SULLE IMPLICAZIONI ETICHE DELLA INTELLIGENZA ARTIFICIALE

Stefano Zamagni

1. E' sul fronte dell'etica pubblica che le conseguenze della diffusione nelle nostre società delle Intelligenze Artificiali (IA) vanno ponendo i problemi più delicati, primo fra tutti quello di capire come la digitalizzazione della nostra vita stia riuscendo a modificare anche il modo di percepirla. Eppure, è proprio su tale fronte che si registra una sorta di "fin de non recevoir" da parte dell'alta cultura, scientifica e filosofica, mentre abbondante è la riflessione sulle implicazioni riguardanti il mercato del lavoro, l'organizzazione delle imprese, il reddito nazionale, l'articolazione sociale. Di due (soli) aspetti particolari desidero qui dire in breve. Il primo concerne la questione della fiducia: può l'intelligenza artificiale creare la fiducia che è necessaria per realizzare uno sviluppo umano integrale? Il secondo aspetto chiama in causa il problema della responsabilità, di cosa significhi essere responsabili nell'era della digitalizzazione. Sono le "smart machines" agenti morali e dunque responsabili? Saranno gli algoritmi a governarci, in tutti i casi in cui le persone non sono in grado di comprendere appieno le questioni sulle quali debbono esprimere valutazioni? Comincio dal primo aspetto.

Generale è il consenso sul fatto che è la fiducia uno dei fattori decisivi per assicurare i vantaggi dell'agire collettivo e, per questa via, sostenere il processo di sviluppo, ma si noti il paradosso tipico dell'attuale fase storica. Mentre la fiducia nei confronti delle istituzioni, sia politiche sia economiche, va declinando per una pluralità di ragioni, il mercato globale è conquistato sempre più da imprese e organizzazioni che chiedono ai loro clienti e utenti prove di fiducia, mai viste in passato. E' come se gli individui stessero imparando la lezione della ben nota vicenda della Tosca di Puccini: la mutua sfiducia genera sempre risultati subottimali. Per Tim Wu, affermato giurista della Columbia University, quello cui stiamo assistendo è un massiccio trasferimento di fiducia sociale: abbandonata la fiducia nelle istituzioni, ci si rivolge alla tecnologia. "La fiducia – scrive R. Botsman (*Di chi possiamo fidarci*, Milano, Hoepli, 2017) – è la nuova valuta dell'economia mondiale. E' un vero moltiplicatore di opportunità di guadagno perché consente di far fruttare beni sottoutilizzati". Si pensi al fenomeno delle criptomonete – la più nota delle quali, ma non certo l'unica, è il bitcoin – che sono valute digitali che si scambiano tra pari. Le transazioni non sono garantite da alcuna autorità centrale, ma convalidate dagli stessi partecipanti alla rete mediante un algoritmo. Al tempo stesso, la forza di queste

criptomonete è che esse consentono di effettuare transazioni anonime non soggette a tassazione e al riparo da confisca da parte dello Stato. L'infrastruttura che ne è alla base è il blockchain, che è un registro di proprietà distribuita su cui sono annotati tutti gli scambi, senza possibilità di modifica. La tecnologia blockchain – finora utilizzata praticamente solo in ambito finanziario – consente già oggi una vasta gamma di applicazioni, da quelle in ambito sociale a quelle di tipo politico-amministrativo. Si pensi alla gestione dei processi amministrativi, dove la blockclain può certificare in modo certo e per sempre un determinato atto senza bisogno di un soggetto Terzo certificatore. Si consideri anche che le Nazioni Unite stanno progettando di avvalersi della medesima tecnologia per la gestione degli aiuti, di varia natura, ai profughi e migranti. E così via.

Il cuore del paradosso odierno è in ciò che l'economia di mercato contemporanea ha ancora più bisogno di quella del passato di fiducia reciproca per poter funzionare al meglio. Al tempo stesso, però, gli straordinari livelli di efficienza finora raggiunti dai nostri sistemi economici sembrano far dimenticare che è necessario rinforzare le reti fiduciarie tra persone perchè il mercato mentre "consuma" sempre più fiducia non riesce, stante l'attuale assetto istituzionale, a produrne a sufficienza. Di qui l'inquietante dilemma sociale: chiediamo sempre più efficienza per accrescere il benessere materiale, la ricchezza, la sicurezza, ma per conseguire un tale obiettivo decumuliamo irresponsabilmente il patrimonio di fiducia che abbiamo ereditato dalle generazioni passate.

2. Che fare per sciogliere questo dilemma? E' nota la proposta di David Hume. Per il fondatore dell'empirismo filosofico (e iniziatore del non cognitivismo etico) la disposizione ad accordare fiducia, e a ripagare la fiducia concessa, trova il proprio fondamento nei vantaggi personali che scaturiscono da una buona reputazione. "Possiamo soddisfare i nostri appetiti meglio in un modo indiretto e artificiale … E' così che imparo a prestare un servigio ad un altro senza provare per lui una vera benevolenza. Infatti io prevedo che egli mi renderà il servigio attendendosene un altro dello stesso tipo, per conservare la medesima *reciprocità* di buoni uffici con me o con altri" (*Trattato sulla natura umana* [1740] 1971, pp. 552-3). E' quasi incredibile che un grande filosofo come Hume sia potuto cadere in una così patente svista concettuale, quella di confondere la reciprocità con una sequenza di scambi auto-interessati. La reciprocità, a differenza dello scambio di equivalenti, è un insieme di relazioni di dono tra loro interrelate.

Anche la soluzione dell'imperativo categorico kantiano non ci è di grande aiuto agli scopi presenti. "Segui la regola che, se ognuno la seguisse, tu potresti volerne il risultato". E' questo un principio di eguaglianza del dovere. Tuttavia, la teoria di Kant

soffre di una evidente aporia quando si cercasse di porla in pratica. Infatti, l'individuo kantiano sceglie la regola (la massima) che va ad applicare assumendo che anche tutti gli altri la applichino. Ma poiché persone diverse, in generale, hanno preferenze diverse circa il risultato finale, anche le regole kantiane da esse preferite saranno a priori diverse. Ne consegue che ciascuno seguirà la sua regola preferita (da cui la sua azione) assumendo che gli altri agiscano nel modo in cui in realtà essi non agiranno affatto. Ciò significa che il principio kantiano non può applicarsi a se stesso; non può validare se stesso: davvero una seria incongruenza logica per una dottrina morale che ambisce ad essere universale. Solamente se tutti gli individui fossero tra loro identici quanto al loro sistema preferenziale l'aporia in questione scomparirebbe. Ma è evidente che se così fosse il principio kantiano perderebbe tutta la sua rilevanza pratica.

La recente ricerca nell'ambito delle neuroscienze va oggi suggerendo la seguente via d'uscita dal dilemma sopra indicato. In un lavoro collettaneo pubblicato sulla prestigiosa rivista *Science* (2006) si legge che se si disattiva, mediante stimolazione magnetica transcranica, una particolare zona della corteccia cerebrale, i soggetti aumentano notevolmente il loro comportamento prosociale, il che conduce ad un sostanziale incremento del loro grado di fiducia. In particolare, somministrando per via nasale una certa quantità di ossitocina (un ormone naturalmente prodotto dall'organismo di molti mammiferi) si è scoperto che esso deattiva l'attività cerebrale di una specifica regione del cervello (l'amigdala) deputata a controllare il comportamento degli individui nei rapporti fiduciari. (D. Narvaez, *Neurobiology and the Development of Human Morality*, Norton, New York, 2014). Si pensi anche agli interventi volti al potenziamento cognitivo che agiscono su capacità come l'attenzione, la memoria, la tendenza all'affaticamento intellettuale. Già vengono praticate tecniche come la stimolazione celebrale profonda (*deep brain stimulation*) che prevede l'impianto di un microchip nel cervello; come la stimolazione transcronica a corrente diretta (*transcranical direct current stimulation*) che prevede la stimolazione dell'encefalo con dosi di corrente elettrica.

Pochi anni fa, un gruppo di ricercatori dell'Università di Berkeley hanno testato su un campione di trentacinque soggetti un farmaco "in grado di produrre artificialmente sentimenti di bontà e di benevolenza verso gli altri" (*Current Biology*, 3, 2014). I risultati ottenuti confermerebbero che il tolcapone, altro ormone umano, contribuisce ad accrescere il tasso di equanimità nei confronti anche di sconosciuti e ad accrescere, per questa via, la propensione alla fiducia. (Si tratta di tentativi che mirano al *mood enhancement* delle persone, per modificarne il carattere e aumentarne il benessere psicologico, contrastando la disposizione alla tristezza e all'introversione). Non si può non discutere della plausibilità di risultati simili e di giudicare l'efficacia, nella pratica, di proposte come quella di somministrare per via chimica molecole atte a potenziare la nostra moralità. Mi limito ad osservare che il

tentativo di attribuire l'origine del senso morale alla biologia, tentativo che riduce tale senso a mera chimica cerebrale, se da un lato può sortire effetti desiderati rispetto a ciò che è funzionale al buon andamento degli affari, dall'altro riduce lo spazio della libertà (positiva) e quindi della responsabilità individuale. Vedere il pensiero morale come intrinseco al cervello umano, piuttosto che come prodotto di volontà e di virtù, comporterebbe un serio e pericoloso arretramento.

3. Non v'è chi non veda come approcci del genere si collochino, al di là delle apparenze o delle dichiarazioni ufficiali, nell'ambito di quel grande progetto, politico e filosofico insieme, che è il transumanesimo, la cui ambizione è sia fondere l'uomo con la macchina per ampliarne le potenzialità in modo indefinito sia (e soprattutto) arrivare a dimostrare che la coscienza non è un ente esclusivamente umano. L'obiettivo qui non è tanto commerciale o finanziario: è politico, e in un certo senso religioso e ciò nel senso che il progetto ambisce a trasformare – non tanto a migliorare - il nostro modo di vivere, oltre che i nostri valori di riferimento. Il transumanesimo è l'apologia di un corpo e di un cervello umani "aumentati", arricchiti cioè dall'intelligenza artificiale, il cui utilizzo consentirebbe di separare la mente dal corpo e quindi di affermare che il nostro cervello per funzionare non avrebbe necessità di avere un corpo, e questo permetterebbe di sviluppare argomenti riguardanti il significato della persona e della sua unità.

La strategia perseguita da Ray Kurzweil, responsabile del progetto che Google va da qualche tempo implementando, mira alla produzione di cyborg dotati di sembianze fisiche e capacità cognitive simili a quelle dell'*homo sapiens*. E' l'obiettivo del *playing God* (recitare la parte di Dio) che nasconde il desiderio di prendere in mano le redini dell'evoluzione. L'approccio fisicalista (secondo cui esisterebbe soltanto una realtà – quella fisica – che le scienze cognitive cercano di comprendere per spiegare come si genera la conoscenza), accolto dalle neuroscience pone in discussione il nesso tra responsabilità e libertà. Veniamo da una lunga stagione durante la quale era assodato ritenere che alla libertà come espressione della responsabilità corrispondesse la responsabilità come consenso all'applicazione della stessa libertà. Cosa significa per un operatore lavorare tutto il giorno con un robot collaborativo? Sappiamo già come l'avvento dei social network e l'uso degli smartphone stiano cambiando le nostre abitudini e i nostri stili di vita. Ma possiamo pensare un futuro in cui l'uomo trascorre tutta la sua giornata lavorativa "dialogando" – si fa per dire – con un robot, senza cadere in forme nuove e più gravi di alienazione? Si prenda nota della radicale differenza tra automazione e Intelligenza Artificiale. Mentre la prima agisce sull'oggetto (si pensi a Internet of things), la seconda interviene sul soggetto che lavora.

4. Passo ora al tema della responsabilità. Come sappiamo, la responsabilità possiede significati diversi. Si può dire responsabilità per significare una libertà che possiede il senso della responsabilità. Ma si può dire responsabilità in senso molto diverso quando si è incaricati di un compito di cui si deve rispondere. (E' il concetto americano di "accountability"). Infine, si può dire responsabilità per indicare che si è colpevoli di un'attività portata a compimento. In tal senso, "io sono responsabile" significa che sono colpevole di qualcosa. Responsabilità e libertà risultano pertanto fortemente correlate, anche se, in tempi recenti, sull'onda degli avanzamenti registrati sul fronte delle neuroscienze, si tende ad allentare il nesso tra libertà e responsabilità. Si considerino gli interventi di potenziamento sull'uomo. Il soggetto potenziato prenderebbe le sue decisioni non sulle ragioni pro e contro, ma in seguito all'influsso causale esercitato sul suo cervello dai mezzi di manipolazione biotecnologica. Quanto a dire che per migliorare la performance degli esseri umani li si priva della loro autonomia morale, che è il bene più prezioso.

Mentre sembra relativamente facile identificare la responsabilità diretta degli agenti – come quando il proprietario di uno *sweatshop* sfrutta il lavoro minorile per trarne un vantaggio – che dire dell'azione economica che è intrapresa con l'intenzione di non svantaggiare nessuno e tuttavia provoca effetti negativi in capo ad altri? Ad esempio, di chi è la responsabilità della disoccupazione, della povertà, delle disuguaglianze, etc? Le risposte tradizionali, in economia, consistono nel sostenere che si tratta di conseguenze non volute delle azioni intenzionali (le "unintended consequences of intentional actions" di cui ha parlato la Scuola dei moralisti scozzesi del 18° secolo). E dunque l'unica cosa da fare è di attribuire alla società il compito di porre rimedio (o di alleviare) le conseguenze negative. E infatti il welfare state è sorto e si è sviluppato precisamente per rendere collettiva e impersonale la responsabilità dei singoli. Ma è veramente così? Siamo sicuri che i meccanismi del libero mercato siano inevitabili e che gli effetti che ne derivano siano inattesi, come si tende a far credere? (Si veda la *Caritas in Veritate* di Benedetto XVI, capp. 3 e 4).

Valga un solo esempio. Albert Carr nel saggio "Is business bluffing ethical?" (*Harvard Business Review,* 1968) – il saggio più citato di sempre in teoria della finanza – scrive: "La finanza, dove si cerca di fare agli altri ciò che non si desidera gli altri facciano a noi (*sic!*) dovrebbe essere guidata da un insieme di standard etici diversi da quelli della morale comune o della religione: gli standard etici del gioco. Se un'azione non è strettamente illegale, e può dare un profitto, allora compierla è un *obbligo* dell'uomo d'affari". E' questo modo di pensare – fondato sulla tesi della doppia moralità – che è all'origine di tutti i grandi scandali finanziari, tra cui quelli dell'ultimo ventennio. Come, tra i primi, aveva notato Z. Bauman, l'organizzazione sociale della seconda modernità è stata pensata e disegnata per neutralizzare la responsabilità

diretta e indiretta degli agenti. La strategia adottata – di grande raffinatezza intellettuale – è stata quella, per un verso, di allungare la distanza (spaziale e temporale) tra l'azione e le sue conseguenze e, per l'altro verso, di realizzare una grossa concentrazione di attività economica senza una centralizzazione di potere. E' in ciò il carattere specifico dell'impresa adiaforica, una figura di impresa ignota alle epoche precedenti la seconda guerra mondiale e il cui fine è quello di annullare la questione della responsabilità morale dell'azione organizzata. Adiaforica è la responsabilità "tecnica" che non può essere giudicata in termini morali di bene/male. L'azione adiaforica va valutata in termini solamente funzionali, sulla base del principio che tutto ciò che è possibile per gli agenti sia anche eticamente lecito, senza che si possa giudicare eticamente il sistema, come Luhmann ha insegnato.

Ebbene, la responsabilità adiaforica ha ricevuto, in tempi recenti, nuovo impulso proprio dalla Intelligenza Artificiale, la quale va producendo "mezzi" che sono alla ricerca di "domande" o di problemi da risolvere. Esattamente il contrario di quanto era accaduto con le precedenti rivoluzioni industriali. Invero, cosa ne è del principio di responsabilità nella società degli algoritmi? Dalle nuove tecnologie industriali alla diagnostica medica, dai social networks ai voli degli aerei, dai big data ai motori di ricerca: ci affidiamo a complesse procedure cui deleghiamo la buona riuscita di operazioni che gli esseri umani, da soli, non saprebbero eseguire. Eppure, gli algoritmi sono irresponsabili, pur non essendo neutrali, né oggettivi, come invece erroneamente si crede. Quando un programma commette un errore non ne paga le conseguenze, perché si ritiene che la matematica resti al di fuori della moralità. Ma non è così, perché gli algoritmi non sono pura matematica; sono opinioni umane incastonate in linguaggio matematico. E dunque discriminano, al pari dei decisori umani. Ad esempio, il processo delle assunzioni di lavoro si va sempre più automatizzando, perché si pensa di rendere obiettivo il reclutamento del personale, eliminando pregiudizi. Ma le dinamiche discriminatorie, anziché diminuire, stanno aumentando nelle nostre società.

5. Generalizzando un istante, il vero problema delle *smart machines* comincia nel momento in cui queste compiono azioni che coinvolgono la necessità di scegliere oppure di decidere. Il soldato-robot, l'automobile-robot, la scopa-robot potrebbero compiere scelte esiziali per vite non robotiche. Di chi è la responsabilità in questi casi? Quale il messaggio subliminale della recente provocazione di Bill Gates di tassare i robot, che andrebbero dotati di personalità elettronica, al pari delle corporations che sono dotate di personalità giuridica? Come ha lucidamente spiegato Gunther Anders, il XXI secolo ha inaugurato l'era dell'irresponsabilità umana, immunizzando i soggetti dalle loro relazioni. Le "smart machines" (quelle dotate di intelligenza artificiale) sono in grado di prendere

decisioni autonome, che hanno implicazioni sia sociali sia morali. (Si veda il caso dell'auto senza pilota Tesla, creata da Elon Musk, che nel maggio 2016 uccise un passeggero). Come assicurare, allora, che le decisioni prese da tali oggetti siano eticamente accettabili? Dato che queste macchine possono causare danni di ogni sorta, come fare in modo che esse siano poste in grado di differenziare tra decisioni corrette" e "sbagliate" ? E nel caso in cui un qualche danno non possa essere evitato - si pensi al caso dell'auto senza conducente che deve scegliere se gettarsi contro un altro veicolo uccidendone i passeggeri oppure investire dei bambini che attraversano la strada -, come istruire (nel senso di programmare) tali macchine a scegliere il danno minore? Gli esempi in letteratura sono ormai schiera. E tutti concordano sulla necessità di dotare l'IA di un qualche canone etico, per sciogliere dilemmi morali del tipo "guida autonoma".

Le divergenze nascono nel momento in cui si deve scegliere il modo (cioè l'approccio) secondo cui procedere: *top-down* (i principi etici sono programmati nella macchina intelligente: l'uomo trasferisce all'intelligenza artificiale la sua visione etica del mondo) oppure *bottom-up* (la macchina impara a prendere decisioni eticamente sensibili dall'osservazione del comportamento umano in situazioni reali). Entrambi gli approcci pongono problemi seri, che non sono tanto di natura tecnica quanto piuttosto concernono la grossa questione se le macchine intelligenti debbano o meno essere considerate agenti morali (cioè *moral machines*). Siamo appena agli inizi di un dibattito culturale e scientifico che già si preannuncia affascinante e preoccupante ad un tempo. Si veda, ad esempio, la recente presa di posizione di A. Etzioni e O. Etzioni ("Incorporating Ethics into Artificial Intelligence", *Journal of Ethics,* March, 2017) che negano la possibilità di attribuire lo status di agente morale all'Intelligenza Artificiale e dunque negano ogni fondamento al programma di ricerca della *Internet Ethics* che studia gli aspetti etici della *Internet Communication* nelle sue varie articolazioni..

Da ciò essi traggono la conclusione che non vi sarebbe alcun bisogno di insegnare etica alle macchine, anche se ciò potesse essere fattibile. Non è del medesimo parere, ad esempio, il gruppo di ricerca che opera per la Neurolink Corporation, in California, che da qualche tempo sta sviluppando tecnologie digitali per realizzare connessioni tra computer e mente umana e che sta progettando un uomo-cyber con microchip nel cervello. Sulla intricata e delicata questione concernente la possibilità di attribuire "personalità elettronica" ai robot intelligenti e, più in generale, la opportunità di favorire il passaggio dalla selezione naturale darwiniana alla scelta deliberata del processo di selezione mediante la scorciatoia biotecnologica, si veda S.M. Kampowski, D. Moltisanti, a cura di, *Migliorare l'uomo? La sfida dell'enhancement,* Cantagalli, Siena, 2011.

6. Nel dibattito corrente, due diversi modi di concettualizzare l'IA si vanno confrontando. Il primo concerne quei software che cercano di ragionare e di prendere decisioni cognitive al modo in cui gli umani lo fanno. Per tale concezione, l'IA aspirerebbe a rimpiazzare l'uomo. (Il famoso test di Turing ha a che vedere con questo tipo di IA). Il secondo modo mira invece a fornire un'assistenza smart agli attori umani. Si tratta di una IA partner dell'uomo, spesso indicata come "Intelligence Augmentation" ovvero "Cognitive augmentation". Nel concreto, Google si sta muovendo nella prima direzione; l'obiettivo dichiarato è quello di arrivare a fondere l'uomo con la macchina per accrescerne, senza limite, le competenze; IBM, con il suo *cognitive computing*, nella seconda. Nel 2013, IBM ha lanciato il sistema di Intelligenza Artificiale "Thomas Watson" in omaggio al nome del suo primo presidente. Watson risponde alle domande poste in linguaggio naturale su qualsiasi tematica. Nella prima pagina del sito dedicato a Watson si legge: "Watson è una tecnologia cognitiva che può pensare come un essere umano". Si tratterà di vedere se la macchina potrà diventare più intelligente dell'uomo. In ogni caso, resta vero che le risposte standardizzate che Watson (o altra macchina) potrà dare non saranno più efficaci di quelle che possono dare persone in grado di comprendere i problemi di altre persone. Le macchine, per quanto intelligenti, mai saranno capaci di empatia, perché non capaci di sentimenti morali.

Generalizzando un istante, il problema serio che sorge è che l'IA costituisce un divario tra agire e intelligenza. Invero, l'IA è una trasformazione che ha a che fare non tanto con l'intelligenza, quanto piuttosto con l'agire. Il *machine learning*, il *naturale language processing,* la robotica sono straordinarie soluzioni per far svolgere alla macchina compiti propri della persona umana, in maniera più efficace, ma senza intelligenza. Si sta separando la capacità di risolvere problemi dalla necessità di essere intelligenti nel farlo. Questa è la grande e rischiosa novità dell'oggi. Mancando l'intelligenza, non c'è intenzionalità, senso, eticità dell'agire. La grande minaccia è che l'IA si trascini dietro un'etica artificiale.

E' giunto il tempo di porsi a pensare seriamente, e in modo sistematico la governance del digitale. Ma per questo, ci vuole un'idea di progetto umano, occorre aver chiaro dove si sta andando. Chi deve assumersi un tale compito di orientamento se non la Chiesa?