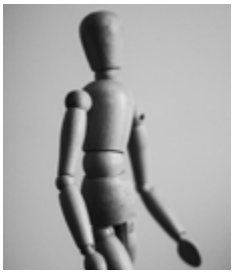

La dignità della persona nell'epoca della machina sapiens

di Paolo Benanti

Francescano, esperto in Teologia morale e bioetica; docente alla Pontificia Università Gregoriana

Nuovi artefatti: la machina sapiens

L'avvento della ricerca digitale, dove tutto viene trasformato in dati numerici, porta alla capacità di studiare il mondo secondo nuovi paradigmi gnoseologici: quello che conta è solo la correlazione tra due quantità di dati e non più una teoria coerente che spieghi tale connessione. Oggi la correlazione viene usata per predire con sufficiente accuratezza, pur non avendo alcuna teoria scientifica che lo supporti, il rischio di impatto di asteroidi anche sconosciuti in vari luoghi della Terra, i siti istituzionali oggetto di attacchi terroristici, il voto dei singoli cittadini alle elezioni presidenziali USA, l'andamento del mercato azionario nel breve termine.

Quello che appare come esito di questa *nuova rivoluzione* è il dominio dell'informazione, un labirinto concettuale la cui definizione più diffusa è basata sull'altrettanto problematica categoria di *dati*. Questa interpretazione dell'informazione come connessa al concetto di *dato* ha portato a sviluppare la cosiddetta Definizione Generale di Informazione (DGI) espressa in termini di d*ati + significato*. La DGI è ormai uno standard operativo, in particolare nei campi in cui i dati e le informazioni sono trattate come entità reificate[1].

L'evoluzione tecnologica dell'informazione e del mondo inteso come una serie di dati si concretizza nelle Intelligenze Artificiali (AI) e nei robot: siamo in grado di costruire macchine che possono prendere decisioni autonome e coesistere con l'uomo. Si pensi alle macchine a guida autonoma che Uber, il noto servizio di trasporto automobilistico privato, già utilizza in alcune città come Pittsburgh, o a sistemi di radiochirurgia come il Cyberknife o ai robot destinati al lavoro in affiancamento all'uomo nei processi produttivi in fabbrica. Le AI, queste nuove tecnologie, sono pervasive. Stanno insinuandosi in ogni ambito della nostra esistenza. Tanto nei sistemi di produzione, *incarnandosi* in robot, quanto nei sistemi di gestione sostituendo i server e gli analisti. Ma anche nella vita quotidiana i sistemi di AI sono

1 Non abbiamo qui modo di affrontare la questione. Rimandiamo al testo di L. Floridi, *La rivoluzione dell'informazione*, Codice, Torino 2012.

sempre più pervasivi. Gli smartphone di ultima generazione sono tutti venduti con un assistente dotato di Intelligenza Artificiale: Cortana, Siri o Google Hello – per citare solo i principali –, che trasforma il telefono da un *hub* di servizi e applicazioni a un vero e proprio partner che interagisce in maniera cognitiva con l'utente. Sono in fase di sviluppo sistemi di Intelligenza Artificiale, i bot, che saranno disponibili come partner virtuali da interrogare via voce o in chat e che sono in grado di fornire servizi e prestazioni che prima erano esclusiva di particolari professioni: avvocati, medici e psicologi, sono sempre più efficientemente sostituibili da bot dotati di Intelligenza Artificiale.

Il mondo del lavoro conosce oggi una nova frontiera: le interazioni e la coesistenza tra uomini e Intelligenze Artificiali. Prima di addentrarci ulteriormente nel significato di questa trasformazione dobbiamo considerare un implicito culturale che rischia di sviare la nostra comprensione del tema. Nello sviluppo delle Intelligenze Artificiali la divulgazione dei successi ottenuti da queste macchine è sempre stata presentata secondo un modello competitivo rispetto all'uomo. Per fare un esempio IBM ha presentato Deep Blue come l'Intelligenza Artificiale che nel 1996 riuscì a sconfiggere a scacchi il campione del mondo in carica, Garry Kasparov; sempre IBM nel 2011 ha realizzato Watson che ha sconfitto i campioni di Jeopardy, un noto gioco televisivo sulla cultura generale. Queste comparse mediatiche delle AI potrebbero farci pensare ad esse come a sistemi che competono con l'uomo e che tra *homo sapiens* e questa nuova *machina sapiens*/macchina autonoma si sia instaurata una rivalità di natura evolutiva che vedrà un solo vincitore e condannerà lo sconfitto a una inesorabile estinzione. In realtà, queste macchine non sono mai state costruite per competere con l'uomo, ma per realizzare una nuova simbiosi tra l'uomo e i suoi artefatti: *(homo+machina) sapiens*[2]. Non sono le AI la minaccia di estinzione dell'uomo, anche se la tecnologia può essere pericolosa per la nostra sopravvivenza come specie: l'uomo ha già rischiato di estinguersi perché battuto da una macchina *molto stupida* come la bomba atomica. Tuttavia esistono sfide estremamente delicate nella società contemporanea in cui la variabile più importante non è l'intelligenza ma il poco tempo a disposizione per decidere e le macchine cognitive trovano qui grande interesse applicativo.

Si aprono a questo livello tutta una serie di problematiche etiche su come validare la cognizione della macchina alla luce proprio della velocità della risposta che si cerca di implementare e ottenere. Tuttavia, il pericolo maggiore non viene dalle AI in se stesse ma dal non conoscere queste tecnologie e dal lasciare decidere sul loro impiego a una classe dirigente assolutamente non preparata a gestire il tema.

Se l'orizzonte di esistenza delle persone nel prossimo futuro – in realtà già del nostro presente – è quello di una cooperazione tra intelligenza umana e Intelligenza Artificiale e tra agenti umani e agenti robotici autonomi, diviene urgente cercare di

2 Cf. J.E. Kelly - S. Hamm, *Macchine intelligenti. Watson e l'era del cognitive computing*, Egea, Milano 2016, pp. 5-42.

capire in che maniera questa realtà mista, composta da agenti autonomi umani e agenti autonomi robotici, possa coesistere.

Primum non nocere

Il primo e più urgente punto che le Intelligenze Artificiali pongono nell'agenda dell'innovazione del lavoro è quello di adattare le nostre strutture sociali a questa nuova e inedita società fatta di agenti autonomi misti. Una prima sfida è di natura filosofica e antropologia. Queste frontiere dell'innovazione, la realizzazione di queste macchine sapiens, per utilizzare un termine molto evocativo, ci interrogano in profondità sulla specificità dell'*homo sapiens* e in particolare su quale sia la specifica componente e qualità umana del lavoro rispetto a quella macchinica: le rivoluzioni industriali hanno dimostrato che non è l'energia, non è la velocità e, ora, che anche la cognizione e l'adattabilità alla situazione non sono specifiche solamente umane.

La ricerca di risposte su questo tema è quanto mai urgente e importante per non sancire un declino dell'uomo negli orizzonti del *postumano*. Gli appartenenti a questa corrente di pensiero propugnano l'idea di un uomo in crisi, incapace di saper gestire le macchine che lui stesso ha creato. L'uomo sarebbe destinato a essere confinato in un passato fatto di residui archeologici[3]. Il *postumano* si configura, quindi, attorno all'idea centrale di un'umanità *sconfitta* dal suo stesso progresso[4].

Un secondo e altrettanto urgente tema è quello di definire come e in che maniera si può garantire la coesistenza tra uomo e AI, tra uomo e robot. Per rispondere a questa domanda, in primo luogo cercheremo di formulare una direttiva fondamentale che deve essere garantita dalle AI e dai robot, poi cercheremo di definire cosa questi sistemi cognitivi autonomi *devono imparare* per poter convivere e lavorare cooperativamente con l'uomo.

La prima e fondamentale direttiva da implementare può essere racchiusa nell'adagio latino *primum non nocere*. La realizzazione di tecnologie controllate da sistemi di AI porta con sé una serie di problemi legati alla gestione dell'autonomia decisionale di cui questi apparati godono. La capacità dei robot di mutare il loro comportamento in base alle condizioni in cui operano, per analogia con l'essere umano, viene definita *autonomia*. Per indicare tutte le complessità che derivano dal tipo di libertà decisionale di queste macchine si è introdotto il termine Artificial Moral Agent (AMA): parlando di AMA si indica quel settore che studia come definire dei criteri informatici per creare una sorta di *moralità artificiale* nei sistemi AI, portando

3 Cf. P. Benanti, *The Cyborg. Corpo e corporeità nell'epoca del postumano*, Cittadella, Assisi 2012.
4 Il tema per quanto affascinante non può essere affrontato in questa sede, rimandiamo a F. Occhetta – P. Benanti, *La politica di fronte alle sfide del postumano*, in *La Civiltà Cattolica*, 3954, I (2015), pp. 572-584.

alcuni studiosi a coniare l'espressione *macchine morali* per questi sistemi[5]. Quando si usa il termine *autonomia* legato al mondo della robotica, si vuole intendere il funzionamento di sistemi di AI la cui programmazione li rende in grado di adattare il loro comportamento in base alle circostanze in cui si trovano a operare[6]. Un esempio classico di applicazione di questa direttiva fondamentale, chiamato "situazione dei due carrelli", è stato formulato da Philippa Foot nel 1967 mentre si sperimentavano i primi sistemi di guida automatica dei mezzi per il trasporto di passeggeri negli aeroporti. Nel caso presentato dalla Foot un veicolo si avvicina a un incrocio e realizza che un altro veicolo, con cinque passeggeri, in direzione opposta è in traiettoria di collisione. Il primo veicolo può continuare sulla sua traiettoria e urtare l'altro uccidendo i cinque passeggeri a bordo o sterzare e colpire un pedone uccidendolo. La Foot si chiedeva: se noi fossimo alla guida del veicolo cosa faremmo? E un sistema robotizzato cosa dovrebbe fare? Giungendo alla conclusione che la macchina autonoma deve essere programmata per evitare assolutamente di ferire o uccidere l'essere umano e che, se in situazioni estreme non fosse possibile evitare di nuocere all'uomo, avrebbe dovuto scegliere il male minore[7].

Tuttavia, racchiudere la questione degli agenti morali autonomi e dell'utilizzo di robot cognitivi in un ambiente di lavoro misto umano-robotico non può esaurirsi in questa direttiva primaria. Sfruttando un linguaggio evocativo potremmo dire che le macchine *sapienti/autonome* per poter coesistere con i lavoratori umani devono *imparare* almeno quattro cose: Intuizione, Intellegibilità, Adattabilità, Adeguatezza degli obiettivi. Questi quattro elementi possiamo capirli come una declinazione operativa della dignità insita nel lavoratore umano. Solo se le macchine sapranno interagire con l'uomo secondo queste direzioni allora non solo non nuoceranno alla persona, ma ne sapranno tutelare la dignità e l'inventività senza mortificarne l'intrinseco valore.

Intuizione

Quando due esseri umani cooperano normalmente, l'uno riesce ad anticipare e assecondare le intenzioni dell'altro perché riesce a intuire cosa sta facendo o cosa vuole fare. Si pensi alla situazione in cui vediamo una persona che cammina con le braccia piene di pacchi. Istintivamente capiamo che la persona sta trasportando quei pacchi e la aiutiamo rendendole il lavoro più semplice o trasportando per lei parte del fardello che le ingombra le braccia. Questa capacità umana è alla base della grande duttilità che caratterizza la nostra specie e che ci ha permesso di organizzarci fin dai tempi più antichi riuscendo a cooperare nella caccia, nell'agri-

5 Cf. W. Wallach - C. Allen, *Moral Machines: Teaching Robots Right from Wrong*, Oxford University Press, New York 2008, pp. 55-79.
6 Cf. E. Yudkowsky, *Levels of Organization in General Intelligence*, in *Artificial General Intelligence (Cognitive Technologies)*, a cura di B. Goertzel, C. Pennachin, Springer, Berlin 2007, pp. 389-498.
7 Cf. W. Wallach - C. Allen, *Moral Machines...* cit., p. 13 e R. Arkin, *Governing Lethal Behavior in Autonomous Robots*, Chapman & Hall, Boca Raton 2009, pp. 37-47.

coltura e poi nel lavoro. In un ambiente misto uomo-robot le AI devono essere in grado di *intuire* cosa gli uomini vogliono fare e adattarsi alle loro intenzioni cooperando. Solo in un ambiente di lavoro in cui le macchine sapranno capire l'uomo e assecondare il suo agire, potremo veder rispettato l'ingegno e la duttilità umana. La macchina si deve adattare all'uomo e alla sua unicità e non viceversa.

Intellegibilità

I robot in quanto macchine operatrici funzionano secondo algoritmi di ottimizzazione. I software ottimizzano l'uso energetico dei loro servomotori, le traiettorie cinematiche e le velocità operative. Se un robot deve prendere un contenitore cilindrico da una fila di contenitori, il suo braccio meccanico scarterà verso il contenitore prescelto secondo una traiettoria di minimo consumo energetico e temporale. Un uomo, di contro, se deve prendere lo stesso barattolo si muoverà verso quello in una maniera che fa capire a chi gli è intorno cosa stia tentando di fare. L'uomo è in grado, nel vedere un altro uomo che compie un'azione, di capire cosa stia per fare in forza non dell'ottimizzazione dell'azione altrui ma della sua intellegibilità. Il modo di compiere le azioni rende l'agito intellegibile e prevedibile. Se vogliamo garantire un ambiente di lavoro misto in cui l'uomo possa coesistere con la macchina, il modo di compiere le azioni della macchina dovrà essere *intellegibile*. Dovremmo far sì che la persona che condivide con la macchina lo spazio di lavoro possa sempre essere in grado di intuire qual è l'azione che la macchina sta per compiere. Questa caratteristica è necessaria, tra l'altro, per permettere all'uomo di coesistere in sicurezza con la macchina non esponendosi mai a eventuali situazioni dannose. Non è l'ottimizzazione dell'agito della macchina la più importante finalità che deve caratterizzare i suoi algoritmi, ma il rispetto dell'uomo.

Adattabilità

Un robot dotato di AI si adatta all'ambiente e alle circostanze per compiere delle azioni autonome. Tuttavia non si tratta di progettare e realizzare algoritmi di Intelligenza Artificiale che siano in grado di adattarsi solo all'imprevedibile condizione dell'ambiente donando alla macchina una sorta di consapevolezza sulla realtà che la circonda. In una situazione di cooperazione e lavoro mista tra uomo e macchina il robot deve *adattarsi* anche alla personalità umana con cui coopera. Per esemplificare questa caratteristica proviamo a fare un esempio. Supponiamo di avere un'automobile a guida autonoma. La macchina dovrà adattarsi alle condizioni del traffico: in condizioni di intenso traffico, se la macchina non possiede degli efficienti algoritmi di adattabilità, rischia di rimanere sempre ferma perché gli altri veicoli a guida umana le passeranno sempre avanti cercando di evitare l'ingorgo. Oppure, se non fosse abbastanza adattabile, rischierebbe di causare degli incidenti non capendo l'intenzione furtiva di cambiare corsia del guidatore che ha davan-

ti. Tuttavia vi è un ulteriore e più importante adattamento che la macchina deve saper fare: quello alla sensibilità dei suoi passeggeri. Qualcuno potrebbe trovare la lentezza della macchina nel cambiare corsia esasperante o, al contrario, potrebbe trovare il suo stile di guida troppo aggressivo e vivere tutto il viaggio con l'insostenibile angoscia che un incidente sia imminente. La macchina deve *adattarsi* alla personalità con cui interagisce. L'uomo non è solo un essere razionale ma anche un essere emotivo e l'agire della macchina deve essere in grado di valutare e rispettare questa unica e peculiare caratteristica del suo partner di lavoro. La dignità della persona è espressa anche dalla sua unicità. Saper valorizzare e non mortificare questa unicità di natura razionale-emotiva è una caratteristica chiave per una coesistenza che non sia un detrimento della parte umana.

Adeguatezza degli obiettivi

Un robot è governato da algoritmi che ne determinano delle linee di condotta. Si pensi a uno di quei robot casalinghi in vendita nei negozi di elettrodomestici che in maniera autonoma pulisce il pavimento raccogliendo la polvere. I suoi algoritmi sono programmati per questo, ma il robot è programmato per raccogliere la polvere o per raccogliere il massimo della polvere possibile? Se in un ambiente di sole macchine l'assolutezza dell'obiettivo è una policy adeguata, in un ambiente misto di lavoro uomo-robot questo paradigma non sembra essere del tutto valido. Se il robot vuole interagire con la persona in una maniera che sia conveniente e rispettosa della sua dignità, deve poter aggiustare i suoi fini guardando la persona e cercando di capire qual è l'obiettivo adeguato in quella situazione. Si pensi a una situazione in cui un lavoratore e un robot cooperino nella realizzazione di un artefatto. Il robot non può avere come unica policy l'assolutezza del suo obiettivo, come se fosse la cosa più importante e assoluta, ma deve saper *adeguare* il suo agire in funzione dell'agire e dell'obiettivo che ha la persona che con lui coopera. In altri termini si tratta di acquisire, ci si perdoni il termine, una sorta di *umiltà artificiale* che, tornando all'esempio del robot aspirapolvere, consenta alla macchina di comprendere se deve aspirare tutta la polvere possibile o in questo momento aspirare solo un po' di polvere e poi tornare a compiere questa funzione più tardi perché sono sorte altre priorità nelle persone che in quel momento sono nella stanza. Si tratta di stabilire che la priorità operativa non è nell'algoritmo ma nella persona che è luogo e sede di dignità. In un ambiente misto è la persona e il suo valore unico ciò che stabilisce e gerarchizza le priorità: è il robot che coopera con l'uomo e non l'uomo che assiste la macchina.

Se queste quattro direttrici possono essere quattro dimensioni di tutela della dignità della persona nella nuova e inedita relazione tra uomo e macchina *sapiens/autonoma*, bisogna poterle garantire in maniera certa e sicura. Si devono allora sviluppare degli algoritmi di verifica indipendenti che sappiano in qualche modo quantificare e certificare questa capacità di intuizione, intellegibilità, adattabilità

e adeguatezza degli obiettivi del robot. Questi algoritmi valutativi devono essere indipendenti e affidati a enti terzi certificatori che si facciano garanti di questo. Serve implementare da parte del governo un framework operativo che, assumendo questa dimensione valoriale, la trasformi in strutture di standardizzazione, certificazione e controllo che tutelino la persona e il suo valore in questi ambienti misti uomo-robot. Non bastano standard ma servono algoritmi che sappiano valutare in maniera *intelligente* l'adeguatezza delle Intelligenze Artificiali destinate a coesistere e cooperare con il lavoratore umano. Solo in questa maniera potremmo non subire l'innovazione tecnologica ma guidarla e gestirla nell'ottica di un autentico sviluppo umano anche nell'era dei robot e delle Intelligenze Artificiali.