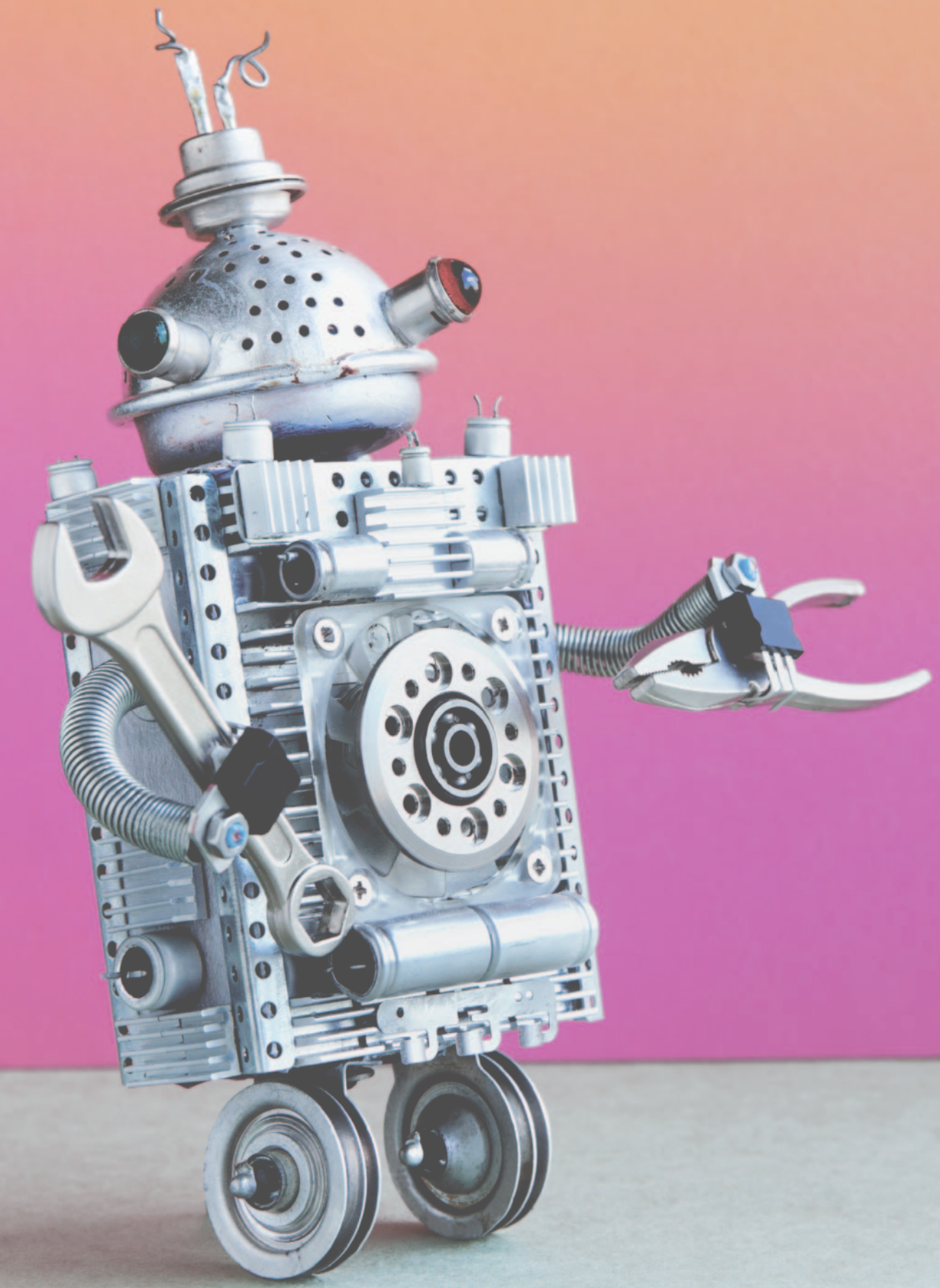


MACHINA SAPIENS
INTELLIGENZE ARTIFICIALI E SFIDE ETICHE

Paolo Benanti

Guardando alle innovazioni e trasformazioni che si stanno verificando grazie all'implementazione di nuove forme di intelligenze artificiali e alla loro diffusione sempre più ampia, dobbiamo riconoscere che questi nuovi sistemi possono cambiare radicalmente il mondo che conosciamo. In maniera provocatoria potremmo dire che oggi, per la prima volta nella nostra storia, è la macchina che ci interpella. Il contributo cerca di mettere a fuoco quali direttrici etiche possano e debbano accompagnare l'innovazione tecnologica, consapevoli che solo se sapremo includere le humanities *nella creazione di nuovi strumenti potremo sperare di non produrre, in un futuro più o meno prossimo, società disumane.*

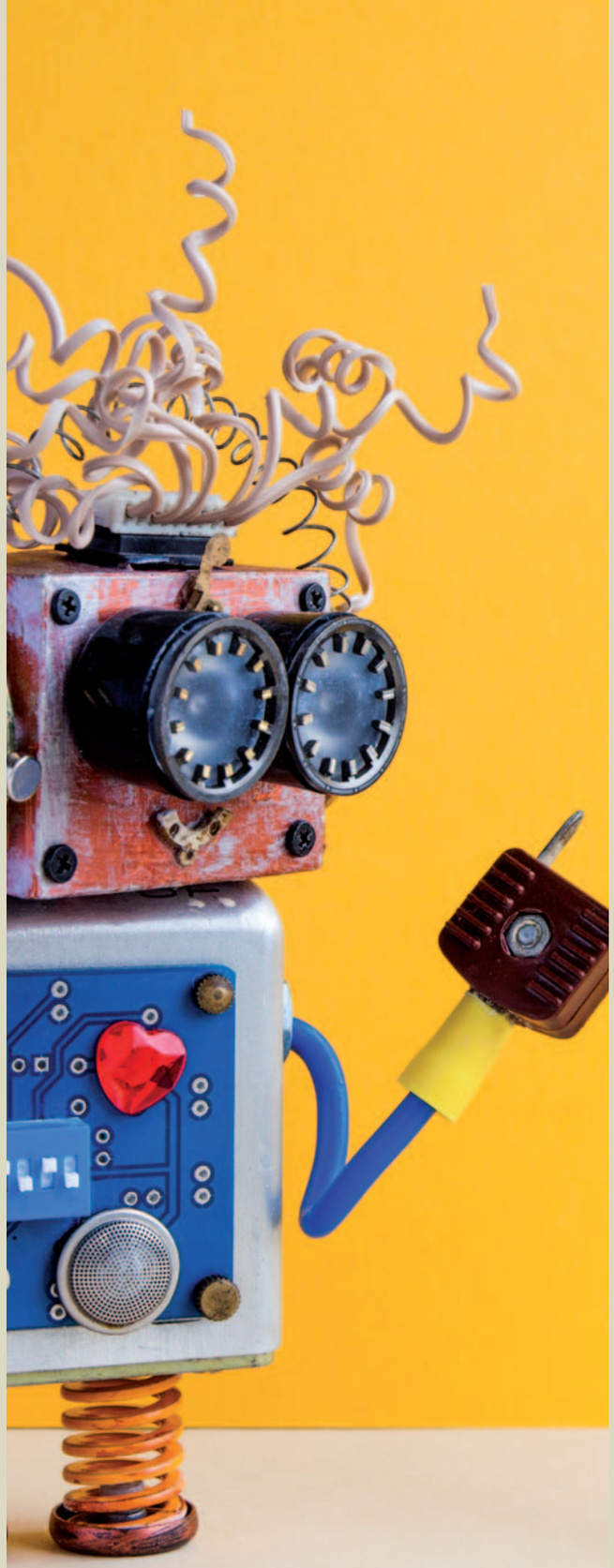
L'avvento della ricerca digitale, dove tutto viene trasformato in dati numerici, porta alla capacità di studiare il mondo secondo nuovi paradigmi gnoseologici: quello che conta è solo la correlazione tra due quantità di dati e non più una teoria coerente che ne spieghi la correlazione. Oggi la correlazione viene usata per predire con sufficiente accuratezza, pur non avendo alcuna teoria scientifica che lo supporti, il rischio d'impatto di asteroidi anche sconosciuti in vari luoghi della Terra, i siti istituzionali possibile 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 dati + significato. La Dgi è ormai uno standard operativo, in particolare nei campi in cui i dati e le informazioni sono trattati come entità reificate[1].

1. Non potendo affrontare la questione in questa sede, si rimanda a Floridi 2012.

L'evoluzione tecnologica dell'informazione e del mondo compreso come 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 radio chirurgia come il Cyberknife o ai robot destinati al lavoro a fianco 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 si diffondono in maniera sempre più penetrante. Gli smartphone di ultima generazione sono tutti venduti con un assistente dotato di AI – *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 AI, i *bot*, che saranno disponibili come partner virtuali da interrogare via voce o in *chat*, 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 AI.

L'innovazione conosce oggi una nuova frontiera: le interazioni e la coesistenza tra uomini e AI. Prima di addentrarci ulteriormente nel significato di questa trasformazione, bisogna considerare un implicito culturale che rischia di sviare la comprensione del tema. Nello sviluppo delle AI la divulgazione dei successi ottenuti da queste macchine è sempre stata presentata secondo un modello competitivo rispetto all'uomo. Solo qualche esempio: Ibm ha presentato *Deep Blue* come l'AI 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 un noto gioco televisivo sulla cultura generale *Jeopardy!*; Deep Mind, sussidiaria di Google, con *Alpha Go* ha mostrato al mondo come un computer possa sconfiggere il campione umano di Go. Le comparse mediatiche potrebbero far pensare che le AI competono con l'uomo, e che tra *Homo sapiens* e la nuova *Macchina sapiens* si sia instaurata una rivalità di natura evolutiva che vedrà un solo vincitore e condannerà lo sconfitto all'inesorabile estinzione. In realtà queste macchine non sono mai state costruite per competere con l'uomo ma per realizzare una nuova simbiosi tra lui e i suoi artefatti: (*homo+machina*) *sapiens*[2].

2. Kelly – Hamm 2016, pp. 5-42.



Non sono le AI la minaccia di estinzione dell'uomo, anche se la tecnologia può essere pericolosa per la sua sopravvivenza come specie: egli 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 apre a questo punto tutta una serie di valutazioni etiche su come validare la cognizione della macchina alla luce proprio della velocità della risposta che si cerca di implementare e ottenere. Il pericolo maggiore non viene dalle AI in sé stesse ma dalla non conoscenza di queste tecnologie e dal lasciar decidere sul loro impiego una classe dirigente assolutamente impreparata a farlo. Se l'orizzonte lavorativo del prossimo futuro – in realtà già del nostro presente – è quello di una cooperazione tra intelligenza umana e AI e tra agenti umani e agenti robotici autonomi, diviene urgente cercare di capire in che maniera questa realtà mista possa coesistere.

«PRIMUM NON NOCERE»

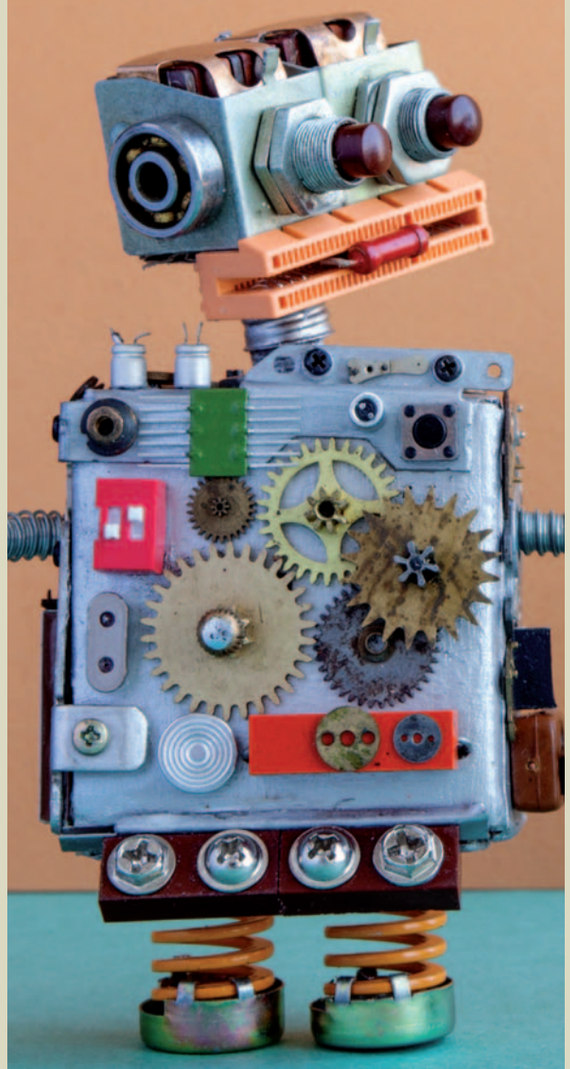
Il primo e più urgente quesito che le intelligenze artificiali pongono è quello di adattare le nostre strutture sociali a questa nuova e inedita società fatta di agenti autonomi misti. Una primissima sfida è di natura filosofica e antropologica. Le frontiere delle innovazioni, la realizzazione di queste 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 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 sul 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].

3. Benanti 2012 e 2018.
4. Il tema, per quanto affascinante, non può essere affrontato in questa sede; si rimanda a Occhetta – Benanti 2015.

Un secondo, ma altrettanto urgente tema, è definire come e in che maniera si possa garantire la coesistenza tra uomo e AI, tra uomo e robot, e per rispondere procediamo nel modo che segue. In primo luogo cercheremo di formulare una direttiva fondamentale che debba essere garantita dalle AI e dai robot, e poi si cercherà di definire cosa questi sistemi cognitivi autonomi *debbano 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 le complessità che ne derivano è stato introdotto il termine *Artificial Moral Agent* (Ama), con il quale s'individua il settore deputato allo studio della definizione dei criteri informatici volti a creare una sorta di *moralità artificiale* nei sistemi AI, inducendo 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 approssima a un incrocio e realizza che un altro veicolo, con cinque passeggeri, provenendo dal verso opposto è in traiettoria di collisione. Il primo veicolo può continuare sulla sua traiettoria e cozzare contro l'altro uccidendo i cinque passeggeri, oppure può sterzare, ammazzando un ignaro pedone. La Foot si chiedeva: è lecito sacrificare la vita di pochi per salvarne molti? 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 in situazioni estreme, qualora non fosse possibile evitare di nuocere all'uomo, avrebbe dovuto scegliere il male minore[7].

5. Wallach – Allen 2008, pp. 55-79.
6. Yudkowsky 2007.
7. Wallach – Allen 2008; Arkin 2009, pp. 37-47.



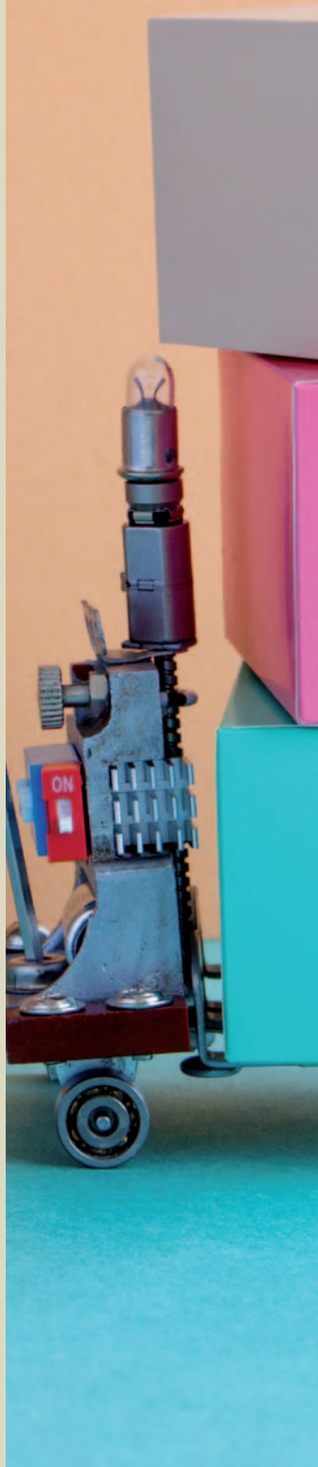
Tuttavia, la questione degli agenti morali autonomi, dell'utilizzo di robot cognitivi in un ambiente misto umano-robotico non può esaurirsi in questa direttiva primaria. Sfruttando un linguaggio evocativo potremmo dire che le macchine *sapiens*, per coesistere con i lavoratori umani, devono *imparare* almeno quattro cose. Quattro elementi che possiamo comprendere come una declinazione operativa della dignità insita nell'uomo che la macchina deve rispettare. Solo se le macchine sapranno interagire con l'uomo secondo queste direzioni, esse non solo non nuoceranno alla persona ma ne sapranno tutelare la dignità e l'inventività senza mortificarne l'intrinseco valore.

a. *Intuizione*

Quando due esseri umani cooperano normalmente, l'uno riesce ad anticipare e assecondare le intenzioni dell'altro perché in grado di intuire cosa stia facendo o cosa vuole fare. Pensiamo alla situazione in cui vediamo una persona che cammina con le braccia piene di pacchi. Istintivamente capiamo che la persona li sta trasportando e la aiutiamo rendendole il lavoro più semplice o trasportando per lei parte del fardello. 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'agricoltura e poi nel lavoro. In un ambiente misto uomo-robot le AI devono essere in grado di *intuire* cosa gli uomini vogliano fare e adattarsi alle loro intenzioni cooperando. Solo se le macchine saranno in grado di comprendere l'uomo e assecondare il suo agire potremo vedere rispettato l'ingegno e la duttilità umana. La macchina deve adattarsi all'uomo e alla sua unicità e non viceversa.

b. *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 di 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 comprendere cosa abbia intenzione di fare in forza non dell'ottimizzazione dell'azione

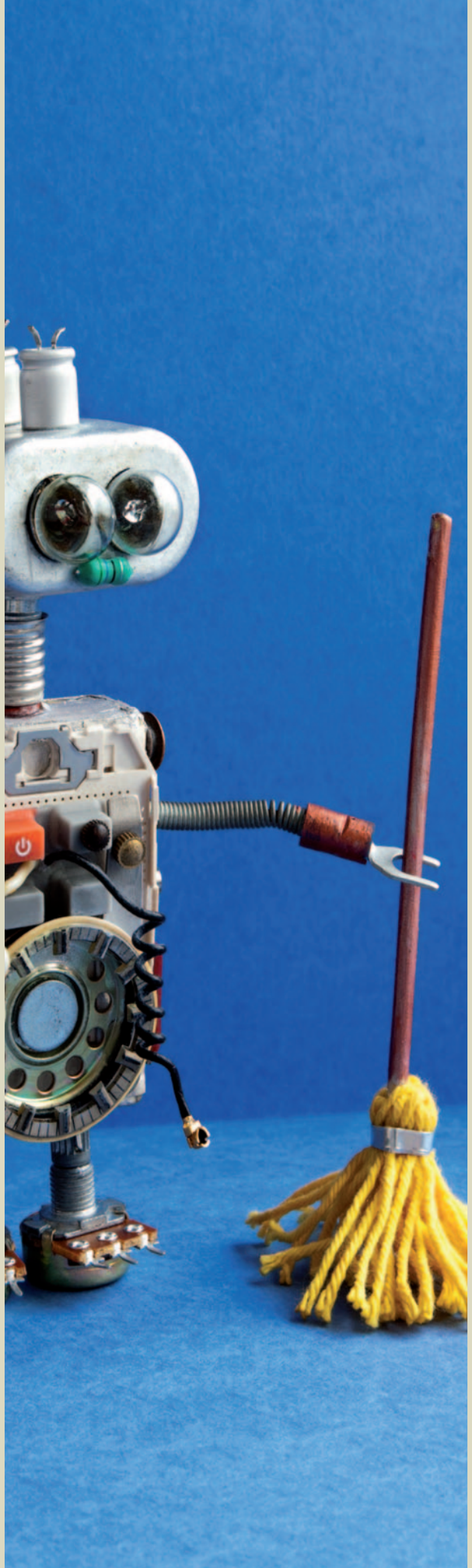


altrui ma della sua intellegibilità. Il modo di compiere le azioni rende l'agito intellegibile e prevedibile. Se vogliamo garantire un ambiente misto, in cui l'uomo possa coesistere con la macchina, il modo di compiere le azioni della macchina dovrà essere *intellegibile*. Dovremmo far sì che la persona che condivide con la macchina lo spazio vitale possa sempre essere in grado di intuire qual è l'azione che la macchina stia per compiere. Questa caratteristica è necessaria 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.

c. *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 AI 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, il robot deve *adattarsi* anche alla personalità umana. Proviamo a fare un esempio. Supponiamo di disporre di un'automobile a guida autonoma. La macchina dovrà adattarsi alle condizioni del traffico: in condizioni di traffico intenso 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 davanti. Ma 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.

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.



d. *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 il massimo della polvere possibile? Se in un ambiente di sole macchine l'assolutezza dell'obiettivo è una policy adeguata, in un ambiente misto uomo-robot questo paradigma non sembra essere del tutto valido. Se il robot vuole interagire con la persona in maniera 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 circostanza 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, solo un po' e poi tornare ad assolvere questa funzione più tardi perché sono sorte altre priorità nelle persone che si trovano 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*, bisogna poterle garantire in maniera certa e sicura. Si devono allora sviluppare 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 se ne facciano garanti. 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 negli ambienti misti uomo-robot. Si tratta di realizzare organismi che siano in qualche modo analoghi a quanto già in essere per la *Direttiva Macchine*: con l'entrata in vigore del Dpr 24 luglio 1996, n. 459, l'Italia era entrata a far parte dell'insieme degli stati europei che, avendo recepito la Direttiva, garantiscono la libera circolazione nel Mercato comune europeo soltanto alle macchine che, rispettando determinati requisiti di sicurezza, possiedono la marcatura CE di conformità, la quale può essere rilasciata dal fabbricante o cer-

tificata da un organismo verificatore ufficiale. Ora non si tratta semplicemente di fare controlli sulla sicurezza di installazione e delle condizioni operative delle macchine, ma di garantire che la componente autonoma di questi nuovi artefatti *intelligenti* rispetti sempre e in ogni condizione le direttive etiche fondamentali che abbiamo tracciato. Per cui 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.

LA GOVERNANCE DELLO SVILUPPO

Una corretta impostazione del dibattito etico dovrà quindi tener conto tutti i criteri che possano favorire o orientare verso il bene comune le innovazioni tecnologiche. Sembra importante l'intuizione della necessità di creare organismi o istituzioni che garantiscano la governance delle tecnologie legate alle intelligenze artificiali. Solo realizzando dei luoghi istituzionali dove queste forme di dialogo etico e di regolamentazione delle biotecnologie possano confrontarsi si potrà affrontare una reale ricerca oggettiva del bene. Solo se le riflessioni e il dialogo per un discernimento etico trovano una struttura politica che abbia realmente il potere di gestire le tecnologie legate alle intelligenze artificiali si può pensare a gestire, secondo una sincera e oggettiva ricerca del bene, la complessità del mondo tecnologico con tutte le problematiche connesse. L'alternativa, nella migliore delle ipotesi, è formulare proposte o valutazioni che si risolvano in un *flatus vocis* privo di efficacia storica. La gestione della tecnica-tecnologia e il suo sviluppo in un prossimo futuro richiede pertanto un approccio a carattere politico-economico. Per questo tipo di gestione si è soliti parlare di «governance»[8], termine riferito all'esistenza di un nuovo modo di organizzare e amministrare territori e popolazioni[9]. Il legame che s'instaura tra governance e sviluppo è biunivoco: da un lato apporre il termine sviluppo affianco del termine governance indica rimettere al centro del vivere sociale, come fine, la persona; contemporaneamente, indicare che lo sviluppo necessita di una governance significa assumere la dimensione etica non come un elemento giustapposto nella gestione e nell'indirizzo dell'innovazione tecnologica, ma riconoscere che questa porta una serie di domande di senso che si collocano proprio nel cuore di ogni autentico sviluppo.

8. Il termine anglosassone *governance* – dal francese antico e privo di un sostantivo corrispondente nella lingua italiana – negli ultimi venti anni è diventato popolare nel dibattito politico e accademico e tende a sostituire l'uso di *government*: cfr. Organisation for Economic Co-operation and Development 2006, p. 236.
9. La stessa definizione del concetto di *governance* ha subìto cambiamenti e integrazioni, seppure in generale si può sostenere che economisti, politologi ed esperti di relazioni internazionali lo hanno impiegato soprattutto per marcare una distinzione e una contrapposizione con *government*, inteso quale istituzione, apparato e organizzazione: cfr. Commissione delle Comunità Europee 2001.

Quindi una vera governance della tecnica-tecnologia non si fonderà su considerazioni di ordine morale che si collochino

> ai margini dello sviluppo e si [... concretizzino] nell'elaborare strumenti correttivi, sia a livello individuale, o comunque privato, sia a livello istituzionale [... ma cercherà] l'efficacia, anche dal punto di vista della produzione, di un'azione che coinvolga singoli e gruppi nella complessità di un impegno non solo settoriale, un impegno che non perda di vista la persona nella sua interezza[10].

La governance dello sviluppo si presenta, per i significati che il termine assume, come l'attuazione possibile e la corretta prassi di governo, frutto di quelle analisi etiche sul mondo della tecnica-tecnologia. La governance è lo spazio ove le considerazioni antropologiche ed etiche, in un mutuo dialogo, devono divenire forze efficaci per plasmare e guidare l'innovazione tecnologica, rendendola autentica fonte di sviluppo umano. È evidente, per la natura stessa dell'innovazione tecnologica, che una governance sarà efficace solo se si configura come momento di confronto tra le diverse competenze fornite dalle scienze empiriche, dalla filosofia, dall'etica e da ogni altra forma di sapere umano coinvolto nei fenomeni descritti[11]. Infine possiamo pensare al ruolo specifico della riflessione etica in questo processo di governance: l'etica non è chiamata a individuare direttamente soluzioni tecniche ai vari problemi ma deve rendere presente, nel dibattito, la domanda critica sul senso dell'umano che l'innovazione tecnologica media e sulle modalità che possano garantire uno sviluppo umano autentico.

10. Lacorte et al. 2004.
11. Rigobello e Latouche 2004.

BIBLIOGRAFIA

R. Arkin, *Governing Lethal Behavior in Autonomous Robots*, Chapman & Hall, Boca Raton 2009.
P. Benanti, *The Cyborg. Corpo e corporeità nell'epoca del postumano*, Cittadella, Assisi 2012.
Idem, *Le macchine sapienti. Intelligenze artificiali e decisioni umane*, Marietti, Bologna 2018.
Commissione delle Comunità Europee, *La governante europea. Un libro bianco*, «Gazzetta Ufficiale dell'Unione Europea» CCLXXXVIII (2001) 1, pp. 1-29.
L. Floridi, *La rivoluzione dell'informazione*, Codice, Torino 2012.
B. Goertzel – C. Pennachin (eds.), *Artificial General Intelligence*, Springer, Berlin 2007.
J.E. Kelly – S. Hamm, *Macchine intelligenti. Watson e l'era del cognitive computing*, Egea, Milano 2016.
P. Lacorte et al., *La governance dello sviluppo: etica, economia, politica, scienza*, Editrice AVE, Roma 2004.
S. Latouche, *Altri mondi sono possibili, non un'altra mondializzazione*, in Lacorte – Scarafine (a cura di) 2004, pp. 25-42.
F. Occhetta – P. Benanti, *La politica di fronte alle sfide del postumano*, «La Civiltà Cattolica» MMMDXCIV (2015) I, pp. 572-584.
Organisation for Economic Co-operation and Development, OECD *Economic Glossary. English-France*, OECD, Paris 2006.
A. Rigobello, *Dinamiche interne di un'etica coinvolta nella governabilità dello sviluppo*, in Lacorte – Scarafine (a cura di) 2004, pp. 43-48.
W. Wallach – C. Allen, *Moral Machines: Teaching Robots Right from Wrong*, Oxford U.P., New York 2008.
E. Yudkowsky, *Levels of Organization in General Intelligence*, in Goertzel – Pennachin (eds.) 2007, pp. 389-498.