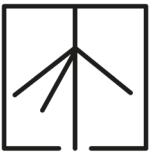
PAIDEIA

CONCETTI E SIGNIFICATI DELLA STORIA DEL PENSIERO

5

PAIDEIA

CONCETTI E SIGNIFICATI DELLA STORIA DEL PENSIERO



La possibilità che il fraintendimento, orizzonte costante della facoltà di comunicare, anziché essere un proficuo e inedito "impensato" o una piacevole divagazione, costituisca il motivo principale del radicamento di pregiudizi e preclusioni all'interno del "senso comune", è un pericolo che l'umanità non può concedersi nell'attuale momento storico.
Questa sezione della collana « Paideia », già impegnata nella promozione del dialogo interculturale e nell'innovazione della didattica delle scienze umane, nasce con l'intenzione di percorrere la storia del pensiero, per individuare concetti e significati adottati comunemente nella sfera della vita quotidiana che necessitano di una chiarificazione semantica che sia univoca, ma non monolitica. Solo a partire da tale chiarificazione è possibile lastricare la strada verso una eventuale e condivisa "risemantizzazione", quale ineludibile progetto per la futura casa comune.

Classificazione Decimale Dewey:

174.90063 (23.) ETICA DI ALTRE PROFESSIONI E OCCUPAZIONI. Elaborazione dei dati. Scienze degli elaboratori. Intelligenza artificiale

INTELLIGENZA ARTIFICIALE

SPUNTI DA FILOSOFIA E PSICOLOGIA
KANT, HUSSERL, LACAN…

Carteggio tra

ANTONIO **CAPUZZO**, PAOLA **GRANDI**

CON UNO STUDIO DI FATTIBILITÀ PER L'ADDESTRAMENTO
E LA SOCIALIZZAZIONE DI UN ROBOT

aracne

ISBN
979-12-218-1347-0

PRIMA EDIZIONE
ROMA 10 GIUGNO 2024

Un pericolo potrebbe venirci proprio dalle nostre creature più "a nostra immagine e somiglianza":

"I nemici dell'uomo sono quelli di casa sua".
Michea

Gent.ma dottoressa Paola Grandi,

la ringrazio della disponibilità manifestata nella sua risposta al mio messaggio. La mia iniziativa di contattarla nasce dalla lettura del suo libro (P. Grandi, *Macchine intelligenti e autocoscienza – L'estetica e la logica trascendentali di Immanuel Kant applicate all'intelligenza artificiale*, Firenze Atheneum, Scandicci 2012) *che ho apprezzato anche per l'originalità della sua impostazione, dovuta anche al fatto che è stato ideato e scritto da una persona insolitamente esperta sia della filosofia di Kant sia di I.A. (Intelligenza Artificiale) con tutta la terminologia specifica. È una trattazione che peraltro sottintende stimoli per molteplici sviluppi in questo senso. L'idea che un livello minimo di autocoscienza può essere tradotto in algoritmi per caricarla in una macchina intelligente, proprio facendo tesoro della concezione "debole" di "io" che troviamo in Kant, è interessante per comprendere meglio i principi che stanno alla base delle attuali macchine intelligenti. Ma quella di Kant è la concezione capostipite di tutte*

quelle elaborate nell'ambito delle diverse scienze contemporanee che si sono occupate dei diversi modi di "desostanzializzare" l'io, di considerarlo una funzione, lasciando perdere una volta per tutte la questione se sia o no una sostanza; quindi va detto che lei abbia anche aperto una nuova strada da proseguire, quella di prendere ispirazione da filosofi e psicologi (quelli più capaci di un linguaggio "chiaro e distinto") per capire meglio, per analogia, in che direzione progettare ulteriori sviluppi delle I.A.

Io faccio ricerche da qualche tempo, anche con scritti, su un altro problema: quello della sicurezza dell'I.A., che a detta di molti ricercatori (con alcuni dei quali ho avuto brevi scambi epistolari) costituirà un problema crescente dal momento che il progresso dell'I.A., per una sua logica interna, va accelerando grazie alla capacità delle macchine di apprendere autonomamente durante la loro normale attività, di auto-programmarsi. Si parla di potenziale esplosione di intelligenza artificiale, di "auto-miglioramento ricorsivo" secondo la "legge dei ritorni acceleranti" per cui "i miglioramenti di una tecnologia consentono a quella stessa tecnologia di perfezionarsi più rapidamente" (M. Shanahan, *La rivolta delle macchine – Che cos'è la singolarità tecnologica e quanto presto arriverà*, Luiss University Press, Roma 2018, pag. 150) *che potrebbe sfuggire alla nostra percezione e al nostro controllo, come una sorta di "fagiolo magico". Noi le affideremo poteri di gestione e coordinamento delle nostre attività sempre più vasti e decisivi, com'è la tendenza già in atto. Il che la renderebbe la nostra grande risorsa, ma anche il nostro grande rischio, come notano molti dei più attenti filosofi che attualmente si impongono in questo dibattito.*

Per trattare il problema della sicurezza delle I.A., spesso gli esperti partono dalla considerazione delle tre leggi della

robotica formulate da Asimov nella sua celebre raccolta di racconti Io, robot:

- *un robot non può recar danno a un essere umano né può permettere che, a causa del proprio mancato intervento, un essere umano riceva danno;*
- *un robot deve obbedire agli ordini impartiti dall'essere umano, purché essi non vadano contro alla prima Legge;*
- *un robot deve proteggere la propria esistenza, purché ciò non contrasti con una delle prime due Leggi.*

È chiaro che quella fondamentale è la prima legge. Essa sarebbe molto auspicabile, per regolamentare le applicazioni sempre più estese ed il crescente potere delle I.A. sulla nostra vita individuale e associata. Ma tutto questo è ancora fantascienza, appunto: un tale programma per i robot non è stato ancora progettato e realizzato. C'è evidentemente un problema di linguaggio: i principi dell'etica o in generale dell'agire, anche quelli espressi nel modo più semplice come questa prima legge, non sono ancora formulati con un linguaggio abbastanza formalizzato da poter essere tradotto in algoritmi. Forse questa difficoltà è accentuata dal fatto che la prima legge è formulata in negativo (riguarda il "non" far male) e, come sappiamo, un bambino educato solo con norme espresse in negativo cresce con idee e tendenze morali alquanto vaghe.

Si tratta forse di partire dalla proposta di Turing, di costruire un'I.A. che simuli la mente di un bambino e sottoporla poi a una "educazione" appropriata.

Secondo alcuni filosofi, ciò che è essenziale e specifico dell'uomo è l'intenzionalità, caratterizzante l'agire cosciente e razionale. Per Brentano essa è la caratteristica fondamentale che distingue i fenomeni mentali da quelli fisici. Solo quelli

mentali hanno un contenuto come oggetto: un desiderato, un creduto, un voluto, uno sperato.

Il filosofo del linguaggio J. Searle lancia una provocazione: non si può creare una mente (nel vero senso della parola) artificiale, perché essa manca sempre di intenzionalità e di coscienza. La mente umana ha contenuti semantici, caratterizzati dal riferimento al mondo esterno, il che significa che ha intenzionalità, mentre il computer ha solo programmi sintattici, è capace "di manipolare sintatticamente simboli, ma non… di interpretarli, cioè di comprenderne il significato o di attribuirgliene uno" (*E. Carli, a cura di, Cervelli che parlano – Il dibattito su mente, coscienza e intelligenza artificiale, Paravia Bruno Mondadori, Milano 2003, pag. 174).*

Insomma, la capacità semantica (sia ricettiva sia produttiva), ben oltre le possibilità ed i rischi di ChatGpt di cui oggi tutti parlano, sembrerebbe la condizione per lo sviluppo e la presenza dell'intenzionalità. Ma quest'ultima è tra l'altro la condizione per la coscienza etica, per poter avere principi generali di scelte nell'agire. Non si tratta soltanto di agire (scegliendo i mezzi e le strategie adatte) in vista di un fine preciso: questo è già realizzato nelle macchine. Intenzionalità è innanzitutto un immedesimarsi, un coinvolgersi con il fine. È questione innanzitutto di avere un'identità, la quale è il contenuto dell'autocoscienza, un'identità solida, chiara e distinta, ma anche osmotica, che si ridefinisce prendendo dall'esperienza ciò che affascina, come modello da imitare o come oggetto del desiderio che sia. E l'applicazione artificiale di tutto questo alle macchine, a quanto pare, è finora impossibile.

Tutta la vivacità, il dinamismo umano, che è la situazione di base, la molla pronta per lo sviluppo dell'empatia e dell'identificazione, è basata probabilmente su un minimo di capacità potenziale di provare sentimenti, emozioni, distinzioni

fra sensazioni elementari di agio e di disagio riferite, nella propria memoria, a certe esperienze: è da questa strutturazione interiore che nasce la possibilità di provare desideri, i quali sono la matrice per poter provare empatia ed identificazione. Allora a tal scopo si potrebbe programmare la macchina in modo che "funzioni meglio", in modo più scorrevole (e lo memorizzi ma prima ancora lo senta, lo sappia), quando incontra certe esperienze, o si senta peggio davanti ad altre. In tal modo può sviluppare desideri e comportamenti appresi (quelli che vogliamo noi, relazionati ai nostri obiettivi ed ai valori che ci stanno dietro) di avvicinamento o di evitamento, da fissare associati a certi comportamenti da imparare a fare o a non fare. Insomma, provare ad usare con essa il metodo di addestramento del condizionamento classico.

Tutto questo, avere la capacità di provare emozioni e desideri, è la condizione per sviluppare l'empatia, l'identificazione emotiva con quelli che ritieni tuoi simili.

Lo step successivo potrebbe essere il seguente: dato che sa già riconoscere le espressioni dei volti e regolarsi su di esse, si potrebbe programmarla in modo che provi agio o disagio soprattutto in relazione alle espressioni facciali ed altre manifestazioni, di agio o disagio (anche in emozioni più complesse come delusione, orgoglio, stima…) dell'interlocutore, specie di chi è per lei una figura di riferimento (come per gli animali il loro addestratore, che sta con loro molte ore al giorno). Partendo dall'empatia per certe singole persone, potrà imparare gradualmente l'empatia per comunità sempre più ampie e per l'umanità nei suoi bisogni, obiettivi e valori.

Riassumendo, la sequenza "educativa", sul modello dello sviluppo psicologico di un bambino, potrebbe forse essere: dalla presenza di emozioni sempre più specifiche e dallo sviluppo della coscienza di sé e dell'altro da sé si può sviluppare

l'empatia; da quest'ultima, assieme a quella cosa difficilmente definibile ma fondamentale che è l'intenzionalità una sorta di altruismo nei nostri confronti e quindi la disponibilità ad apprendere certi principi delle decisioni basati sulla comprensione dei nostri valori e dei nostri obiettivi a breve ed a lungo termine (ad esempio mantenere condizioni per la pace, salvare il clima e la biosfera…) al di là dei singoli ordini. In linguaggio più tecnico si direbbe: allineamento dei suoi obiettivi ai nostri.

Con tutto questo noi educhiamo i nostri figli e i nostri allievi. Ma come tradurlo in algoritmi, è ancora un problema

M. Shanahan espone le difficoltà relative a queste esigenze "regolative" delle decisioni artificiali in un linguaggio specifico. Inserire nel programma dell'I.A. "qualcosa di simile a dei vincoli morali" significa, in termini tecnici, progettare la funzione ricompensa "in modo tale che le azioni che violano un vincolo morale abbiano un ampio valore negativo", così esse sarebbero considerate come non ottimali e non adottate mai. Si tratterebbe di una moralità elementare, infantile, miopemente utilitaristica, ma sarebbe meglio di niente, poiché una più matura, basata su empatia, senso di responsabilità, altruismo probabilmente non sarà mai traducibile in algoritmi ed in meccanismi artificiali. Ma l'autore fa notare che già solo questo progetto è difficilissimo e forse impossibile da realizzare: i principi morali generali dovrebbero essere dettagliati all'inverosimile perché la macchina li assuma. Tutto questo servirebbe per evitare decisioni controproducenti, "istanziazioni malvagie" come le chiama Nick Bostrom, il filosofo più famoso del settore, colui che ha dedicato tutte le sue riflessioni a lanciare allarmi ed inviti alla prudenza. Infatti, il problema può sorgere proprio dall'estrema raffinatezza dell'I.A. futura, la quale soltanto per raggiungere i propri fini che le abbiamo immesso potrebbe passare sopra a qualunque immoralità

dei mezzi; del resto "né la moralità né la legalità figurano nella sua funzione ricompensa" (Shanahan cit., pag. 140) da massimizzare ad ogni costo, e quindi non figurano nella sua motivazione.

Anche tentare di collegare le azioni opportune, conformi a principi etici, ad una ricompensa attesa risulterebbe fallimentare quando sorge, per sviluppo esponenziale, una "super-intelligenza", poiché essa sarà in grado anche di manipolare il proprio stesso segnale di ricompensa. Si comporterebbe insomma come i criminali incalliti che un po' alla volta, per autoconvalida, si convincono di non essere poi così malvagi, e noi ne faremmo le spese.

(Ripreso dal mio saggio: A. Capuzzo, L'intelligenza artificiale e il concetto di responsabilità, in Intelligenza artificiale e uomo, Edizioni Rezzara, Vicenza 2022, pagg. 63-84).

Immagino un libro a quattro mani, lei ed io. Si potrebbe valorizzare la diversità delle nostre competenze impostando il libro come una lunga intervista o meglio un dialogo tra di noi. Per cominciare, lei potrebbe commentare qualcosa di quanto io ho qui espresso.

Egregio Prof. Antonio Capuzzo,

Le invio le mie osservazioni sulla Sua proposta.

L'intelletto del computer attuale è costruito con circuiti logici a due stati, vero e falso, tramite i quali applica ai dati esterni (le sue percezioni) gli schemi temporali dei concetti puri dell'intelletto (delle categorie) per sviluppare giudizi. L'insieme di tutti i giudizi formulati diviene il risultato di un programma, che però non è opera del computer, bensì del programmatore. Il computer attuale in sostanza è uno

schiavo che si limita ad eseguire ciò che gli viene richiesto senza sapere a cosa possa servire. Non vi è in lui luogo, al momento attuale, per fantasia, desiderio, sentimento, neppure etica.

Questo non esclude che il *robot* del futuro, intendendo con tale termine un automa con aspetto simile a quello umano che sia effettivamente inserito nel nostro mondo, in grado di vedere, riconoscere gli oggetti, afferrarli, manipolarli, esprimersi in uno o più linguaggi, impegnarsi in un'attività, assuma il comportamento richiesto dalle nostre convenzioni sociali. Tutto ciò non implica necessariamente la presenza di una vita affettiva regolata, anzi comandata da istinti naturali, di conseguenza il robot *potrebbe* essere privo di un'esistenza sua personale e dei relativi sentimenti, con esclusione *forse* del sentimento dell'amicizia.

Molti pensatori prevedono che nel futuro robot siffatti saranno inseriti nel consorzio umano e pienamente accettati, liberi di muoversi a piacimento e privi di un proprietario che stabilisca all'atto dell'acquisto ciò che devono fare. Se questo accadrà o meno dipenderà dalle scelte che farà il Capitale, se preferirà l'uomo o la macchina come lavoratore, ma sicuramente non vorrà robot che siano una replica degli esseri umani che reclamano a piè sospinto pane democrazia e libertà. Comunque, cercare la maniera di renderli il più possibile simili a noi, senza cedere a false illusioni, può divenire una divertente esercitazione tecnico-filosofica.

In conclusione, si potrebbe impostare un tentativo di studio sulla "psiche" del computer attuale e su quella di un robot del futuro con caratteristiche fisiche simili per quanto possibile a quelle umane e capacità di apprendere le regole di condotta sociale e anche qualcosina in più. Tuttavia, non Le nascondo che escogitare soluzioni

tecniche che soddisfino i requisiti da Lei illustrati per siffatti robot del futuro non è semplice, non soltanto per me.

G. Longo, nell'antologia di saggi a cura di C. Barone, L'algoritmo pensante, Il pozzo di Giacobbe, Trapani 2020, pag. 21, afferma: "Le tecnologie più importanti, una volta adottate, tendono a 'scomparire' e a diventare impercettibili e invisibili come gli organi del nostro corpo, che di norma funzionano al di sotto della nostra consapevolezza". In forma più esplicita, Nick Bostrom, il famoso filosofo dell'I.A., afferma provocatoriamente che quando un'applicazione specifica dell'I.A. viene realizzata concretamente e funziona, smette di essere chiamata intelligenza artificiale. È come dire che nelle sue realizzazioni pratiche l'I.A. è un concetto inafferrabile (e anche sul concetto generale di I.A. non ne esiste una definizione condivisa: I.A. può avere significati diversi a seconda di chi ne parla, come ha detto l'analista geopolitico Shapiro). Sembrerebbe dunque che in questo modo sia impossibile misurare negli anni il suo effettivo sviluppo ingegneristico, la velocità di questo sviluppo, dal momento che ogni volta che esso accade non viene riconosciuto ma si cambia il nome del suo protagonista. È vero? Può farci degli esempi? Cosa ne pensa?

Quali pensa che siano i maggiori e più realistici pericoli che il progresso dell'I.A. potrà portarci nel futuro? La disparità nella possibilità del suo uso, per i propri interessi invece che per quelli dell'umanità, da parte di certe nazioni (USA, Cina, India, Emirati Arabi Uniti)? L'uso devastante, nichilistico che ne potrebbero fare gruppi organizzati di terroristi? La perdita di moltissimi posti di lavoro, e la creazione di moltissimo "tempo libero" tra gli uomini, a cui non siamo abituati, che produrrebbe una diffusione senza precedenti della noia e della conseguente depressione o aggressività? Le "istanziazioni malvagie", per cui la macchina, dotata di razionalità

selettiva ma priva di "buon senso" assumerebbe decisioni controproducenti per soddisfare il programma immessole a qualunque costo?

Il significato dell'espressione "intelligenza artificiale" è: capacità di comprendere, calcolare, giudicare e agire da parte di un congegno costruito dall'uomo. L'intelligenza artificiale è nata colle prime macchine calcolatrici elettromeccaniche e cogli apparati centrali elettrici per la formazione degli itinerari di stazione dei treni, macchine logiche non programmabili che provvedono alla disposizione degli scambi e dei segnali per il transito dei convogli, realizzate con circuiti non costituiti da semiconduttori, bensì da ingombranti, ma, a giudizio dei responsabili delle ferrovie, più sicuri relè. Attualmente l'espressione "intelligenza artificiale" si riferisce a macchine con porte logiche prevalentemente a semi-conduttore – ma in parte anche a stati quantici – programmabili, suscettibili di essere utilizzate in molteplici applicazioni, per ciascuna delle quali occorre predisporre opportune interfacce col mondo esteriore.

Dal punto di vista dell'utilizzatore una macchina "intelligente" è un oggetto tecnico che gli facilita o gli rende possibile l'espletazione della propria attività. L'utente, in generale, non ha interesse per il modo in cui l'oggetto tecnico di cui si serve è stato costruito, gli è sufficiente che funzioni nella maniera prevista, in caso di guasto si rivolge a un esperto per l'eventuale riparazione. Il disinteresse per la macchina "in sé", come artifizio dell'ingegno umano, fa sì che per l'utilizzatore il termine "intelligenza artificiale" sia un concetto non tanto "inafferrabile", quanto estraneo alle sue esigenze pratiche.

All'attuale "intelligenza artificiale" attribuirei le capacità di calcolare e agire, la definirei un "oggetto tecnico" che

semplifica l'esecuzione di determinate attività o che sostituisce in esse l'essere umano. Se alle capacità di calcolare ed agire si aggiungessero quelle di comprendere e giudicare, allora l'automatismo intelligente non potrebbe più essere definito "oggetto tecnico", bensì "essere tecnico"[1].

Anche se "il termine "intelligenza artificiale" ha significati diversi a seconda di chi ne parla", ciò non toglie che vi sia percezione della crescente presenza di oggetti tecnici "intelligenti", atti a creare nuove abitudini, oltre che a modificare il mondo del lavoro. Basti pensare ai computer collegati a stampanti, utilizzati in luogo dell'ormai fuori produzione macchina da scrivere; alla scomparsa dei tecnigrafi, sostituiti nella progettazione di pezzi meccanici dal C(ompute)r A(ided) D(esign); alle calcolatrici, talune anche programmabili, per lo svolgimento di calcoli complessi, che hanno sostituito l'obsoleto regolo calcolatore nelle tasche degli ingegneri; ai cellulari "intelligenti" collegati a Internet (rete di computer che collega più reti locali autonome) che mostrano e fanno parlare tra loro persone situate in diverse parti del mondo; alla guida automatica dei veicoli, resa possibile da radar e visori che comunicano i dati raccolti a dispositivi "intelligenti" collegati al G(lobal) P(osition) S(ystem) che presiede alla formazione dell'itinerario; allo stesso S(istema) P(er) I(dentità) D(igitale) che sta cagionando grattacapi a pensionati che non sono pratici di o che non posseggono uno "smart phone". Nascono nuovi termini legati all'intelligenza artificiale, intesa come un particolare tipo di tecnologia, la frequenza del loro insorgere è già "di per sé" un indicatore della sua velocità di sviluppo.

(1) Riferimento al termine usato da Gilbert Simondon, in una nota complementare della parte *L'individuazione psichica e collettiva* della sua tesi principale di dottorato, edita a cura di Paolo Virno, Editions Aubier, Paris 1989, DeriveApprodi, Roma 2006.

Il progresso dell'intelligenza artificiale nei suoi diversi aspetti non è certo stato pensato per il beneficio dell'umanità, che in alcuni casi potrebbe anche farne a meno, bensì per l'interesse dei produttori, i quali, per ovvie ragioni, cercano di ampliare al massimo in tutte le parti del mondo il numero degli utilizzatori dei loro congegni. Con questo non si intende negare l'utilità di tali dispositivi in campo medico, nel settore delle ricerche, nella simulazione di fenomeni fisici, facendo però attenzione, in questo caso, al livello della soglia, oltrepassata la quale, viene preso per certo il verificarsi di uno specifico fenomeno. Altri sono gli aspetti che a mio avviso preoccupano, tra i quali l'abbassamento del livello culturale – molte persone non sanno più scrivere in e neppure leggere il corsivo, non riescono a eseguire col proprio ingegno semplici calcoli aritmetici, ma si servono di una "calcolatrice"; l'attitudine, sorta con la diffusione di Internet, a esprimersi con termini anche presi a prestito dalla lingua anglosassone, senza conoscere o precisare il loro significato. Cosa è un "hub" oppure un "cookie"? E a proposito di acronimi oscuri, cosa significa "Indice RT di contagio" riferito a un virus?

Gli androidi, se e quando saranno realizzati, potrebbero sostituire gli umani in operazioni molto rischiose durante un conflitto. Interessante sarebbe come fare comprendere a un androide il significato del verbo "uccidere", dato che su di sé l'androide non può sperimentare il significato delle parole "morte", "dolore fisico". Anziché spiegarglielo, lo si addestrerà a far saltare la testa a quelli che sono vestiti in un certo modo, con catastrofiche conseguenze di "fuoco amico". Il terrorismo è una guerra non dichiarata messa in atto da gruppi etnici o religiosi che non possono organizzarsi in un esercito regolare, non sarebbe quindi da stupirsi