelligenza artificiale: gli algoritmi sono degli *aiuti* all'azione umana.

Ma è proprio nella complementarità tra essere umano e algoritmo che si annida la vera sfida che, secondo me, richiede di essere tematizzata più di molte altre che si rifanno in certo qual modo alla cosiddetta algoretica: nel momento in cui la complementarietà tra essere umano e macchina si sposta dal piano della forza motrice a quello della causalità formale, è ancora possibile parlare in termini di complementarietà? Cerco di spiegarlo più chiaramente.

Se per quanto riguarda l'attività pratico-manuale (come il sollevare dei blocchi di cemento o il catalogare dei testi) è abbastanza semplice affidarsi a uno strumento meccanico o elettronico che compia il nostro stesso lavoro in minor tempo (e molto probabilmente con più accuratezza), si può dire lo stesso quando quello strumento entra a pieno titolo nel nostro processo decisionale?

Per comprendere la mia posizione è utile richiamare cosa è necessario perché un essere umano pervenga a una decisione deliberata, e quindi libera, e della quale pertanto può essere ritenuto responsabile. Risulta necessario che egli sia in grado di:

1. comprenderne le conseguenze attraverso un ragionamento controfattuale (requisito dell'*agency*),
2. raffigurarsi i diversi corsi di azioni disponibili (requisito delle *alternative possibilities*),
3. assumersi la piena responsabilità di quell'azione, dal momento che essa discende dalla sua volontà e non dalla volontà altrui, come nel caso di una scelta coartata (requisito del *controllo*).

I giuristi sanno bene come qualora mancasse uno di questi tre elementi l'azione non sarebbe pienamente imputabile al soggetto:

come nel caso di demenza, o ignoranza, o qualora ci si trovasse durante stati attenuati di coscienza. In contesti usuali – mancando particolari patologie neurologiche o stati di coscienza alterata – gli esseri agenti hanno tutti e tre i requisiti e sulla base di essi possono operare delle scelte[29]. A quanti, tuttavia, non è mai capitato di compiere una scelta, a seguito di lunga ponderazione, e poi pentirsene perché si era fatta una valutazione dei rischi e delle conseguenze non pienamente in linea e non pienamente soddisfacente? Questo scenario, se è all'ordine del giorno per gli esseri umani, non lo è per le intelligenze artificiali che hanno raggiunto una capacità di previsione strabiliante, almeno in alcuni settori specifici. Per far comprendere ancora meglio il suggerimento che qui invito ad assumere, riporto un caso che saltò subito agli onori della cronaca.

6. *Interazione 'dentro' la decisione*

Soffermarsi in questa sede sul processo decisionale degli esseri umani sarebbe impossibile. Rifacendomi, tuttavia, a due casi che attengono alla giurisprudenza, cercherò di addurre alcune riflessioni conclusive.

Nel 2013 i giornali americani portarono alla ribalta il caso di Eric Loomis, condannato da un giudice di un tribunale locale del Wisconsin, il quale – per l'emissione del suo verdetto – aveva fatto ricorso a un rapporto, contenente la storia dell'indagato (precedenti penali, contesto familiare e sociale, ecc.), all'interno del quale era presente anche una valutazione operata dall'algoritmo COMPAS[30] che catalogava l'imputato afroamericano come

[29] Per un'approfondita discussione sul tema, mi sia permesso rimandare a C. Calì, *Il libero arbitrio in questione. Una ricerca tra filosofia, scienze e intelligenza artificiale*, Mimesis, Milano 2023, in particolare il cap. 1.

[30] Sebbene l'algoritmo di COMPAS sia rimasto segreto, in quanto prodotto da una società privata e tutelato dalle leggi sul diritto d'impresa, è possibile dire su quali dati esso si basasse per giungere al proprio verdetto: precedenti penali, utilizzo pregresso di sostanze stupefacenti, e un insieme di altri dati che lo stesso indagato forniva all'algoritmo mediante un test di 137 domande, volte a indagare la personalità e la storia dell'individuo. Questi dati erano quindi combinati con profili e circostanze similari. Per un approfondimento sulla

ad "alto rischio di recidiva". Nonostante i due gradi di giudizio ai quali fece ricorso Loomis, nel 2016 la *Corte suprema del Wisconsin* rigettava il ricorso dell'imputato e definiva che il verdetto del primo tribunale non poteva essere considerato invalido o lesivo del diritto al giusto processo perché il *report* dell'algoritmo aveva costituito soltanto un "ausilio" al giudice *umano,* ma non era stato in alcun modo decisivo ai fini della sentenza definitiva[31].

Nel momento in cui faccio emergere le criticità di quello che ho definito come *approccio complementarista* è doveroso, tuttavia, far notare come al giorno d'oggi, epoca in cui si parla del dominio dei dati, risulti del tutto impossibile a un agente umano gestire una quantità tale di dati in un tempo conveniente. L'analisi dei dati a nostra disposizione, infatti, costituisce – come già ho avuto modo di accennare – un momento significativo del processo decisionale. Ora è proprio quest'ultimo processo che sembra non potersi più condurre senza l'ausilio di algoritmi specifici. Una tale situazione richiede pertanto una seria riflessione se non, addirittura, un cambiamento del nostro modo di intendere e di gestire il processo decisionale[32].

Tornando al caso di Eric Loomis, non mi è possibile entrare qui nella miriade di complessità che questa sentenza porta con sé. È doveroso sottolineare però che nel maggio 2016 (prima della condanna definitiva) in un'inchiesta condotta da *ProPublica* si asseriva, sulla base di ricerche statistiche, che l'algoritmo COMPAS

vicenda e i suoi risvolti etici cfr. C. Calì., *L'imparzialità del giudicante. Alcune implicazioni etiche dell'utilizzo dell'intelligenza artificiale in giurisprudenza*, in *La pubblica amministrazione del futuro. Tra sfide e opportunità per l'innovazione del settore pubblico*, a cura di A. Alù, A. Ciccarello, Editoriale Scientifica, Napoli 2021, pp. 121-134

[31] Cfr. *State of Wisconsin v. Loomis*, cit., 4. 7-8.

[32] Per la discussione se le cosiddette *alternative possibilities* amplino o riducano la nostra capacità decisionale, mi sia permesso rimandare a C. Calì, *Algoritmi e processo decisionale. Alle origini della riflessione etico-pratica per le IA*, in «Scienza&Filosofia», 27, 2022, pp. 69-87 e Id., *Il libero arbitrio in questione. Una ricerca tra filosofia, scienze, intelligenza artificiale*, cit., pp. 53-65.

presentava forti pregiudizi nell'attribuire agli afroamericani un tasso di rischio sempre più elevato (*bias* razziale confermato da un'ulteriore inchiesta sul medesimo algoritmo condotta dal Dartmouth College)[33].

Si vede quindi come i giudici elettronici non siano perfetti. Non sembra migliore la situazione con i giudici umani.

Nel 2011 lo studioso israeliano Shai Danziger, insieme ad alcuni colleghi, ha deciso di valutare come i giudici pervenissero a un determinato verdetto: se si limitassero ad applicare argomenti giuridici ai fatti del caso concreto in modo razionale, meccanico e deliberativo (formalismo giuridico) o se il processo deliberativo fosse influenzato da una serie di fattori del tutto esterni alla legge ma interni alla mente del giudicante (realismo giuridico). Attraverso uno studio sperimentale – i cui risultati sono confluiti in un articolo pubblicato sulla rivista scientifica *PNAS* della National Academy of Sciences statunitense – sono state osservate le decisioni dei giudici nell'arco temporale di una giornata lavorativa (in cui i giudici fanno due *break*). L'articolo ha rilevato che:

> La percentuale di decisioni favorevoli al reo scendeva gradualmente dal 65% a quasi zero all'interno di ciascun segmento della giornata [in misura

[33] «Was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants. White defendants were mislabeled as low risk more often than black defendants» J. Angwin, J. Larson, S. Mattu, L. Kirchner, *Machine Bias. There's Software Used Across the Country to Predict Future Criminals. And It's Biased against Blacks*, ProPublica, 23 maggio 2016 www.propublica.org. All'inchiesta è seguito un acceso dibattito, al quale hanno partecipato sia la società produttrice di COMPAS con uno studio [cfr. W. Dieterich, C. Mendoza, T. Brennan, *COMPAS Risks Scales: Demonstrating Accuracy Equity and Predictive Parity*, 2016, go.volarisgroup.com] sia un *team* di ricercatori [cfr. A. Flores, K. Bechtel, C. Lowenkamp, *False Positives, False Negatives, and False Analyses: A Rejoinder to 'Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks'*, in «Federal probation», 80, 2016, pp. 38-46; J. Jung, C. Concannon, R. Shroff, S. Goel, D.G. Goldstein, *Simple Rules for Complex Decisions*", in «arXiv», 1702, 2017, p. 04690]. A conclusioni simili a quelle edite da *ProPublica* nel 2016 sono giunti invece, nel 2018, due ricercatori del Dartmouth college, Julia Dressel e Hany Farid, i quali hanno condotto uno studio per mostrare che nel valutare la potenziale recidività di un individuo, COMPAS non è più affidabile di un gruppo di volontari scelti a caso su internet. Cfr. J. Dressel, H. Farid, *The Accuracy, Fairness and Limits of Predicting Recidivism*, in «Science Advances», IV, 2018, p. eaao5580.

> direttamente proporzionale a quanto ci si allontanava dalla colazione del giudice]; subito dopo una pausa, la percentuale risaliva bruscamente a 65%[34].

In altre parole, il verdetto dipendeva «da quello che il giudice aveva mangiato a colazione»[35].

L'apporto di questi due casi in un contesto così delicato com'è quello del giudizio in tribunale, ci consente di comprendere come la vera sfida per la ricerca scientifica e per i decisori politici non sia tanto quella di normare le "macchine intelligenti" o di fornire strumenti etici ai programmatori delle stesse, quanto piuttosto quella di formare le coscienze affinché forse realizzando il sogno spinoziano in merito alla libertà – si prenda atto di alcuni dei meccanismi messi in atto dagli algoritmi.

Un tale compito è legato al fatto, già concreto e non futuro, che l'intromissione delle AI nei processi decisionali cogenti sta già operando un ri-ontologizzazione non del mondo attorno a noi ma di quella che sinora è stata la dimensione più nostra: la capacità cognitiva di prendere azioni libere informate dalla ragione. Questa a mio modo di vedere è la sfida per quanti a diverso titolo progettano gli algoritmi di AI – ingegneri, programmatori, filosofi – ma, ancor di più, per tutti noi che ogni giorno ci avvaliamo di alcuni sistemi – dai *cookies* del *browser* alle *newsletter* per la profilazione del cliente ai navigatori satellitari – per compiere le nostre azioni e informare le nostre decisioni, dalla più banale alla più significativa.

Cristiano Calì è Assegnista di Ricerca in Filosofia Morale presso l'Università di Torino e Docente a Contratto di Filosofia della Medicina presso L'Università di Roma "La Sapienza"

cristiano.cali@unito.it

[34] S. Danziger, J. Levav, L. Avnaim-Pessoa, *Extraneous factors in judicial decisions*, in «PNAS Proceedings of the National Academy of Sciences of the United States of America», 108, 2011, pp. 6889–6892 [traduzione mia].

[35] «What the judge ate for breakfast», *Ibid.*, p. 6889.