CRISTIANO CALÌ

ALGORITMI E PROCESSO DECISIONALE.
ALLE ORIGINI DELLA RIFLESSIONE ETICO-PRATICA PER LE IA

ABSTRACT: *ALGORYTHMS AND DECISIONAL PROCESS. AT THE ORIGINS OF ETHICAL-PRACTICAL REFLECTION FOR AI*
This contribution aims to investigate not so much the ethical implications of utilizing intelligent machines in specific contexts, (human resources, self-driving cars, robotic hospital assistants, et cetera), but the premises of their widespread use. In particular, it aims to analyze the complex concept of decision making - the cornerstone of any ethical argument - from a dual point of view: decision making assigned to machines and decision making enacted by machines. Analyzing the role of algorithms in decision making, we suggest a redefinition of the ethical vocabulary for AI, specifically regarding the concept of agency.

1. *Introduzione*

Nel 2018 usciva la prima monografia in lingua italiana che trattava, in modo quasi esclusivo, del problema, abbastanza recente, dell'etica delle macchine intelligenti[1], preceduto di diversi anni dal suo corrispettivo in lingua inglese[2]. Non erano di certo mancati articoli che avevano avviato la discussione, ma soltanto negli ultimi cinque anni, a seguito della rinnovata "estate" che l'intelligenza artificiale (IA) sta vivendo, vi è stata un'innumerevole produzione di testi legati

[1] Cfr. P. Benanti, *Le macchine sapienti. Intelligenze artificiali e decisioni umane*, Marietti, Bologna 2018.
[2] Cfr. M. Anderson, S. Leigh Anderson, *Machine ethics*, Cambridge University Press, New York 2011.

all'impatto etico delle nuove tecnologie[3]. Il catalizzarsi dell'attenzione verso quest'ultime non è stato operato esclusivamente dal mondo accademico, ma anche da quello politico. Dal 2016, infatti, «quando i governi nazionali, le organizzazioni non governative e le aziende private hanno iniziato ad assumere un ruolo di primo piano nel dibattito sull'IA e sugli algoritmi "equi" ed "etici"», i lavori sull'etica degli algoritmi sono aumentati in modo significativo[4] e a partire dal 2017, quando furono definiti alcuni principi nel documento di Asilomar, le pubblicazioni intorno all'etica delle IA e i documenti da parte di governi e commissioni internazionali si sono moltiplicati a dismisura[5].

Se si dà un rapido sguardo alle prime, le pubblicazioni, gli argomenti trattati sono quasi sempre gli stessi: macchine a guida autonoma e problemi correlati, armi autonome e politiche che le normino, IA e mondo del lavoro, cybersicurezza e privacy, impatto ambientale dell'IA e sex robot, diritto *delle* IA e IA *nel* diritto[6]. Indubbiamente questi sono i temi più cogenti e anche i

[3] L. Floridi, *Etica dell'intelligenza artificiale. Sviluppi, opportunità, sfide*, tr. it. Raffaello Cortina, Milano 2022, p. 147.

[4] Cfr. C. Sandvig, K. Hamilton, K. Karahalios, C. Langbort, *When the algorithm itself is a racist: Diagnosing ethical harm in the basic components of software*, in «Internationl Journal of Communication», 10, 2016, pp. 4972-4990; R. Binns, *Fairness in machine learning. Lessons from political philosophy*, ArXiv 2018, 1712.03586 [cs]; A.D. Selbst, D. Boyd, S.A. Friedler, S. Venkatasubramanian, J. Vertesi, *Fairness and abstraction in sociotechnical systems*, in «Proceedings of the Conference on Fairness, Accountability, and Transparency-fat*'19», ACM Press, Atlanta/GA, pp. 59-68; R. Ochigame, Rodrigo, The Invention of "Ethical AI", in https://theintercept.com/2019/12/20/mit-ethical-ai-artificial-intelligence/ (13/03/2022).

[5] Oltre ai 23 principi frutto del convegno ad Asilomar in California, si pensi anche alla *Partnership on AI*, finalizzata allo sviluppo di *best practices*, e al progetto l'*AI for Good*, con il coinvolgimento dell'ONU. Per una rassegna commentata cfr. Floridi, *op. cit.*, pp. 94-105.

[6] Cfr. N. Turi, *Guida per umani all'intelligenza artificiale. Noi al centro di un mondo nuovo*, Giunti, Milano 2019; G. Tamburrini, *Etica delle macchine. Dilemmi morali per robotica e intelligenza artificiale*, Carocci, Roma 2020; A. Longo, G. Scorza, *Intelligenza artificiale. L'impatto sulle nostre vite, diritti e libertà*, Mondadori università, Milano 2020; M. Chiriatti, *Incoscienza artificiale. Come fanno le macchine a prevedere per noi*, Luiss University Press, Roma 2021; F. Fossa, V. Schiaffonati, G. Tamburrini (a cura di), *Automi e persone. Introduzione all'etica dell'intelligenza artificiale e della robotica*, Carocci, Roma 2021; F. Pasquale, *Le nuove leggi della robotica. Difendere la competenza umana nell'era dell'intelligenza artificiale*, Luiss

più attraenti. Pensare a un veicolo a guida autonoma (e a tutti i possibili scenari in cui questo potrebbe ritrovarsi coinvolto) rimanda ai classici dilemmi etici come quello del *trolley problem*, e a concetti tipici della riflessione morale, primo fra tutti quello di responsabilità morale. Prendere le mosse da questo tipo di IA è altresì giustificato dal fatto che, al giorno d'oggi, le macchine a guida autonoma, sperimentate per prime da Google e poi da Waymo (ai quali sono seguito Tesla, Uber, Toyota e Nvidia), sono «un vero concentrato di IA»[7] perché esse devono avere

> la capacità di "vedere" la strada, i segnali, gli eventuali pedoni e gli ostacoli, perciò hanno bisogno di tecniche di interpretazione delle immagini. In più, devono prevedere le azioni dei pedoni e delle altre auto per evitare gli incidenti. Per questo hanno bisogno di ragionare statisticamente e gestire l'incertezza delle possibili situazioni che possono capitare tramite tecniche di IA come la pianificazione e la ricerca ottimizzata[8].

Se ci si rivolge quindi ai documenti, si notano invece i principi ispiratori dell'etica delle IA e quelli che debbono guidare la realizzazione degli algoritmi: trasparenza, inclusione, responsabilità, imparzialità, sicurezza e privacy, affidabilità[9]. Da questa velocissima disamina si nota come la riflessione si sia focalizzata in modo preponderante sul cosiddetto campo delle etiche applicate eludendo – o semplicemente dando per presupposta – una riflessione che si collochi piuttosto alla base del discorso etico, che ne indaghi la sua possibilità prima ancora della sua applicabilità. Senza volere (né potere) entrare qui nel dibattito tra IA debole (o leggera) o IA forte (o sarebbe meglio dire AGI), bisogna notare, infatti, che nel momento in cui parliamo di etica delle macchine si assume tacitamente (anche da parte di quegli autori che a buon diritto sono scettici se non del tutto

---

University Press, Roma 2021; P. Severino (a cura di), *Intelligenza artificiale. Politica, economia, diritto, tecnologia*, Luiss University Press, Roma 2022.

[7] F. Rossi, *Il confine del futuro. Possiamo fidarci dell'intelligenza artificiale*, Feltrinelli 2019, p. 22.

[8] *Ibid*.

[9] Cfr., *Rome Call for AI Ethics*, in https://www.academyforlife.va/content/dam/pav/documenti%20pdf/2020/CALL%2028%20febbraio/AI%20Rome%20Call%20x%20firma_DEF_DEF_con%20firme_.pdf (ultimo accesso 12/3/2022).

negazionisti della IA forte[10]) che le macchine siano come gli esseri umani. Questo dato traspare già dal vocabolario al quale si ricorre per parlare delle macchine e delle loro operazioni.

Si parla di «*autonomia* delle macchine»[11], la quale deve «essere limitata e resa intrinsecamente reversibile, qualora l'autonomia umana debba essere protetta o ristabilita»[12]; si fa riferimento alle «*capacità* di *operare* in maniera intelligente»[13]; si parla di vero e proprio *agire*, non di un suo surrogato, al punto tale che per Floridi l'IA deve essere intesa primariamente come forma di un «agire artificiale»[14] più che di un pensare artificiale.

Anche guardando a una delle possibili (seppur abbastanza generica) definizioni di robot, si nota come essi vengano definiti in quanto «dispositivi hardware supportati da un software che permette loro di funzionare e *agire* secondo determinati *scopi* nel mondo fisico reale»[15]. L'attribuzione di elementi tipici dell'agire umano alle macchine è pertanto prassi comune e non si limita al puro ambito lessicale. Piuttosto, quest'operazione lessicale è susseguente alle applicazioni in ambito pratico in cui prodotti sofisticati come, ad esempio, Shakey, uno dei primi robot mobili, è «in grado di *pianificare* da solo le sue azioni»[16].

Se è vero, allora, che ci troviamo dinanzi a un uso metaforico del linguaggio, è altrettanto vero – e non è un fattore da sottovalutare – che ci troviamo sprovvisti di un dizionario adeguato ad affrontare questi argomenti. Questo dato richiede quindi di essere assunto come sintomo di una patologia da indagare: quando attribuiamo alle IA nozioni come responsabilità,

[10] Penso sia opportuno rammentare che, sebbene il progetto dell'IA forte non sia l'orientamento maggioritario, a esso lavorano le big dell'automazione come DeepMind.
[11] L. Floridi, *op. cit.*, p. 99, corsivo mio.
[12] *Ibid.*
[13] M. Ienca, *Intelligenza. Per un'unione di intelligenza naturale e artificiale*[2], prefazione di A. Scoccia Pappagallo, Rosenberg & Sellier, Torino 2019, p. 13, corsivo mio.
[14] L. Floridi, *op. cit.*, p. 12.
[15] F. Rossi, *op. cit.*, p. 40, corsivo mio.
[16] *Ibid.*, p. 39.

libertà, autonomia, capacità (e volutamente per questo discorso non chiamo in causa coscienza e intelligenza), sinora appannaggio esclusivo degli agenti umani, dobbiamo chiederci se esse mantengano ancora il loro classico significato, ovvero siano una semplice trasposizione dal piano naturale a quello artificiale dei medesimi concetti, o se in questo *slittamento contestuale* vengano risignificate e, conseguentemente, aiutino a ricomprendere l'umano.

Questo contributo vuole dedicarsi all'analisi di una di queste nozioni, anzi di quella nozione che può essere individuata come minimo comune multiplo di tutte le altre: la nozione di decisione e/o scelta (in questo contesto la distinzione tra le due perde di significato in quanto la seconda sarebbe l'atto conclusivo, il momento volitivo, al quale conduce la prima). Quando, infatti, intavoliamo un qualsiasi discorso etico, non facciamo altro che indicare i principi normativi dai quali una scelta dovrebbe essere diretta o forniamo un giudizio di valore sulla scelta compiuta. Cercherò quindi di circoscrivere alcuni caratteri peculiari della scelta per notare *come* e *se* le macchine abbiano influito dall'esterno sul nostro processo interno di deliberazione, per analizzare poi se quegli stessi elementi possano essere attribuiti anche alla "scelte" operate dalle IA. In altre parole, non andrò a indagare le applicazioni pratiche (e quindi etiche) della scelta nei suoi vari ambiti, ovvero gli ambiti nei quali le macchine possono dirsi responsabili e libere, tenterò piuttosto di individuare *se* esse possano essere "agenti capaci di compiere scelte".

Un tale discorso mi appare tutto fuorché inutile dal momento che, se è vero che molti pensano che gli algoritmi non potranno mai prendere delle decisioni importanti perché queste richiederebbero una dimensione etica (che gli algoritmi non possono né potranno mai comprendere), è altrettanto vero che

> non c'è ragione di presumere che gli algoritmi non saranno in grado di superare l'uomo medio anche nell'etica. Già oggi, mentre apparecchi

> come gli smartphone o i veicoli a guida autonoma prendono decisioni che un tempo erano monopolio dell'uomo, altri simili apparecchi iniziano ad affrontare problemi etici analoghi a quelli che tormentano gli uomini da millenni[17].

2. *La decisione demandata alle macchine*

Se parlare di *self driving cars* può sembrare ancora qualcosa di distopico – benché ormai non lo sia affatto – l'IA ritenuta oggigiorno in grado di prendere decisioni non si limita a quell'ambito di applicazione e, ben lungi dall'essere aliena dalle nostre vite, le permea ormai completamente. Oggi l'IA, infatti, non è perennemente presente soltanto nella vita della generazione Z e di quelle successive ma anche in quella delle generazioni precedenti. GPS, assistenti vocali, *software* per ogni tipo di lavoro, sistemi di identità digitale per accedere ai più basilari servizi al cittadino, sono ormai imprescindibili e fruibili da chi, anche superata la soglia degli ottant'anni, ha acquisito un minimo di dimestichezza digitale. L'IA è «entrata a far parte delle nostre abitudini e dei nostri gesti quotidiani, talvolta modificandoli profondamente»[18].

Questo ingresso, massiccio e non graduale, non ha trasformato soltanto le nostre abitudini ma ha cambiato anche il nostro cervello, o per essere più precisi, ha modificato le nostre capacità cognitive. Si pensi a come l'utilizzo costante del GPS (anche per muoversi nella propria città) abbia comportato – dati alla mano – una perdita notevole del senso di orientamento, inteso come la capacità di avere punti di riferimento con i quali orientarsi nello spazio[19]. Viviamo ormai in uno spazio *smart*, interattivo, ma l'interazione non è soltanto unidirezionale (da noi al *social*, o ad Alexa o a Siri, ecc.) ma bidirezionale: noi

[17] Y.N. Harari, *21 lezioni per il XXI secolo* (2005), tr. it. Bompiani, Milano 2018, p. 68.
[18] F. Rossi, *op. cit.*, p. 26.
[19] Per un'introduzione all'argomento, cfr. *Il GPS ha cambiato il nostro mondo e sta cambiando la nostra testa!*, in https://www.solotablet.it/lifestyle/il-gps-ha-cambiato-il-nostro-mondo-e-sta-cambiando-la-nostra-testa.

usiamo i *social* e loro "usano" noi; noi adoperiamo Siri e lei (o lui, a seconda della voce che abbiamo definito sui nostri dispositivi Apple) "adopera" noi, per apprendere ad esempio, ma non solo. Un primo dato che è necessario far emergere, allora, risiede nel cogliere come la nostra decisione di agenti razionali sia stata modificata dai sistemi di IA. Quando adottiamo l'IA e il suo agire *smart*, infatti,

> cediamo volontariamente parte del nostro potere decisionale ad artefatti tecnologici. Per questo, affermare il principio di autonomia nel contesto dell'IA significa trovare un equilibrio tra il potere decisionale che ci riserviamo e quello che deleghiamo agli agenti artificiali. Il rischio è che la crescita dell'autonomia artificiale possa minare il fiorire dell'autonomia umana[20].

L'impatto dell'intelligenza artificiale per le nostre vite è stato diffusamente e magistralmente descritto da Luciano Floridi, che non si limita a vedere nei progressi della *computer science* una nuova rivoluzione industriale né una rivoluzione epistemologica (sulla base della rivoluzione copernicana di Kant) ma una vera e propria rivoluzione ontologica dal momento che «il digitale ri-ontologizza il reale»[21]. Assunta questa prospettiva, allora, è di cruciale importanza comprendere cosa questa riontoligizzazione implichi per l'agire in generale e per l'agire morale in particolare. L'intelligenza artificiale, infatti, ha iscritto nel nostro universo «un nuovo agire non naturale che ha sempre maggior successo nelle sue interazioni con la realtà»[22]. L'impatto dell'IA sulle *nostre* decisioni, infatti, è già in atto.

A chi non è capitato di essere totalmente confuso quando si deve scegliere quale film guardare tra le migliaia di titoli offerti da Netflix, o quando bisogna selezionare una tra le centinaia di offerte di Glovo? È proprio in questi casi che qualcuno avrà gradito quella sezione delle nostre applicazioni che ha il titolo di *scelti per te*, una selezione *ragionata* in base ai contenuti

[20] L. Floridi, *op. cit.*, p. 98.
[21] Id., *Agere sine intelligere. L'intelligenza artificiale come nuova forma di agire e i suoi problemi etici*, in L. Floridi, F. Cabitza, *Intelligenza artificiale. L'uso delle nuove macchine*, Bompiani, Milano 2021, p. 66.
[22] *Ibid.*

precedentemente scelti (ma anche a quelli non scelti e che pertanto potrebbero destare il nostro interesse) volti a facilitare la nostra decisione. Se qualcuno non si è ancora trovato in queste situazioni paradigmatiche, basti pensare allora che quotidianamente i nostri *browser* di ricerca per il web decodificano i nostri interessi per selezionare i siti internet di maggior interesse per noi, cosicché se io studio filosofia in Italia alla digitazione delle lettere *cin* apparirà subito *cineca* e non *cinema*. Questo procedimento – si noti bene – non avviene nel momento in cui noi digitiamo nella barra di ricerca una qualunque parola (che il browser andrà poi a rintracciare tra miliardi di pagine web) ma avviene ancor prima che noi digitiamo le parole. Per spiegarlo meglio: la profilazione dei nostri interessi e la possibile nostra volontà di cucinare un tiramisù è calcolata dall'algoritmo molti mesi prima di quando ci vorremmo industriare nell'impresa culinaria, per esempio quando – mentre facevamo un giro per Roma – abbiamo digitato sul motore di ricerca del nostro smartphone 'migliori ristoranti a Trastevere'. Questa dinamica, che appare tanto consueta quanto comoda, modifica, anche se non ce ne rendiamo immediatamente conto, la nostra capacità di scelta, o sarebbe meglio dire di scelta *libera*.

È necessario allora fare una breve precisazione. Non sarebbe possibile neppure accennare al problema se una scelta possa dirsi libera o no, e in quale senso si debba intendere questo predicato[23]. Al fine della mia argomentazione si darà per scontato che la scelta sia libera nel senso comune che conferiamo a questo attributo, ovvero che appartiene all'agente, senza ulteriori specificazioni. Di fatto, l'oggetto di ogni riflessione etica sono sempre le scelte ritenute *libere*, e non quelle coartate o impedite.

[23] Per un primo approccio alla questione, cfr. S.F. Magni, *Teorie della libertà. La discussione contemporanea*, Carocci, Roma 2017, cap. 2.

Ora, perché si possa parlare di scelta libera è indispensabile che si abbia la possibilità – oltre che la capacità (altro distinguo necessario ma che qui può essere eluso senza che l'argomentazione ne abbia detrimento) – di scegliere tra due o più possibili opzioni. Perché possa operare una scelta tra i diversi contenuti offerti da Netflix è necessario che quei contenuti siano due o più, perché qualora avessi uno e un solo film a disposizione potrei sicuramente scegliere, ma senza che quella scelta possa ritenersi davvero genuina e libera. Nel momento in cui mi accomodo sul divano, allora, e Netflix seleziona per me 30 tra le migliaia di titoli, quale ruolo sta giocando l'algoritmo delle preferenze del colosso americano sulla mia scelta? sta forse riducendo la mia capacità di scelta, dal momento che limita le altre possibili opzioni? Stessa cosa si potrebbe dire del GPS: quando imposto un tragitto – nella mia stessa città – e il navigatore satellitare mi fornisce due o più percorsi possibili, ma io so che ve ne sono altri senza che il navigatore li indichi (a volte solo perché più lunghi di poche decine di metri anche se più lineari o più usuali per l'utente autoctono) sta forse riducendo la mia capacità di discernimento e di valutazione dei possibili corsi d'azione? In definitiva, si dovrebbe riconoscere che «la tecnocrazia invita a non scegliere, introducendoci in un orizzonte nel quale convivono "troppe" possibilità, che non siamo in grado di vagliare nemmeno volendolo, e in cui soltanto il sistema può stabilire cosa è meglio per noi»?

Reputo che la situazione non sia così drammatica e rammento una mia studentessa che – seguendo un ragionamento del tutto controintuitivo – mi disse "più scelte si hanno e meno si è liberi". Seguendo questa suggestione potremmo quindi dire che gli algoritmi designati a selezionare e discernere alcuni tra i molteplici corsi d'azione al posto nostro, piuttosto che essere identificati come un ostacolo alla nostra libertà di scelta si configurano come un *auxilium*, e non è forse questo che ricerchiamo

quando impostiamo il navigatore dal nostro smartphone? Questa dinamica, tuttavia, necessita di un ulteriore *focus*.

Nel momento in cui permettiamo all'algoritmo di *scegliere per noi*, seppur circoscrivendo la vasta gamma di opzioni percorribili, è utile chiederci cosa stiamo demandando, all'interno del processo deliberativo, alle IA. La deliberazione, infatti, caratterizzata dal soppesare diversi motivi per percorrere A o B, e quindi i *pro* e i *contro* sia di A che di B, è sempre stata un procedimento tipico del nostro apparato cerebrale, al punto tale che Thomas Hobbes riconosceva in questo procedimento la sede della libertà[24]. Oggi quel processo tipicamente umano viene demandato in parte alle IA, le quali, attraverso i loro algoritmi, selezionano per noi soltanto un *range* limitato di opzioni. Il vero problema è che le altre opzioni non vengono neanche proposte e quindi, di fatto, sono a noi del tutto precluse. Come si è avuto modo di notare, questo procedimento ci appare un aiuto all'esercizio della nostra libertà, in quanto ci aiuta a discernere solo su *concreti* e *possibili* corsi d'azione perché confacenti ai nostri interessi. Un aspetto, tuttavia, rimane però da essere rilevato; e per far ciò viene in aiuto uno dei *social* più nuovi che sta spopolando tra i *teenager*: TikTok. Il social, nato soltanto nel 2016, che si basa né su contenuti testuali (com'era in origine Facebook) né su foto (come il più recente Instagram) ma su video. A differenza dei sue social sopracitati, in cui sono gli utenti a individuare gli account o le pagine di proprio interesse che desiderano seguire (su Instagram o Facebook, ad esempio all'utente appaiono soltanto le *stories* e i contenuti dei profili che l'utente *ha scelto* di seguire), TikTok è pensato e progettato perché sia l'algoritmo a stabilire cosa tu debba vedere, in base ai tuoi interessi e ai video sui quali ti sei soffermato per più secondi. Anche in questo caso risulta particolarmente comoda una selezione dei contenuti

[24] Cfr. T. Hobbes, *Il corpo* (1655), in T. Hobbes, *Elementi di filosofia. Il corpo e l'uomo*, tr. it. Utet, Torino 1972, IV, 25, 13, p. 393.

secondo i nostri gusti, ma a cosa condurrebbe un mondo virtuale in cui siamo sottoposti solo ed esclusivamente ai contenuti in linea col nostro pensiero? Se siamo di destra vedremo solo contenuti del nostro orientamento politico, se amiamo i cani vedremo sempre e solo video con questi animali a quattro zampe e mai con felini, se siamo annoverabili tra i *novax* non ci verranno mai proposti contenuti dell'orientamento opposto. Osservato da questa prospettiva l'*auxilium artificialis* smette di avere un ruolo di *adiutorium*, per divenire un vero e proprio *impedimentum*. Se è vero che una scelta è sempre nostra perché discende dal soggetto, è altrettanto vero – benché alcune posizioni compatibiliste in merito alla libertà non concorderebbero – che quando siamo orientati, o sarebbe meglio dire influenzati solo e soltanto in una direzione, quella scelta da noi compiuta perde di efficacia e, sebbene non arriviamo a dire che essa sia stata costretta, potremmo dire che essa è stata sicuramente indotta.

Senza voler demonizzare, tuttavia, questa tipologia di algoritmi preposti alla "scelta per noi", (atteggiamento spesso portato avanti soprattutto da una certa saggistica che si concentra sugli scenari distopici) e tenendo conto che non ne potremmo fare a meno (sarebbe impossibile navigare sul web senza un algoritmo che ci orienti), rimane allora da capire *come* la decisione di selezione delle opzioni percorribili venga esercitata dall'algoritmo. Questa disamina si rende necessaria perché, mentre per gli agenti umani la valutazione dei possibili corsi d'azione è un momento previo alla scelta, per l'algoritmo la valutazione diventa vera e propria scelta, e quindi vero e proprio atto. Rimane quindi da capire in che modo l'IA prenda le sue decisioni.

3. *Le decisioni delle macchine*

È ormai famoso il caso di Eric Loomis, un cittadino statunitense che nel 2013 fu condannato dalla Corte suprema del Wisconsin, la quale – per emettere il verdetto – si basò su una valutazione del

rischio di recidiva formulata da un algoritmo: COMPAS, *Correctional Offender Management Profiling for Alternative Sanctions*. Quel caso divenne paradigmatico per le riflessioni sull'utilizzo in ambito giurisprudenziale delle IA, ma in questa sede assume rilevanza per un duplice ordine di fattori (che peraltro furono già fatti emergere dagli avvocati della difesa): l'algoritmo di COMPAS è a tutt'oggi segreto, in quanto prodotto da una società privata e tutelato dalle leggi sul diritto d'impresa, e quindi, benché si sappia che esso giunge a un verdetto combinando alcuni dati standard, non si ha conoscenza del procedimento esatto attraverso il quale viene emessa la valutazione dell'imputato[25]. Il secondo fattore rilevante è emerso a seguito di un'inchiesta condotta da *ProPublica* nel 2016, la quale, avendo condotto diverse ricerche statistiche, mostrò come l'algoritmo COMPAS presentasse forti pregiudizi nell'attribuire agli afroamericani un tasso di rischio sempre elevato[26].

[25] Per un'analisi particolareggiata del caso di Loomis, mi sia permesso rimandare a 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 e A. Ciccarello, Editoriale Scientifica, Napoli 2021, pp. 121-134.

[26] «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 et al., *Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks*, ProPublica, 23 maggio 2016, in www.propublica.org (8 gennaio 2021). 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*, in go.volarisgroup.com (6 gennaio 2021)] 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"*. Federal probation, 80 (2016) 2, pp. 38-46; J. Jung, et *al.*, *Simple rules for complex decisions*, in «arXiv», 1702 (2017), 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*, Science Advances, 4 (2018), eaao5580.

Dinanzi all'assenza di trasparenza del procedimento algoritmico che conduce all'azione sembrerebbe configurarsi la prima grande differenza col processo decisionale degli animali razionali. Il *libero accesso* al processo con cui si è addivenuti a un *output*, infatti, se è facile da ottenere con un'IA basata su un approccio simbolico, è pressoché impossibile da ottenere quando si adoperano IA basate sul *machine learning*. Questi sistemi, infatti, «non descrivono i passi necessari per risolvere un certo problema, ma sanno solo apprendere come risolverlo a partire dall'osservazione di una grande quantità di esempi»[27]. Si deve notare il fatto per cui mentre questo tipo di IA non sono in grado di riconsegnare il procedimento che ha condotto a un determinato esito, un agente è sempre capace – salvo casi di lesioni cerebrali o di ipnosi, o altri stati di coscienza attenuata – di dare conto della propria decisione e del perché sia addivenuto a tale verdetto, fosse anche che quest'ultimo suoni come "ho voluto perché ho voluto", "sono stato spinto dalla foga del momento" o altre argomentazioni che dicono molto poco dei reali motivi per i quali si è agito. Il nostro cervello, inoltre, non soltanto ha la capacità di ricostruire i processi decisionali ma, nientemeno, anche di inventarli *ex post*. Per dirla con un gergo più tecnico, il cervello riesce a *confabulare*, ovvero a razionalizzare *a posteriori* un'azione intrapresa non consapevolmente o, quantomeno, un'azione che abbiamo preso in modo leggermente differente da quella di cui poi ci viene richiesto di dare conto[28]. Come si è accennato, questo processo con l'IA è al momento impossibile, e questo fattore, definito come argomento della *black box*, è stato considerato un problema talmente inaggirabile da essere utilizzato sovente per gettare discredito sull'IA, tant'è che è stato

[27] F. Rossi, *op. cit.*, p. 52.

[28] Gli esperimenti in merito sono innumerevoli, ne riporto due particolarmente significativi: R.E. Nisbett, T.D. Wilson, *Telling More Than We Can Know*. Verbal Reports on Mental Processes, in «Psychological Review», 84, 1977, pp. 231-259. P. Johansson, L. Hall, S. Sikström, A. Olsson, *Failure to Detect Mismatches Between Intention and Outcome in a Simple Decision Task*, in «Science», 310, 2005, pp. 116-119.

fortemente contrastato dalla normativa europea secondo la quale un soggetto deve poter accedere al meccanismo col quale l'IA ha preso una decisione, almeno in quelle circostanze in cui quest'ultima ha un impatto sulla vita del soggetto[29]. È altresì vero che questo problema si sta avviando, seppur lentamente, a una risoluzione dal momento che diversi ricercatori stanno lavorando a fornire all'IA la capacità di spiegare perché abbia preso una decisione (es. ha interpretato un'immagine come un possibile tumore o no), ovvero della capacità di *spiegare* le proprie decisioni. Non entro qui nella questione, perché lo sviluppo di questa capacità algoritmica è appena incominciato e io stesso sto conducendo alcune ricerche in merito: solo un appunto – neanche troppo originale – ritengo però possa essere mosso[30]. Mancando all'IA delle preferenze, delle inclinazioni, in quanto le IA sono prive di carattere, in base a cosa la scelta sarebbe operata? Sarebbe soltanto un risultato ottimale a seguito di un calcolo? Anche ammettendo quest'ultima possibilità, viene da chiedersi quante volte noi calcoliamo approfonditamente e soppesiamo accuratamente i pro e i contro prima di deciderci ad agire[31]. Sembra ancora una volta come l'IA, lungi dal fare qualcosa di anche vagamenti simile all'attività umana, crei qualcosa di profondamente diverso, una nuova forma di *agere*, un *agere* che Floridi definisce come «agere sine intelligere»[32]. Rimane quindi da analizzare il secondo elemento che ho fatto emergere portando l'esempio di COMPAS e di Eric Loomis: i *bias* impliciti presenti anche negli algoritmi.

[29] Per i casi giuridici e i riferimenti normative dell'Unione Europea, cfr. T. Cassauwers, *Opening the "black box" of artificial intelligence*, in https://ec.europa.eu/research-and-innovation/en/horizon-magazine/opening-black-box-artificial-intelligence (12/3/2022).

[30] Ho già suggerito una simile riflessione nel momento in cui si valuta l'utilizzo delle IA per scelte e previsioni in ambito giuridico, cfr. C. Calì, *op. cit.*, pp. 132-134.

[31] Si noti per altro che questa critica è spesso mossa contro coloro che difendono posizioni libertarie in ordine alla libertà metafisica. Si veda a tal proposito, R. Kane, *The Significance of Free Will*, Oxford University Press, New York 1996.

[32] L. Floridi, *Etica*, cit., p. 13.

Un anno dopo la condanna di Loomis con l'ausilio di COMPAS, nel 2014, si riscontrò un errore nel *software* utilizzato da Amazon per lo screening dei *curricula vitae* dei candidati da assumere. Si notò come l'algoritmo penalizzasse i candidati di sesso femminile soprattutto quando si trattava di posizione nel settore tecnologico dell'azienda[33]. In quel caso «il modello ha riconosciuto in modo automatico un pattern che delineasse i migliori candidati, inglobando tra le caratteristiche ideali il genere maschile, e incorrendo così in un bias»[34]. L'errore – viene detto in molti dei commenti sulla vicenda – «era dovuto ai dati con cui il modello è stato addestrato: dati reali, contenenti i curricula ricevuti dalla società nei 10 anni precedenti; CV prettamente maschili, data la maggioranza di uomini nel settore tecnologico»[35].

Nel 2018, una giovane ricercatrice per Microsoft, poi assunta da Google, insieme con un collega del MIT, notò come certi algoritmi modellassero le elaborazioni linguistiche e il riconoscimento facciale su basi razziste, sessiste e offensive[36] (una precedente versione del software di Google, ad esempio, etichettava come *gorilla*, i cittadini di pelle scura[37]). L'indagine si estendeva ad alcuni algoritmi adoperati da altri fra i GAFAM. Gli algoritmi che stanno alla base del funzionamento di Google Cloud Vision, Microsoft Azure Computer Vision e Amazon Rekognition, ad esempio,

[33] Cfr. J. Dastin, *Amazon scraps secret AI recruiting tool that showed bias against women*, in https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G, (ultimo accesso 22/4/2022).

[34] Cfr. *Amazon: intelligenza artificiale discriminava curriculum donne*, in https://www.ansa.it/sito/notizie/tecnologia/tlc/2018/10/10/amazon-stampa-intelligenza-artificiale-discriminava-donne_1b1ebaca-f000-48a6-89b4-ff9854ef75e7.html (ultimo accesso 22/4/2022).

[35] G. Rizzi - M.T. Cimino, *Bias negli algoritmi: come le macchine apprendono i pregiudizi dagli esseri umani*, in https://ibicocca.unimib.it/bias-negli-algoritmi-come-le-macchine-apprendono-i-pregiudizi-dagli-esseri-umani/ (ultimo accesso 21/4/2022).

[36] Cfr. J. Buolamwini - T. Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, in «Proceedings of Machine Learning Research», 81(2018), pp. 1-15.

[37] Cfr. C. Dougherty, *Google Photos Mistakenly Labels Black People 'Gorillas'*, in https://bits.blogs.nytimes.com/2015/07/01/google-photos-mistakenly-labels-black-people-gorillas/ (ultimo accesso 22/4/2022).

osservavano e valutavano donne e uomini in modo molto differente, stimando le prime su criterii legati all'aspetto estetico e sessuale[38]. Noto *en passant*, come nel 2020 la ricercatrice – dedicatasi all'interno dell'azienda alla *machine ethics* – sia stata licenziata dal colosso della Silicon Valley a causa –stando alla versione ufficiosa – di un articolo scientifico in cui esponeva i numerosi rischi legati allo sviluppo delle IA.

Nel 2021, infine, il tribunale del lavoro di Bologna, con una sentenza storica, ha rintracciato atteggiamenti discriminatori nell'algoritmo di nome Frank, utilizzato da Deliveroo, nota piattaforma di *food delivery*. L'algoritmo discriminava i riders autonomi sulla base delle loro performance senza distinguere tra quelli che si assentavano per motivi futili e quelli che invece ricevevano un annullamento dell'ordine. Anche in questo caso è stato detto che «non è colpa dell'algoritmo: semplicemente Frank ripete e applica, in modo più o meno continuo, le direttive che gli sono state impiantate»[39].

È evidente come se la situazione in ordine alla decisione umana demandata alle IA destava qualche perplessità, la situazione in riferimento alle decisioni *delle* IA non versa in condizioni migliori. Benché in molti ripetano che l'IA sia neutra, in quanto non esiste un'IA buona e una cattiva, diviene sempre più difficile definire chi è il soggetto responsabile per una determinata decisione. Il tribunale di Bologna ha condannato e obbligato al risarcimento Deliveroo, ma è davvero l'azienda londinese di Will Shu a essere responsabile o, piuttosto, i programmatori assunti dalla stessa o, ancora, coloro che 'tra gli umani' procedono troppo spesso a valutazioni sommarie e poco accurate, sulla base delle quali l'IA apprende il proprio *modum operandi*? Anche su

[38] C. Schwemmer, C. Knight, E.D. Bello-Pardo, S. Oklobdzija, M. Schoonvelde, J.W. Lockhart, *Diagnosing Gender Bias in Image Recognition Systems*, in «Socius: Sociological Research for a Dynamic World», 6, 2020, pp. 1-17.

[39] E. Verga, *Intelligenza artificiale, la discriminazione da parte degli algoritmi è un pericolo reale*, in https://tech4future.info/discriminazione-degli-algoritmi-di-intelligenza-artificiale/ (ultimo accesso 22/4/2022).

questo secondo versante, nondimeno, si sta correndo ai ripari, grazie al campo di ricerca definito *algorithmic fairness*,

> volto a mitigare gli effetti di pregiudizi e discriminazioni ingiustificate sugli individui nell'apprendimento automatico, principalmente incentrato sul formalismo matematico e sulla ricerca di soluzioni per questi formalismi. È un ambito di ricerca interdisciplinare che ha l'obiettivo di creare modelli di apprendimento in grado di effettuare previsioni corrette dal punto di vista di equità e giustizia[40].

Anche in questo caso però il compito è particolarmente arduo dal momento che una prima difficoltà si riscontra nell'incapacità di ottenere una definizione esaustiva e universale di correttezza (*fairness*)[41], poiché quest'ultima risulta enormemente variabile in base ai contesti sociali, politici e religiosi di riferimento. È proprio a partire da questo tentativo col quale i programmatori vogliono correre ai ripari che la riflessione etica in riferimento alle IA deve interrogarsi. Ammesso che si possa dare responsabilità senza coscienza e, altresì, responsabilità senza libertà, è ancora opportuno continuare a utilizzare un termine che sin dalla notte dei tempi è stato, è, e sembra ancora essere – nonostante qualsiasi descrizione del mondo la fisica ci fornisca sulla nostra libertà – un termine imprescindibile per l'agire umano?

Floridi, come pocanzi accennato, si è mosso da tempo in questa direzione, comprendendo l'IA come una «nuova forma di agire intelligente, determinata dal disallineamento digitale tra azione e intelligenza»[42], un *agere sine intelligere*, al punto tale che l'ormai consueta sigla di AI (*artificial intelligence*) potrebbe essere anche riformulata in AA (*artificial agency*). Questa riformulazione dell'usuale binomio tra intelligenza e artificio in termini di *agency* penso sia cruciale dal momento che viene separata l'*attività* dalla *razionalità*, un binomio canonico tanto

[40] G. Rizzi, M.T. Cimino, *Bias negli algoritmi*, cit.
[41] A tal proposito, cfr. N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, *A Survey on Bias and Fairness in Machine Learning*, in «ACM Computing Surveys», 54, 2021, pp. 1-35.
[42] L. Floridi, *Etica*, cit., p. 21.

nell'antichità quanto nel periodo della Scolastica. Questo scollamento ha notevoli ripercussioni e fornisce innumerevoli prospettive d'indagine. Qui mi limito a percorrerne soltanto una. Nel momento in cui la decisione artificiale non è informata dall'intelligenza sembra configurarsi, ancora una volta, una situazione ibrida in cui la macchina viene in aiuto all'umano e l'umano corregge o sistema l'operato della macchina. Alla luce di ciò non si parlerà più di *autonomia* ma di meta-autonomia, o modello di decisione di delega:

> Gli esseri umani dovrebbero mantenere il potere di decidere quali decisioni prendere, esercitando la libertà di scelta dove necessario e cedendola nei casi in cui ragioni di primaria importanza, come l'efficacia, possano prevalere sulla perdita di controllo sul processo decisionale. Ma qualsiasi delega dovrebbe anche rimanere in linea di principio rivedibile, adottando come ultima garanzia il potere di decidere di decidere di nuovo[43].

Anche per la scelta, allora, si avrà una decisione composita in cui interagiscono tanto gli esseri umani quanto le macchine. Oggi, dice Marcello Ienca, la nostra specie si trova a un crocevia epocale:

> Davanti a noi, infatti, si profila una transizione storica senza precedenti noti nella storia del pianeta: quella nella quale un'intelligenza biologica inizia a utilizzare ricorsivamente un'intelligenza artificiale da lei stessa creata al fine di unirsi a essa e, in tal modo, autopotenziarsi[44].

Proprio in riferimento alla scelta, tuttavia, ritengo che i giudizi debbano essere molto più cauti di quanto non si è fatto in ordine alle macchine utilizzate in ambito aziendale o per fini medici. Quando deliberiamo, ma soprattutto quando valutazione, un computer è nettamente più affidabile di noi stessi, e, quando ricorriamo a certi algoritmi, di fatto «siamo spinti a rinunciare alla "libera scelta" e ad affidarci al sistema. È questa la promessa di ogni delega tecnocratica: *liberarci dalla libertà*»[45]. Ma dietro a questa efficienza nella valutazione vi è anche una

[43] *Ibid.*, p. 99.
[44] M. Ienca, *Intelligenza²*, cit., p. 14.
[45] N. Bellanca, *La libertà al tempo dell'Intelligenza Artificiale*, in https://www.academia.edu/43738365/La_libertà_al_tempo_dellIntelligenza_Artificiale (ultimo accesso 30/4/2021).

qualità della valutazione? «Se deleghiamo all'intelligenza artificiale alcune delle nostre decisioni, siamo certi che ne sappia valutare le conseguenze e quindi sappia prendere le decisioni migliori? Se in alcune attività la tecnologia diventa più *capace* di noi, quale finirà per essere il nostro ruolo»[46]?

Penso che questi interrogativi ci impongano di scandagliare non più il nesso intelligenza e artificio ma quello tra libertà e artificio. Nel momento in cui separiamo l'azione della macchina dall'*intelligere*, poiché l'IA «non concerne la capacità di riprodurre l'intelligenza umana, ma in realtà la capacità di farne a meno»[47], a cosa riconduciamo l'azione artificiale? Vedo in questo una riproposizione, *mutatis mutandis*, della *querelle* medievale sulla determinazione della volontà da parte della ragione o da parte della volontà stessa. Come la discussione sull'intelligenza applicata alle macchine ha imposto un ripensamento della nozione di intelligenza, o quantomeno ha riacceso la riflessione, lo stesso si deve dire della libertà, la quale viene illuminata nuovamente e in modo differente dalle discussioni legate all'intelligenza artificiale.

Alla luce di questo dato incontrovertibile, allora, è necessario ripensare il vocabolario e il contenuto dell'etica non adattando le classiche etiche – nate per un'*agency* umana – ma formulando un'etica artificiale per un'"agency" artificiale, sempre ammesso che della macchina si possa predicare un'*agency*.

Cristiano Calì è PhD candidate presso la Facoltà di Teologia di Lugano e Docente invitato presso lo Studio Teologico S. Paolo di Catania
cristianocali30@icloud.com

[46] F. Rossi, *op. cit.*, p. 12, corsivo mio.
[47] L. Floridi, *Etica*, cit., p. 52.