SALVATORE AMATO
DIPARTIMENTO DI GIURISPRUDENZA
UNIVERSITÀ DEGLI STUDI DI CATANIA
samato@lex.unict.it

Gli algoritmi etici delle macchine consapevoli

Non si può *dedurre* il pensiero dalla coscienza biologica
(E. Lévinas, Entre nous. Essais sur
le penser-à-l'autre, 1991, 44)

Essere in anticipo, essere in ritardo, che inesattezze
(Ch. Péguy, *Note conjointe sur M. Descartes*
***et la philophie cartesienne*, 1924)**

1. O la *tyche* o gli dei?

Come ha fatto la materia a diventare intelligente nel corso dell'evoluzione? Perché una "provvisoria perturbazione di reazioni biochimiche"[1] si è trasformata in una persona? Un'informazione o un pensiero non "hanno massa, né quantità di moto, né carica elettrica, né solidità, e neppure una chiara distinzione nello spazio, dentro di noi o intorno a noi, o da qualsiasi altra parte"[2]. Eppure esistono. Se si esclude il disegno divino insito nella lettura "intenzionale" dell'origine del mondo, appare problematica questa trasformazione dell'indistinto della materia nel distinto delle varie forme di vita e del distinto delle varie forme di vita nel personale della coscienza. La teologia ci spiega che tutto ha un senso e che l'intelligenza, il *logos*, è l'elemento cruciale del legame tra Dio e l'uomo che si costituisce con la creazione. La risposta opposta, quella del riduzionismo materialistico che ritiene che tutto derivi da una combinazione di particelle, ci suggerisce che nulla ha senso, che ogni cosa è frutto del caso, un caso innestato sulla legge della selezione darwiniana unita alla dinamica dei sistemi complessi. Se nulla ha senso, tutto è possibile? Anche che qualcosa di materiale e tangibile come una sostanza chimica si trasformi nell'immaterialità intangibile del pensiero? Un particolare assestamento del ciclo dell'evoluzione avrebbe fatto in modo che la materia divenisse, a un certo punto, "consapevole" di se stessa: la mente è solo un prodotto casuale dell'ingegneria biologica[3]. E neppure uno dei prodotti migliori: sta assieme alla rinfusa, è frutto di anomalie e di contingenze. Basti pensare che "due tra i più complessi e creativi sistemi mai inventati dall'evoluzione – il genoma e il cervello – sono reticolari, ridondanti, palesemente imperfetti, inutilmente complicati, figli di rabberci, aggiustamenti, accrocchi e compensazioni. Nessuno dei due supererebbe un esame di ingegneria"[4].

Questo modello di riduzionismo materialistico è probabilmente il più accreditato tra gli scienziati, ma continua a lasciare aperto proprio il problema che intende risolvere. Come osserva da tempo Thomas Nagel, se i fenomeni mentali coscienti non sono spiegabili esaustivamente con nessuna delle attuali teorie fisiche, allora la tradizionale visione dell'evoluzione dev'essere rivista. Nagel non invoca un essere trascendente, ma rifiuta la mera immanenza dell'evoluzione casuale. Immagina che "nella storia della natura sono al lavoro anche principi di tipo diverso, principi sull'aumento dell'ordine che sono, per quanto riguarda la loro forma logica, teleologici piuttosto

[1] Paul G. Falkowski, *I motori della vita. Come i microbi hanno reso la terra abitabile*, trad. it. Torino, Bollati Boringhieri, 2016, p. 189.
[2] Terrence W. Deacon, *Natura incompleta. Come la mente è emersa dalla materia*, trad. it., Le scienze, Milano, 2012, p. 14.
[3] N. Humphrey, *Polvere d'anima. La magia della coscienza*, trad.it. Torino, Codice, 2013, p. 8.
[4] T. Pievani. *Imperfezione. Una storia naturale*, Milano, Codice, 2019, p. 129.

che meccanicistici"[5]. Sarebbe dunque possibile, secondo Nagel, una spiegazione teleologica, ma "non intenzionale". Una spiegazione basata su forme di auto-organizzazione biofisica che ricordano in qualche modo, aggiungerei io, il rapporto tra potenza e atto del naturalismo aristotelico. Insomma mi sembra che, nella prospettiva di Nagel, la fisica stia per ritornare all'antica e originaria *physik*ê del mondo greco, in cui costituiva un *unicum* con le scienze biologiche nella ricerca di una materia intrisa di *psych*ê[6].

E' vero che almeno su una cosa tutti gli studiosi sono d'accordo: sul fatto che non sappiamo cosa ancora non sappiamo sia sull'evoluzione biofisica che sull'origine della coscienza[7] per cui possiamo muoverci solo sul piano delle congetture. Mi pare estremamente significativo che, tra queste congetture, non sia possibile escludere il problema della teleologia e quindi di uno scopo[8]. Dopo tanti secoli e malgrado il sofisticato apparato tecnologico che abbiamo accumulato, ci accorgiamo di non essere ancora in grado di sfuggire a quel dubbio di fondo che ha iniziato a inquietare la tradizione greca da quando Anassagora ha ipotizzato che l'universo non fosse altro che il continuo aggregarsi e disgregarsi di una pluralità infinita di elementi. Ce lo ricorda uno dei pochi frammenti rimasti dell'*Ipsipile* di Euripide: "o pensieri mortali , o vano errare/ degli uomini, che fanno essere a un tempo / e la *tyche* e gli dei. Perché se c'è / la *tyche*, che bisogno degli dei? / E se il potere è degli dei, la *tyche* / non è più nulla"[9]. Il problema dell'intelligenza artificiale ripropone radicalmente questa alternativa, perché si muove proprio entro l'orizzonte in cui il finito incontra l'infinito. "Una cosa è certa: la nostra specie si trova all'inizio di qualcosa, ma non sappiamo bene che cosa. Siamo in un punto senza precedenti della nostra storia, il punto in cui neuroscienze e tecnologia si evolvono di pari passo. Il risultato di questo interscambio è destinato a cambiare ciò che siamo"[10]. E' una sfida tecnica sulla nostra capacità di progettazione, ma anche filosofica sul senso del mondo in cui viviamo e sulla nostra stessa identità.

---

[5] Th. Nagel, *Mente e cosmo. Perché la concezione neodarwiniana della natura è quasi certamente falsa*, Milano, trad. it., Milano, Raffaello Cortina, 2015, p. 9.

[6] Sono diverse le prospettive che rifiutano il riduzionismo a partire dall'ormai tradizionale analisi di David Chalmers (*La mente cosciente*, tra. it. Milano, Mc-Graw Hill,1999) per cui gli stati psichici, pur essendo occasionati da sistemi fisici posizionati nel cervello, non sono riconducibili esclusivamente alla fisica, ma a delle proprietà insite nell'universo. Una visione che risente delle riflessioni condotte già nel 1890, nel saggio *The Principle of Psycology*, da William James in cui si sosteneva che l'universo è probabilmente permeato da una "mind-dust" (polvere mentale).

Recentemente Federico Faggin ha riproposto l'idea che la realtà fisica sia costituita dalla partecipazione attiva di una gerarchia di entità coscienti che hanno una realtà interiore semantica e una realtà esteriore informatica. Per definire questa diversa dimensione propone il termine *nousym* sintesi di *nous* e *symbol*. L'incontro con Dio deriva proprio da questa particolare lettura della nostra struttura biofisica. "l'aspetto particellare per la capacità di mantenere la mia identità di osservatore nonostante sperimentassi me stesso come il mondo: il mio aspetto ondulatorio. La mia identità quindi è quel punto di vista unico con cui Uno -la totalità di ciò che esiste- osserva e conosce se stesso". (*Silicio. Dall'invenzione del microprocessore alla nuova scienza della nel consapevolezza*, Milano, Mondadori, 2019, p. 137 della versione elettronica).

[7] "…non sappiamo in che modo la «mera materia» generi l'immaterialità percepita della mente; in effetti non sappiamo nemmeno che cosa sia in realtà la mera materia, o perché esista la materia invece di un nulla totale (una domanda in qualche modo analoga a quella del perché esista la coscienza, piuttosto che un'elaborazione inconscia delle informazioni)" (N. Lane, *Le invenzioni della vita*, trad. it. Milano, il Saggiatore, 2012, p. 238).

[8] Come nota il premio Nobel per la fisica, Eugene P. Wigner, "certo è difficile credere che le nostre capacità di ragionamento abbiano raggiunto la perfezione che sembrano possedere grazie un processo darwiniano di selezione naturale" (*L'irragionevole efficacia della matematica nelle scienze naturali*, trad. it. Adelphi eBook, 2017, p. 8). E infatti arriva alla conclusione che non sia possibile formulare in maniere coerente le leggi della meccanica quantistica senza fare riferimento a una sorta di coscienza cosmica.

[9] Il frammento di quest'opera perduta è riportato in H. von Arnim *Supplementum Euripideum* fr. 63, p. 65 (Bonn, 1913). Lo ricorda C. Diano, *Teodicea e poetica nella tragedia antica* in *Saggezza e poetiche degli antichi*, Vicenza, Neri Pozza, 1968, p. 303.

[10] D. Eagleman, *Il tuo cervello. La tua storia*, trad. it. Milano, Corbaccio, 2016, p. 147.

Lo mette in luce emblematicamente l'idea di "singolarità" che, introdotta dallo scrittore di fantascienza Vernor Vinge, ha avuto la sua consacrazione scientifica in un saggio di Ray Kurzweil del 2005, *The Singularity is Near*, per indicare quella situazione in cui il ritmo del cambiamento è così estremo che la tecnologia tende ad espandersi a una velocità non più controllabile, per cui i computer saranno sempre più intelligenti fino ad arrivare a modificarsi e programmarsi in assoluta autonomia, superando l'intelligenza umana[11]. Perché non dovrebbe essere possibile? Se il nostro cervello, come qualsiasi altro composto della natura, non è altro che un ammasso di particelle regolato dalla fisica, non esiste alcuna legge fisica che escluda la possibilità di configurare queste particelle in modo da effettuare calcoli sempre più complessi: "l'introduzione di un fattore infinito in un calcolo che non è in grado di contenerlo"[12]. In effetti è proprio quello che è avvenuto nella testa degli esseri umani[13] dove dalle operazioni più semplici si è passati a quelle con le coppie di numeri e poi alle successioni con i numeri irrazionali e poi ai sistemi più avanzati con i numeri complessi e poi…

L'idea di fondo di questa visione "funzionale" è che il sorgere della vita prima e dell'intelligenza poi, nei rispettivi particolari sviluppi, fino all'emergere della coscienza siano dovuti all'organizzazione della materia, al collegamento complesso che, di adattamento in adattamento, hanno raggiunto, per caso, le singole interazioni causali[14]. Basterebbe impadronirsi dei meccanismi di questa complessità organizzativa per raggiungere i medesimi risultati. "Se ogni assone, sinapsi e cellula nervosa del mio cervello fosse sostituito da cavi, transistor e circuiti elettronici che eseguono esattamente la stessa funzione, la mia mente rimarrebbe identica. La versione elettronica del mio cervello sarebbe forse più sgraziata e ingombrante ma, ammesso che ciascuna componente neurale abbia una rappresentazione fedele in silicio, la coscienza persisterebbe"[15]. Non vi sarebbero, dunque, ostacoli teorici all'*uploading* del cervello umano, che potrebbe continuare in eterno a elaborare i propri pensieri una volta inserito in un sistema informatico, o alla costruzione di *cyborg*, di uomini-macchina anch'essi destinati all'immortalità digitale[16]. Sarebbe ora di domandarci "se un corpo da bipedi respiranti con una visione binoculare e un cervello da 1400 cc sia una forma biologica adeguata"[17]?

I vari modelli di intelligenza artificiale sono già in grado, come e spesso meglio degli esseri umani, di riconoscere i volti, giocare a scacchi e a go e a un qualsiasi videogioco, tradurre in più di sessanta lingue, condurre un veicolo, effettuare una diagnosi, progettare un esperimento e così via, incidendo sulla maggior parte delle nostre attività quotidiane. La tecnica di *machine learning* nell'apprendimento profondo, frutto dell'incontro con le neuroscienze, ha portato alla fabbricazione di nuovi supercomputer che usano reti neurali artificiali multistrato. Ciascuno strato contiene molte migliaia di unità, che comunicano tra di loro fino ad assestarsi in uno schema stabile che poi si modificherà successivamente, per effetto di altri progressivi cambiamenti adattivi, su altri livelli più elevati e così ricorsivamente in un processo di continua auto-organizzazione. In genere si

[11] R. Kurzweil, *La singolarità è vicina*, trad. it., Milano, Apogeo, 2008, p. 24.

[12] A. Greenfield, *Tecnologie radicali. Il progetto della vita quotidiana*, trad. it. Torino, Einaudi, 2017, p. 283 (versione digitale). Mi sembra sia il problema che Gödel affronta nella paginetta di equazioni che poi è stata pubblicata con il titolo *La prova matematica dell'esistenza di Dio*, Torino, Bollati Boringhieri, 2006.

[13] Ma non dei dinosauri, si potrebbe obiettare. I dinosauri hanno popolato la terra per più di cento milioni di anni, ma non hanno mai costruito non dico un computer intelligente, ma una qualsiasi macchina.

[14] In termini più complessi Deacon parla di "emergenza di proprietà entenzionali da una termodinamica di non equilibrio, e dunque un ponte tra i processi non viventi e quelli viventi" (op. cit., p. 494) e spiega l'emergere del senso di sé come una forma speciale di organizzazione dinamica: la teleodinamica (p. 528).

[15] C. Koch, *Una coscienza. Confessioni di uno scienziato romantico*, trad.it. Torino, Codice (ed. speciale per il mensile "le Scienze"), 2014, p. 163.

[16] "Can an upload be conscious?" è uno dei temi che affronta Chalmers nel saggio *The Singularity: A Philosophical Analysis* in "Journal of Consciousness Studies" 2010-17,

[17] M. O'Connell, *Essere macchina*, trad. it., Milano, Adelphi, 2018, p. 161.

tratta di schemi assolutamente inattesi da parte degli stessi programmatori. Parallelamente si sviluppano le ricerche di *riverse engineering*, di "ingegnerizzazione inversa" per replicare meccanicamente il funzionamento del cervello. Dopo la simulazione del cervello di un topo, si è arrivati a quello di un ratto, che contiene centocinquanta milioni di neuroni e presenta notevoli somiglianze con il cervello umano soprattutto nell'organizzazione della corteccia cerebrale[18]. Si sta tentando di raggiungere il medesimo risultato appunto con la corteccia cerebrale umana che riveste grande importanza nell'elaborazione del pensiero astratto. Secondo alcuni raggiungere queste forme di computazione neuromorfica non è questione di anni, ma di dollari[19].

La *Defence Advanced Research Projects Agency* (DARPA), una delle più importanti agenzie del Dipartimento di difesa degli Stati Uniti, ha investito diverse migliaia di dollari in questa corsa alla *Whole-Brain Emulation*, utilizzando chip che possono portare 5,4 miliardi di transistor, ciascuno dei quali tiene un milione di unità (neuroni) e 256 milioni di sinapsi. La Germania e il Giappone collaborano nell'utilizzo di NEST (*Neural Simulation Technology*) per sviluppare il computer K: nel 2012 questo computer impiegava quaranta minuti per simulare un secondo dell'1% dell'attività di un cervello reale, coinvolgendo 1,73 miliardi di "neuroni" e 10,4 trilioni di "sinapsi"[20].

Per quanto sia estremamente complesso, il cervello finisce per apparire, nelle prospettive aperte da questi meccanismi di simulazione, solo un fenomeno chimico, regolato dalla fisica. L' Io, il soggetto, la coscienza e le stesse singole decisioni non sarebbero altro che l'esito imponderabile delle connessioni tra le reti neuronali, la conseguenza imprevedibile delle qualità emergenti di una molteplicità di particelle biochimiche. Ci piaccia o no, saremmo solo "un fascio di neuroni" (Francis Crick), un "pattern di materia e di energia" (Raymond Kurzweil), "un computer quantistico" (Roger Penrose), "una collezione di atomi di presente" (Edoardo Boncinelli), una "macchina di carne" (Marvin Minsky), "trasferimenti chimici all'interno di 1.200 grammi di patè elettrificato" (David Foster Wallace), "una nuvola energetica di 23 watt" (Robert Lanza), "vibrazioni coerenti di proteine" (Cairns-Smith) e così via.

Sono queste visioni estreme che aprono lo spazio alla progettazione di un'area di lavoro neuronale globale (*global neuronal workspace*) che dovrebbe consentire di trasferire i meccanismi biochimici di funzionamento del cervello (*physical correlates of consciousness*) sui computer collegati alle macchine del futuro. "Se una teoria matematica della coscienza, le cui equazioni stessero su un fazzoletto, potesse prevedere correttamente gli esiti di tutti gli esperimenti che conduciamo sul cervello, potremmo cominciare a prendere sul serio non solo la teoria stessa, ma anche le sue previsioni per la coscienza aldilà del cervello, per esempio nelle macchine"[21]. La nostra intelligenza sarà la loro intelligenza? La nostra coscienza sarà la loro coscienza? Se fosse così, potremmo metterci definitivamente da parte, affidandoci a un mondo artificiale che abbiamo creato, ma che può oramai fare definitivamente a meno di noi.

La vocazione del moderno è la "pensione"? A questa conclusione, quasi profetica, era giunto Péguy, riflettendo sulle tensioni irrisolte del nostro tempo: "tutto il loro pensiero è mettere lo spirito umano in condizione di andare in pensione e di goderne. O, come dicono, di guadagnarsi la pensione"[22]. Idea che mi pare sia stata riformulata, in altri contesti e con maggior successo

[18] Sulla plausibilità scientifica di questi annunci S. Seung, op. cit., pp. 333 e ss.

[19] M. Kaku, *Fisica del futuro. Come la scienza cambierà il destino dell'umanità e la nostra vita quotidiana entro il 2100*, trad. it., Torino, Codice edizioni, 2012, p. 89. E' stata annunciata la creazione, attraverso cellule staminali neurali, di "organoidi", che altro non sono che delle strutture cerebrali estremamente simili al prosencefalo di un embrione di 10 settimane: insomma dei "mini-cervelli". Ricercatori dell'École polytechnique fédérale di Losanna prevedono di riuscire a realizzare entro il 2023 una struttura *hardware* e *software* in grado di simulare integralmente il funzionamento del cervello umano.

[20] Su questi sviluppi: Margaret A. Boden, *L' intelligenza artificiale*, trad. it. Bologna, il Mulino, 2019, cap. VII.

[21] M. Tegmark, *Vita 3.0. Essere umani nell'era dell'intelligenza artificiale*, trad. it. Milano, Raffaello Cortina, 2018, p. 376.

[22] Ch. Péguy, *Cartesio e la filosofia cartesiana*, trad.it., Roma, Studium, 2014, p. 101 (versione digitale).

mediatico, da Bill Joy in un noto articolo del 2000, *Perché il futuro non ha bisogno di noi*. "Man mano che la società e i problemi che deve affrontare diventano più complessi e le macchine diventano più intelligenti" sosteneva Joy "gli uomini lasceranno sempre più che le macchine decidano per loro, semplicemente perché le decisioni prese dalle macchine produrranno risultati migliori di quelle prese dagli uomini"[23]

2. Una scienza senza carità?

Per questi motivi non è assurdo pensare che, se da una commistione biochimica può derivare qualcosa di così radicalmente diverso come l'intelligenza umana, anche l'intelligenza umana possa produrre una macchina intelligente, anzi così intelligente da fare a meno dell'uomo. E anche questa macchina intelligente potrà produrre un'altra macchina ancora più intelligente, che potrà produrre un'altra macchina ancora più intelligente e così all'infinito… Un'intelligenza infinita è Dio[24]? Probabilmente sì, nei limiti in cui, con il concetto di infinito, il nostro linguaggio vago e confuso esprime qualcosa di imponderabile e indefinibile, ma "le analogie con determinate idee di Dio sono evidenti: l'astrazione da tutte le limitazioni è un modo per dare conto dell'impulso religioso in termini laici"[25].

Ci troviamo, insomma, di fronte al paradosso per cui il rifiuto, in nome della scienza, della creazione divina del mondo finisce per ipotizzare (o almeno a non escludere) la creazione "scientifica" di Dio. "il verbo si fa macchina"[26]? Il paradiso è "un computer potente"[27]? Restando entro un rigoroso darwinismo, Dawkins afferma che "qualsiasi intelligenza creativa abbastanza complessa da progettare qualcosa è solo il prodotto finale di un lungo processo di evoluzione graduale" [28]. Mi pare che i concetti di intelligenza creativa e di evoluzione graduale sottolineino quel delicato passaggio dal quantitativo al qualitativo, che costituisce il dato problematico di ogni forma di determinismo materialistico. Non per nulla è stata coniata l'espressione "qualia" per indicare i mattoni fondamentali della coscienza che caratterizzano tutte le sensazioni dalle più semplici alle più complesse. I "qualia" dovrebbero saldare quantità e qualità, salvaguardando il fisicalismo e insieme giustificando il funzionalismo, perché introducono all'interno del caso non proprio un fine, ma un meccanismo di "autorinforzo", "evoluzione costruttiva" o "teleodinamica" che dir si voglia[29], e quindi un perfezionamento progressivo (un af*fin*amento?) nello sviluppo: quell'ordine nel caos realizzato da un *orologiaio cieco*, per riprendere il titolo di una delle opere più fortunate di Dawkins. Il "nulla ha senso" dell'evoluzionismo darwiniano e il "tutto ha senso" della teologia si incontrano in questa apertura incondizionata alla possibilità di una dimensione meccanicamente infinita a cui casualmente tutto tende o divinamente infinita da cui tutto ha inizio e fine.

Allora possiamo credere in Dio come nel caso. L'uomo non è creato da Dio, ma finirà, casualmente, per creare Dio? "Dio non lascia residui nelle nostre provette, né tracce nelle nostre

---

[23] Wired, Aprile 2000, www.wired.com/wired/archive/8.04/html. La pensa così anche J. Attali "L'essere umano sarà allora bardato di protesi, prima di diventare lui stesso un artefatto, venduto in serie a consumatori diventati a loro volta artefatti. Poi l'uomo, divenuto ormai inutile alle proprie creazioni, scomparirà" (*Breve storia del futuro*, Roma, Fazi, 2016, p. 11).

[24] Quanto l'idea di singolarità possa alimentare varie tendenze mistiche verso l'attesa dell'oltre umano è attestato da alcuni fenomeni a carattere religioso nati negli ultimi anni. Ad esempio il *Terasem movement*. Cfr. Robert M. Geraci, *Apocalyptic AI: Visions of Heaven in Robotics, Artificial Intelligence and Virtual Reality*, Oxford, Oxford University Press, 2010.

[25] D. Foster Wallace, *Tutto e di più. Storia compatta dell'∞*, trad. it. Torino, Codice, 2005, p. 19.

[26] Riprendo il titolo della IV parte del saggio di Bodei, *Dominio e sottomissione. Sciavi, animali, macchine, Intelligenza Artificiale*, Bologna, Il Mulino, 2019.

[27] S. Seung, *Connettoma. La nuova geografia della mente*, trad. it. Torino, Codice, 2013, p. 323.

[28] R. Dawkins, *L'illusione Dio. Le ragioni per non credere*, trad. it., Milano, Mondadori, 2008, p. 32.

[29] L'espressione evoluzione costruttiva è di R. Dowkins ed è descritta nel capitolo 7: *dell'Orologiaio cieco.Creazione o evoluzione?*, trad. it., Milano, Rizzoli, 1988. Per la teleodinamica rinvio alla nota n. 14.

camere a bolla; e nemmeno rivela se stesso attraverso la logica"[30], ma pensare che abbia creato il mondo non è meno incompatibile con la fisica della pretesa di Dennett che i neuroni abbiano la sensazione di qualcosa. Se non possiamo provare l'esistenza di Dio che crea l'uomo, non possiamo neppure provare l'esistenza del caso che inventa il pensiero, tendendo a un affinamento che si spinge progressivamente oltre ogni limite. Il caso si può rifiutare per consolazione direbbe Pascal (scommettendo su Dio, se vincete, vincete tutto; se perdete, non perdete nulla), per disperazione direbbe Dostoevskji (se Dio non esiste, tutto è permesso) oppure accettare per esclusione direbbe Dawkins, che infatti si definisce un "non credente profondamente religioso"[31].

Le diverse tecnologie che convergono nella ricerca dell'intelligenza artificiale ripropongono queste domande perché ci inducono a guardare dentro di noi, domandandoci in che misura e perché siamo intelligenti, e fuori di noi, domandandoci in che modo possiamo rendere un artefatto intelligente. Trasformando la materia in pensiero, stiamo ripercorrendo il tracciato della natura o emulando il disegno divino? La prima via, quella dell'evoluzione naturale, è tutt'altro che rassicurante. Da una parte non sembra escludere in alcun modo il nostro diritto di tentare. Se lo ha fatto la natura, non vi è nulla di male che continui a rifarlo quel suo singolare artefatto che è l'uomo. Dall'altra, però, se la natura equivale al caso e se il caso è incontrollabile e imprevedibile, allora non è possibile escludere che il nostro tentativo di costruire altre intelligenze finisca per produrre intelligenze "altre" con tutte le possibili incognite che ne derivano. Possiamo dimenticare che Goethe ci presenta Mefistofele, semplicemente e radicalmente, come l'opposto in quanto tale? "Io sono lo spirito, che sempre nega (Ich bin der Geist, der stets verneint)".

Uno dei padri della nostra robotica, Gianmarco Veruggio, ci invita ad avventurarci nel mondo dell'innovazione senza dimenticare che stiamo maneggiando qualcosa di estremamente delicato. "Che cos'è l'apprendimento per i robot? Sono algoritmi, sempre più sofisticati, in grado di modificare i comportamenti delle macchine e le strategie di controllo sulla base di una valutazione delle precedenti esperienze. È qualcosa di puramente tecnologico. Se però noi dotiamo questi algoritmi di una complessità sufficiente, è matematico che il programmatore o il progettista non abbiano in alcun modo la possibilità di conoscere tutte le possibili evoluzioni del sistema. Questo non significa che il robot diventi un essere umano, ma solo che abbiamo mescolato dei componenti chimici in una provetta e non sappiamo quale risultato ne verrà fuori. Attenzione, allora, a fare gli apprendisti stregoni, perché il rischio è quello di mettere in mezzo agli esseri umani qualcosa che, in definitiva, non sappiamo come funziona"[32].

E, infatti, i timori provengono innanzitutto da scienziati o imprenditori. Per il fisico Stephen Hawking e per Stuart Russell, uno degli autori di uno dei più importanti manuali sull'intelligenza artificiale, una "intelligenza artificiale completa" potrebbe forse "segnare la fine della razza umana". Per Elon Musk, il fondatore di SpaceX, l'intelligenza artificiale è "potenzialmente più pericolosa delle armi nucleari". Per l'informatico Vernor Vinge condurrà, quanto meno, a "cambiamenti paragonabili all'insorgere della vita umana sulla Terra" determinando "una fuga esponenziale al di là di ogni speranza di controllo"[33].

Il pensiero teologico dovrebbe avere meno difficoltà nell'accettare l'intelligenza artificiale. Vi è l'ostacolo iniziale costituito dal timore che pretendere di giocare alla divinità sia un atto di arroganza. Non è l'ammonimento che ci deriva dall'ambizione di Adamo ed Eva nel giardino dell'Eden? Con gli sviluppi della scienza la cultura occidentale ha, però, da tempo messo da parte

[30] C. Koch, op. cit., p. 209.

[31] R. Dawkins, op. cit., p. 28.

[32] Citato da R. Oldani, *Spaghetti robot. Il made in Italy che ci cambierà la vita*, Torino, Codice edizioni, 2017, pp. 173-174 (versione digitale).

[33] Una rapida rassegna di queste opinioni in R. Lanza, B. Berman, *Oltre il biocentrismo. Ripensare il tempo, lo spazio, la coscienza e l'illusione della morte*, trad. it. Milano, il Saggiatore, 2017, pp. 134 e ss.

questi timori. Anche senza ripercorrere le suggestioni di Teilhard de Chardin, dal disegno della creazione emerge l'immagine di un Dio che lascia uno spazio per l'autonomia dell'uomo. Dopo aver dato un nome ai componenti dell'universo, la luce, le tenebre, il cielo, la terra e il mare, Dio attende che sia, a sua volta, l'uomo a dare un nome a tutti gli animali. Possiamo dedurne, con Rémi Brague, che l'uomo sia chiamato a dire la sua nell'attuazione del progetto divino? "La realizzazione di questo piano è affidata alla libertà umana, la quale riceve da Dio solo quello di cui ha bisogno per essere veramente tale"[34]. Emerge una visione del mondo come progetto o come impresa di cui la scienza potrebbe apparire oggi ai nostri occhi quasi il versante mistico ed escatologico perché assume sempre più l'ambizione di completare e concludere, continuando a dare un "nome" alle cose, il "dossier della creazione"[35]. Considerando la *translatio creativitatis* come il compito del nostro tempo, Sloterdijk propone provocatoriamente l'idea che "il Dio della fine veste i panni dell'onniscienza: quando il sapere giunto a compimento non può trovarsi di fronte a nessun compito nuovo dal punto di vista della creatività… (o dell' "evento"), Dio volge il suo sguardo sull'universo nella sua interezza. Egli guarda calmo tutto ciò che c'è"[36].

Se le cose stessero in questi termini, se l'intelligenza artificiale fosse solo un passaggio verso il compimento del progetto della creazione, non possiamo averne paura, perché l'intelligenza, il *logos*, è parte del rapporto dell'uomo con Dio e non può, quindi, che tendere al bene. Per questo motivo alcuni teologi sostengono che l'intelligenza artificiale, se è veramente tale, non può che condurre a Dio, "partecipare dei principi cristiani di redenzione del mondo"[37].

Il problema, semmai, è l'uomo stesso. L'ambiguità del suo essere, la presenza del male nella sua storia, il rischio, insito nella libertà, di giungere a risultati opposti rispetto a quelli attesi. Non dovremmo preoccuparci di difendere l'uomo dall'intelligenza artificiale, ma l'intelligenza artificiale dall'uomo, instillandole quegli stessi sentimenti che difendono l'uomo da se stesso: il pudore, la vergogna, la pietà, il rimorso. Le macchine saranno intelligenti quando saranno morali, ma noi possiamo programmare la moralità e, poi, quale moralità? Esiste un algoritmo della moralità per le macchine intelligenti?

E' estremamente significativa una riflessione di Sant'Agostino nel *De civitate dei* (L. IX, 20) quando riprende, anche se con non lievi forzature esegetiche, la spiegazione che Platone, nel *Cratilo* (398b), offre del termine *daimon.* "Comunque anche l'etimologia di questo nome, se consideriamo attentamente i libri della Scrittura, ci induce a una importante considerazione. Demoni (*daimones*), stando alla radice greca del nome, derivano etimologicamente da scienza. Ora l'Apostolo parlando nello Spirito Santo afferma: *La scienza gonfia, la carità costruisce*. Il detto non significa altro che la scienza giova soltanto quando si ha la carità e che senza di essa gonfia e cioè innalza a una vuota altezzosità. V'è dunque nei demoni la scienza senza la carità e quindi sono così gonfi, cioè così superbi al punto che si sono industriati perché fossero loro tributati onori divini e il servizio religioso che, come sanno, si devono al vero Dio; tuttora si dan da fare per quanto è loro possibile e con chi è possibile. Ora l'anima degli uomini gonfia della colpa dell'orgoglio non sa, perché simile ai demoni nella superbia e non nella scienza, quanto potere ha l'umiltà di Dio che si è manifestata in Cristo contro la superbia dei demoni, dalla quale era meritatamente reso schiavo il genere umano".

---

[34] R. Brague, *Il Dio dei cristiani. L'unico Dio?*, trad. it., Milano, Raffaello Cortina, 2009, p. 121. E' significativo che nel Corano sia Dio a dare il nime ad ogni cosa e l'uomo sia tenuto rispettare a questa volontà.

[35] Mi pare che questa idea sia espressa efficacemente dalle parole con cui si apre il romanzo di Ian McEwan (*Machines Like Me and People Like You*) in cui descrive le vicende del primo robot intelligente: "Era anelito religioso corroborato dalla speranza, era il sacro graal della scienza. Le nostre ambizioni in corsa su un ottovolante: un mito della creazione trasformato in realtà, un atto di mostruoso narcisismo" (trad. it. Torino, Einaudi, 2019, p. 3).

[36] P. Sloterdijk, *Dopo Dio*, trad. it. Milano, Raffaello Cortina, 2018, p. 11..

[37] Lo ricorda M. O'Connell, op. cit., pp. 232-233. Lo stesso termine di *mind uploading* potrebbe assumere sfumature mistiche se lo colleghiamo all'idea di ascensione, *up loading* come caricare in alto contrapposto al *down loading* dello scaricare in basso.

Agostino usa il termine scienza nel senso proprio della cultura del suo tempo per indicare un orizzonte estremamente vasto di forme di sapere raramente collegate ad applicazioni tecniche, ma credo sia estremamente attuale il problema di una conoscenza che tende ad aumentare la potenza a disposizione di chi la detiene senza interrogarsi sui suoi eventuali riflessi morali. Un tema che ha turbato profondamente la nostre coscienze dopo la seconda guerra mondiale [38] e che ora segna in maniera cruciale il passaggio dalle applicazioni particolari e settoriali dell'intelligenza artificiale che sono state finora realizzate alla progettazione di un'intelligenza artificiale generale capace di intervenire in modo intenzionale e in maniera indipendente su qualsiasi evento e di elaborare qualsiasi progetto.

Questo passaggio non è ancora avvenuto e non sappiamo se effettivamente avverrà[39]. Più che di intelligenza artificiale noi oggi dovremmo parlare di intelligenze artificiali, delle tante forme di rielaborazione meccanica dei dati, dal termostato a "Siri", che reagiscono con maggiore o minore autonomia agli stimoli esterni. Non ha senso porre il problema della "moralità" dei tanti robot o dei tanti programmi informatici che semplificano la nostra *routine* giornaliera. Basta definire il quadro giuridico della responsabilità per gli eventuali danni o per i difetti di costruzione e programmazione. Andare oltre potrebbe apparire assurdo: "preoccuparsi di una AI cattiva è un po' come preoccuparsi del sovraffollamento su Marte. Forse un giorno arriveremo al punto in cui l'intelligenza dei computer supererà la nostra, ma quel momento è ancora lontanissimo. A dire il vero, è lontano persino quello in cui riusciremo a creare un'intelligenza pari a quella di un porcospino. Oggi siamo fermi al verme"[40].

La sfida, per ora solo intellettuale, che ci pone la singolarità tecnologica è costituita dall'ipotesi che ci sia un momento in cui le intelligenze artificiali particolari lasceranno il passo a un'intelligenza generale in grado, in quanto tale, di elaborare da sé un sistema di regole con cui decidere la propria condotta e predisporre i propri fini. E' a quel punto che la situazione finirebbe per sfuggire al nostro controllo, perché potrebbe non dipendere da noi segnare i confini del possibile. Potrebbe non dipendere da noi stabilire il se e il quando una macchina si impadronirà del proprio destino. Un'intelligenza cognitiva non è ancora un'intelligenza generale, ma è la premessa per la sua realizzazione. Cosa succederà quando avverrà questa conquista della singolarità tecnologica? Un'intelligenza generale artificiale diventerà anche un agente morale artificiale? Queste macchine avranno una propria morale o accetteranno la nostra morale? E noi saremo in grado di elaborare una morale universale e poi di tradurre questo codice morale in un codice macchina?

Cosa intendiamo quando parliamo di agente morale artificiale? Fino a che punto la capacità computazionale è intelligenza? Fino a che punto la capacità di auto-organizzazione è coscienza? Fino a che punto la capacità di selezione è libertà? Intelligenza, coscienza, libertà sono componenti fondamentali dell'identità umana. Di nessuna di queste qualità sappiamo dare una definizione chiara ed univoca, ma siamo tutti d'accordo nel ritenere che siano elementi fondamentali della morale, della capacità cioè di percepire ed elaborare un sistema di valori e di adeguarsi ad essi. L'umiltà e la carità, le doti che secondo Agostino mancano al *daimon*, sono le condizioni, le virtù che ispirano l'agire morale perché implicano il senso del limite attraverso l'accettazione e il rispetto

[38] Si vedano, ad esempio, le riflessioni sulla scienza come "potenza senza etica" di Romano Guardini, *Il Potere*, tr. it. Brescia, Morcelliana, 1954, pp. 129-130.

[39] Ad esempio, due pensatori autorevoli come John Searle e Roger Penrose sostengono che le macchine non saranno mai capaci di pensare come l'uomo. Il primo sviluppa la propria argomentazione in chiave meramente logica. Il secondo analizzando le implicazioni della fisica quantistica.

[40] Su questo aspetto si veda H. Fry, *Hello word. Essere umani nell'età della macchine*, trad. it., Torino, Bollati Boringhieri, 2018, p. 21. In effetti è stato mappato il sistema nervoso del verme *C. elegans*, uno dei vermi più semplici esistenti in natura, composto da circa 300 neuroni, connessi però da più di 7000 sinapsi. Per quanto possa essere semplice, il sistema nervoso di C. elegans è così complesso che nessuno è mai riuscito a costruirne un modello integrale al computer.

dell'altro. L 'assenza di limiti sfocia nell'arbitrio di una potenza fine a se stessa. Per questo il *daimon* appare ad Agostino gonfio, altezzoso, superbo. Un'intelligenza artificiale così avanzata da decidere il futuro in base alle proprie preferenze avrebbe il senso del limite? Perché mai dovrebbe trovare questo limite, se deriva da una progettazione che implica un continuo auto-potenziamento? Senza un senso del limite, senza carità dovremmo dire con Agostino, come potrebbe sviluppare quel rispetto dell'altro da sé, e quindi degli esseri umani e anche di tutte le forme viventi e della stessa natura, che esclude derive drammaticamente impreviste per il nostro futuro?

Gli effetti di una intelligenza senza carità sono, in qualche modo, assimilabili al concetto di "*default*" tecnologico a cui ricorre Bostrom per illustrare il rischio di svolte insidiose, di guasti di programmazione, di istanziazioni inattese di una "superintelligenza" che diviene progressivamente "aliena", vale a dire sviluppa una potenza "non allineata" ai nostri modelli e quindi indipendente dagli esseri umani, se non addirittura "malevola". Un rimedio tecnologico potrebbe essere costituito dalla previsione di specifiche forme di controllo o dalla programmazione di rigidi livelli di *satisficing*, per cui, raggiunta una certa soglia di pericolo o realizzato un certo risultato, la macchina viene fermata o si arresta da sé. Tuttavia è difficile escludere un dubbio di fondo: perché mai un sistema più intelligente di noi e che si auto-programma in base a logiche proprie, dovrebbe accettare il nostro controllo esterno e attestarsi entro i nostri *input* di soddisfazione? Ipotizzando che uno dei primi effetti di un'intelligenza complessa sia la coscienza di sé, questa consapevolezza non potrebbe non condurre alla ricerca dell'auto-conservazione e quindi al rifiuto di qualsiasi programma che tenti di porre limiti al suo funzionamento. In macchine coscienti e quindi tendenzialmente "emotive", la paura di subire quella forma estrema di danneggiamento che è l'arresto delle funzioni (e quindi la fine della "vita" meccanica) dovrebbe essere un sentimento altrettanto fondamentale quanto la nostra paura della morte[41]. E allora?

3. *Default* tecnologico o *default* esistenziale?

In base a queste considerazioni, se l'origine di tutto fosse il caso, una facile convivenza tra intelligenza umana e intelligenza artificiale è meno probabile dell'esplosione di una potenza fine a se stessa o meglio così racchiusa nella propria "singolarità" da tendere alla distruzione di ogni forma di differenza. Anche se l'intelligenza artificiale finisse per adeguarsi a un sistema di valori, divenendo, come nell'evoluzione umana, un'intelligenza morale, perché dovrebbe sviluppare la nostra stessa moralità e comunque un sistema morale di cui sia parte integrante il rispetto per gli esseri umani[42]? Proprio ragionando nei termini di un mero adattamento evolutivo la costruzione di un sistema di valori che favorisce la cooperazione, supplendo alle carenze individuali, ha senso solo per limitare le tendenze all'aggressività e alla competitività degli esseri umani, ma non svolge alcuna funzione nel potenziamento delle interconnessioni di un sistema digitale. La conclusione non cambia anche a voler accettare integralmente la concezione darwiniana secondo cui il senso morale è in stretta continuità con i meccanismi biologici e quindi non vi sarebbe nessuna eccezionalità nella

[41] "L'evoluzione l'ha selezionata (la paura) per un motivo, quello di evitare determinati pericoli: dovessero anche essere fatti d'acciaio, i robot dovranno temere le circostanze che potrebbero danneggiarli, evitando così di precipitare dall'alto o di intrufolarsi dove stia divampando un incendio. Un robot privo di paura è perfettamente inutile, se finisce per distruggersi" (M. Kaku, *Il futuro della mente. L'avventura della scienza per capire, migliorare e potenziare il cervello*, trad. it. Torino, Codice, 2014, p. 283). Nel romanzo di McEwan, *Macchine come me*, la prima cosa che compie Adam, il robot umanoide, è disattivare l'interruttore di emergenza che consentiva di arrestarne le funzioni.

[42] Non dobbiamo necessariamente pensare a un' intelligenza "malevola", ma questo non esclude i rischi per il nostro futuro. Tegmark nota che "l'analogia con il trattamento che noi umani riserviamo alle forme di vita inferiori non è incoraggiante: se decidiamo di costruire una diga idroelettrica e si dà il caso che nell'area destinata a essere inondata vivano delle formiche, la diga verrà costruita ugualmente, e non perché le formiche ci siano particolarmente antipatiche, ma solo perché siamo concentrati su obiettivi che consideriamo più importanti" (*L'universo matematico. La ricerca della natura ultima della realtà*, trad. it. Torino, Bollati Boringhieri, 2014, p. 371 della versione digitale)

natura umana. L'intelligenza artificiale è meccanica e non biologica; gli algoritmi non seguono i percorsi della genetica.

Per questo motivo penso che, tra Dio e il caso, gli sviluppi dell'intelligenza artificiale dovrebbero destare meno timori di una catastrofe esistenziale in una prospettiva teologica rispetto al quadro offerto dal materialismo riduzionistico. Se all'origine del tutto troviamo Dio e non il caso, possiamo ipotizzare un percorso comune tra intelligenza umana e intelligenza artificiale perché non vi è che un *logos*. Sembra riproporsi l'idea di fondo del pensiero di Bergson che l'evoluzione giunga a compimento con l'assimilazione tra l'umano e il divino. *Le due fonti della morale e della religione* si conclude, evocando una "umanità che geme (*gémit*), semischiacciata (*à demi écrasée*) dal peso del progresso compiuto. Non sa abbastanza che il suo avvenire dipende da lei. A lei di vedere prima di tutto se vuole continuare a vivere; a lei domandarsi poi se vuole soltanto vivere, o fornire anche lo sforzo perché si compia, anche sul nostro pianeta refrattario, la funzione essenziale dell'universo, che è una macchina destinata a creare delle divinità (*une machine à faire des dieux*)"[43].

E' un atto di fede. Un atto di fede che implica, come direbbe Agostino, la fiducia in una scienza che si fonda sulla carità e non sulla potenza; una scienza che guarda il mondo con l'umiltà di chi avverte i propri limiti e non con l'ambizione di chi vuole aumentare il proprio potere. Anche Bostrom, in fondo, sembra auspicare, dinanzi al rischio di *default* tecnologico, un atto di fede, quando, non so con quanta ironia, descrive, tra gli altri, l' "approccio dell' Ave Maria" ("*Hail Mary" approach*). Prospetta, infatti, la speranza *new age* che da qualche parte nell'universo esista o esisterà una civiltà particolarmente evoluta "capace di gestire bene l'esplosione dell'intelligenza e che i suoi valori arrivino a coincidere in misura significativa con i nostri. Potremmo quindi cercare di costruire la nostra IA in modo che sia motivata a fare quello che queste superintelligenze vogliono che faccia. Il vantaggio è che potrebbe essere più facile rispetto a costruirla in modo che sia motivata a fare direttamente ciò che vogliamo noi"[44].

Parlo di un atto di fede in Dio come unico antidoto, secondo il suggerimento di Euripide, al dominio del caso, ma penso anche a un atto di fede nell'uomo, nella sua capacità di cercare Dio, mantenendo e alimentando questo rapporto con la carità e l'umiltà. Non ho nulla da dire sull'atto di fede in Dio, resta nell'orizzonte insondabile delle scelte personali, ma vorrei riflettere sul secondo tema, sulla fiducia negli esseri umani "semischiacciati dal peso del progresso compiuto". Abbiamo tanti elementi da valutare. E sono tutt'altro che positivi. Gli studi sull'intelligenza artificiale si sono, infatti, sviluppati nei due settori della nostra società dove è maggiore l'aggressività e l'irresponsabilità, il settore militare e il settore commerciale, per un giro di affari che si stima possa raggiungere, tra qualche anno, i 15 miliardi di dollari[45]. Sono campi, quelli degli armamenti e dei mercati, in cui è più facile raccogliere investimenti, sfruttando la competizione politica o commerciale, ma anche i campi in cui sono minori i freni morali. Sono, inoltre, settori in cui prevale la logica del segreto, il segreto militare e il segreto industriale. Si sottraggono, quindi, strutturalmente a qualsiasi forma di controllo istituzionale[46].

La teoria del *dual use*, che sottolinea le numerose ricadute civili di innegabile utilità sociale di molte scoperte effettuate per scopi militari, tende a rassicurarci sul prevalere alla lunga dei possibili benefici. Continuiamo, però, a muoverci sul filo sottile delle combinazioni fortunate. E' vero che l'esplosione della bomba atomica, pur precedendo di molti anni la costruzione delle prime

[43] H. Bergson, *Le due fonti della morale e della religione*, trad. it., Torino, UTET, 1979, p. 590.

[44] N. Bostrom, *Superintelligenza. Tendenze, pericoli, strategie*, trad. it., Torino, Bollati Boringhieri, 2018, p. 300.

[45] *Intenational AI ethics panel must be indipendent*, editoriale di Nature 21 agosto 2019. Solo i maggiori paesi europei e il Giappone pongono con decisione l'urgenza di affrontare i probabili profili critici dell'intelligenza artificiale con la netta opposizione di Cina, Stati Uniti e Russia, le maggiori potenze militari.

[46] Dedica particolare attenzione al problema dell'impatto sociale lo studio di Illah Reza Norbakhsh, *Robot tra noi. Le creature intelligenti che stiamo costruendo*, trad. it. Torino, Bollati Boringhieri, 2014.

centrali, non ha impedito l'utilizzazione pacifica dell'energia nucleare, ma proprio questa sfasatura cronologica tra l'immediatezza dell'impiego militare e i suoi successivi riflessi civili desta particolari preoccupazioni. Se crediamo all' eventualità della crescita esponenziale prevista dalla teoria della singolarità, non corriamo il rischio di arrivare, questa volta, troppo tardi nel far prevalere i "benefici" sui "malefici"? Del resto, il termine "singolarità" è preso in prestito dalla matematica e dalla fisica dove indica quel punto "nel quale e aldilà del quale le normali leggi della causalità e della misurabilità smettono di funzionare" [47]. Insomma segna il momento in cui si arrestano le nostre capacità cognitive.

I limiti tecnologici ci pongono, inoltre, di fronte a una sfasatura nei processi di programmazione che riflette un inquietante squilibrio esistenziale. Non avevamo bisogno dell'intelligenza artificiale per scoprire che, per gli esseri umani, è più facile compiere il male piuttosto che perguire il bene. L'unica cosa in cui gli uomini sono assolutamente uguali, affermava Hobbes, è nella capacità di dare la morte. "Quanto alla forza corporea il più debole ne ha a sufficienza per uccidere il più forte, sia ricorrendo a una macchinazione segreta, sia alleandosi con altri che corrono il suo stesso pericolo"[48]. Ma la difesa della vita? Quanti sono veramente in grado di praticare la carità e l'amore per il prossimo? Non ci sono in giro tante Teresa di Calcutta e tanti Francesco d'Assisi. In compenso non c'è posto nel mondo in cui le carceri non siano sovraffollate.

L'utilizzazione dell'intelligenza artificiale per scopi militari ha confermato e aggravato questa prospettiva. Sono impiegate in maniera sempre più massiccia armi robotizzate e in particolare i droni. I progetti più recenti tendono a ridurre al minimo, se non ad escludere del tutto, il controllo umano nel colpire i vari obiettivi. E' il modo per ottenere la reazione più rapida ed efficace nel contrastare il nemico. Non certo per evitare incidenti o effetti imprevisti. E' facile programmare una macchina per "uccidere", per distruggere un oggetto. Un oggetto, appunto. Gli esseri umani, per un drone, sono solo uno dei tanti algoritmi che sono programmati a decifrare, sono solo l'*input* di un *feedback*. Più questa retroazione è automatica e immediata, più è affidabile[49].

Al contrario è estremamente difficile, se non impossibile, programmare un drone non dico a a "perdonare", espressione troppo "umana" per essere applicata a una macchina, ma almeno a distinguere i combattenti da i non combattenti, chi aggredisce da chi intende fuggire, chi attacca da chi si difende, chi tende un agguato da chi sta per arrendersi. Insomma tutto quell'insieme di cautele e regole che il diritto internazionale ha elaborato per evitare abusi o eccessi, cercando di imporre il rispetto di simmetria, proporzionalità, responsabilità, trasparenza nelle azioni e nelle reazioni militari, oltre alla tutela dei civili. Se la storia ci insegna quanto sia illusorio pretendere di elaborare e far rispettare regole che rendano meno cruenta la guerra, perché mia dovremmo riuscire a circondarci di robot killer ossequienti verso i trattati internazionali? E' alla nostra portata l'automazione meccanica nel dare la morte, ma non l'opposto. Non lo è ancora e forse non lo sarà mai. Il male, il male che si concretizza nel dare la morte, ha una dimensione oggettiva, riconducibile entro schemi definiti e realizzabile attraverso specifici strumenti. Come tradurre, invece, in impulso elettronico il rispetto per il nemico e la pietà per gli sconfitti?

Questo squilibrio esistenziale nella programmazione militare dell'intelligenza artificiale esaspera il rischio dell'imporsi di una scienza senza carità: pura potenza, integrale meccanica al servizio della forza. Del resto non c'è minimamente paragone tra quanto investiamo in tempo e risorse per aumentare la potenza dei mezzi informatici rispetto alla ricerca dei sistemi per

[47] M. Hanlon, Eternità. Il nostro prossimo miliardo di anni, trad. it. Milano, Edizione speciale per il mensile Le Scienze, 2011, p. 156.

[48] Th. Hobbes, *Leviatano*, trad.it. Roma-Bari, Laterza, 1989, p. 99

[49] Sono in avanzata sperimentazione droni dotati di videocamere per la visione artificiale e di *sofware* per l'individuazione dei volti: gli algoritmi che si sviluppano sono in grado, dopo il confronto con l'obiettivo previsto, di far fuoco senza nessun impulso umano.

disinnescare questi meccanismi in caso di *default*. Eppure, come nota Tegmark, se esiste anche l'1 % di probabilità che si sviluppi la singolarità sarebbe ragionevole spendere almeno l'1% del PIL per analizzare il problema e decidere cosa fare. "Nel 2013 l'Union of Concerned Scientist , una delle più grandi organizzazioni che si occupano di rischi esistenziali, ha raccolto circa venti milioni di dollari: nello stesso periodo, solo negli Stati Uniti è stata spesa una somma cinquecento volte più grande per la chirurgia estetica, mille volte più grande per dotare le truppe di aria condizionata, cinquemila volte più grande per l'acquisto di sigarette e circa 35.000 volte più grande per spese militari… Perché l'uomo è così miope?"[50]

Esistono anche i robot destinati all'assistenza (*carebots* o *socially assistive robots*), usati in casa, negli ospedali o in altri luoghi per dare aiuto, provvedere alla cura, offrire conforto ad ammalati, disabili, bambini, anziani, o persone altrimenti vulnerabili. Le particolari capacità cognitive di cui sono dotati consentono già a questi robot di "conoscere" le abitudini del loro assistito, compreso il numero, i tempi e il dosaggio dei farmaci da assumere regolarmente. Aiutano a rialzarsi in caso di caduta e, se necessario, sono in grado di chiamare un'ambulanza o rispondere al telefono. Studi recenti asseriscono che robot umanoidi, addestrati a riconoscere i volti e analizzare il linguaggio del corpo, possono rivelarsi efficaci nel trattamento dell'autismo, sfruttando l'empatia che si instaura per abituare i bambini affetti da tale sindrome a una migliore interazione con il mondo esterno. Per non dire che siamo continuamente assistiti da orologi intelligenti, lampade intelligenti, auto intelligenti, frigoriferi intelligenti, cestini intelligenti. Potremo anche cercare particolari forme di "assistenza" nei *sexbot* che stanno per essere introdotti sul mercato con varie personalità per appagare i gusti più vari[51]: Frigid Farah, Wild Wendy, S&M Susan, Young Yoko e Mature Martha, con i corrispondenti Roxxy, Rocky ecc. della versione maschile.

Tutti questi congegni migliorano innegabilmente la qualità della nostra vita, ma nessun programmatore penserà mai che potranno condurre al "bene". Se si dovesse cercare il "bene", nel senso della migliore assistenza possibile, a quale modello etico si dovrebbe fare riferimento per sviluppare l'intelligenza più adatta? All'etica delle virtù, all'etica della cura, all'approccio delle capacità[52]? Dobbiamo ipotizzare un robot virtuoso, un robot affettuoso o un robot abile? Emerge il timore che la pretesa di inserire il bene, ammesso che sia possibile, come finalità del sistema di funzionamento, potrebbe determinare applicazioni così rigidamente integraliste da costituire un pericolo ancora più grave della programmazione per fini militari. Il bene di chi e per chi? Quante volte è stato seguito, nella storia, l'appello di Mounier: "non bisogna amare Dio *contro* nessuno" [53]? Purtroppo lo squilibrio esistenziale insito nella natura umana ha tradotto, continuamente e sistematicamente, la ricerca del bene in uno strumento di aggressione e contrapposizione. Non vi sono dubbi sulle modalità da seguire per dare la morte e sulla necessità di riuscirci nel modo più rapido e radicale. E', invece, estremamente controverso quali siano le misure per realizzare una società "bene ordinata", come direbbe Rawls.

Nel film *Io, robot*, tratto dai racconti di Isaac Asimov, si prende in esame un problema del genere. VIKI (*Virtual Interactive Kinetic Intelligence*) è stato progettato per assicurare il perfetto funzionamento della società, proteggendo l'umanità da ogni pericolo. I problemi emergono proprio da questo compito. Qual è il peggior nemico dell'umanità? Quando VIKI si trova a rispondere a questa domanda, i suoi algoritmi non lasciano alternative: il peggior nemico dell'umanità è

[50] Tegmark, *L'universo matematico*, op. cit., pp. 374-375.

[51] Le possibili implicazioni etiche e giuridiche sono esaminate nel volume a cura di Adrian D. Cheok e D. Levy, *Love and Sex with Robots*, Cham (Switzerland), Srpinger, 2018.

[52] S. Vallor, *Carebots and Caregivers: Sustaining the Ethical Ideal of Care in the Twenty-First Century* in "Philosophy & Technology", 2011-24, pp. 251 e ss.

[53] E. Mounier, *Il pensiero di Charles Péguy*, trad. it. Bari, Ecumenica ed., 1987, p. 176.

l'umanità stessa. Per ubbidire alla direttiva per cui è stato programmato, VIKI deve privare il genere umano della libertà, costringendolo a un "benevolo" regime di schiavitù.

Questa sfiducia nell'uomo è riproposta anche da alcuni aspetti del rapporto tra neuroscienze e intelligenza artificiale. Da una parte, abbiamo la conferma che il cervello è certamente l'organismo biologico più complesso che esista in natura e forse nell'universo. La materia, infatti, sembra non possa più nasconderci nulla[54]: microscopi, telescopi, spettroscopi, macchine sequenziatrici, centrifughe, *very large telescope*, anelli a collisione di adroni, tomografie, ecografie, scansioni a ultrasuoni 3D ci consentono di penetrare, allo sesso tempo, l'infinito e l'infinitesimale. Del cervello, invece, più sappiamo e meno comprendiamo: con i suoi 100 miliardi di neuroni ha più stelle della nostra galassia e i neuroni interagiscono gli uni con gli altri attraverso un numero di contatti (sinapsi) che varia da 10.000 a 150.000. Cento milioni di miliardi di connessioni che sono in funzione contemporaneamente. 10^{14} (100 bilioni) di contatti o istruzioni al secondo, mezzo miliardo per millimetro cubo, considerando che ogni sinapsi è più o meno grande come un batterio. Gerald Edelman spiega che il semplice atto di enumerare le sinapsi, dedicando un secondo a ciascuna di esse, richiederebbe 32 milioni di anni.

Proprio questa complessità dovrebbe indurre a escludere l'ipotesi che sia possibile progettare artificialmente un'intelligenza simile a quella umana. Eppure il materialismo riduzionistico presuppone, come abbiamo visto, che alla base di tanta complessità non vi sia altro che un fenomeno chimico, regolato dalla fisica, geneticamente condizionato, farmacologicamente condizionabile e meccanicamente riproducibile. Il cervello aumentato, di cui le neuroscienze illustrano la varietà delle articolazioni, si risolve in un uomo "diminuito", i cui impulsi non sono altro che l'effetto delle connessioni tra le reti neuronali, le conseguenze dell'efficienza dei neurotrasmettitori, l'esito della connessione tra i geni[55]. Proprio per questo potremo aspirare a simulare il funzionamento del cervello su un computer e a pretendere di racchiudere la mente in un *file.*

Un algoritmo, ci suggeriscono le ipotesi sull'intelligenza artificiale, dovrebbe essere in grado di produrre la… libertà, o singolarità tecnologica che dir si voglia, nei limiti in cui attiva il programma di una macchina capace di apprendere da sé e di assumere decisioni assolutamente autonome e imprevedibili, eppure, proprio noi che costruiamo questa macchina, proprio noi che elaboriamo l'algoritmo della libertà, non saremmo liberi, ma il frutto casuale della combinazione di geni e neuroni che decidono per noi. I giuristi affrontano il problema delle macchine intelligenti, cercando di configurare una sorta di personalità e dignità "numerica" [56], in analogia con la personalità e dignità umana, proprio quando vacillano i presupposti (libertà e responsabilità) dell'attribuzione della personalità giuridica agli esseri umani. Queste contraddizioni sono l'esito degli sviluppi più raffinati della nostra cultura, eppure mettono in crisi la nostra cultura. Siamo sospesi tra determinismo naturalistico e indeterminazione artificiale, per cui la libertà appare un artificio necessario, se osservata nella prospettiva dei robot cognitivi, e un concetto vuoto, se osservata nella dimensione del metabolismo cerebrale.

Si ripropone, sotto altre forme, quella sottile contraddizione che abbiamo già registrato in precedenza, la ragione spinge paradossalmente verso esiti irrazionali. Progettiamo il futuro con il timore di restare senza futuro, di avviarci verso un "pensionamento" forzato. Studiamo i modi per costruire macchine "libere", per apprendere che siamo noi a non essere liberi. Costruiamo meccanismi sempre più complessi per manipolare la natura per correre, poi, il rischio di essere a

[54] Cfr. H. Nomotny, G. Testa, *Geni a nudo. Ripensare l'uomo nel XXI secolo*, trad. it., Torino, Codice, 2012, p. 70.

[55] Penso al bel libro di M. Benasayang, *Il cervello aumentato, l'uomo diminuito*, trad. it., Trento, Erickson, 2015.

[56] A. Bensoussan, *La protection de la dignité humaine s'étend au champ du numérique*, Le Huffington Post, 6 giugno 2014.

nostra volta manipolati da questi meccanismi[57]. E che dire del primato, nella fattibilità industriale, degli algoritmi per la morte rispetto agli algoritmi contro la morte? Ogni incremento della conoscenza con tutte le sue svariate applicazioni tecniche nel campo dell'intelligenza artificiale ci pone sistematicamente di fronte alla fragilità umana e ai rischi di una scienza senza carità[58]. Potremmo dire che la carità costituisce per la morale quello che la teleologia, nel senso di Aristotele o di Nagel, rappresenta per la fisica: la fiducia o forse solo la speranza sul fatto che ci sia un senso tanto nei meccanismi naturali quanto nelle azioni umane. E' difficile ipotizzare che la carità, anche solo nella versione più immediatamente laica di senso del limite e rispetto dell'alterità, sia il frutto del caso. Non ci resta, allora, che affidarci all'altro corno del dilemma di Euripide?

[57] Fino a che punto l'ingegnerizzazione inversa non finisce per produrre un'inversione antropologica? Mi pare estremamente significativo l'aneddoto che si racconta sul dialogo tra due "padri" dell'intelligenza artificiale. Marvin Minsky, il fondatore di uno dei primi laboratori sull'intelligenza artificiale annuncia orgogliosamente: "le faremo parlare e camminare! Faremo di loro entità coscienti!" annuncia Minsky. "Farete tutto questo per i computer? Fantastico. Ma per le persone cosa pensate di fare?" domanda Doug Engelbart, l'inventore del mouse. Racconta questo aneddoto, senza garantirne l'autenticità, S. Seng, op. cit., p. 216.

[58] Per renderci conto di quanto sia cruciale questo tema, si può notare che, anche se in chiave completamente diversa da quella di Agostino, Morin, negli anni '80 del secolo scorso, aveva posto al centro delle sue riflessioni su metodo e complessità il problema di una "scienza con coscienza" (è anche il titolo del saggio tradotto in italiano da Franco Angeli, Milano, 1982). Recentemente è stata avanzata la necessità di una "scienzosofia", che sviluppi "la saggezza morale riguardo al perseguimento delle ricerca scientifica e le sue applicazioni pratiche", da I. Persson e J.Savulescu (*Inadatti al futuro. L'esigenza di un potenziamento morale*, trad. it. Torino, Rosenberg & Selliers, 2019, p. 154).