Alessio Montagner

PERSONA ED INTELLIGENZA ARTIFICIALE: LE MACCHINE HANNO UN'ANIMA?

Abstract. Articolo divulgativo pubblicato su Club Theologicum il 12 dicembre 2022. La prima sezione collega il problema dell'identità personale al paradosso delle sorite. La seconda approfondisce il problema, rifacendosi agli esperimenti mentali di Parfit in *Ragioni e Persone*. Si conclude sposando un ecceitismo coerente con la teologia cristiana. La terza parte espone alcune argomentazioni tipiche di filosofia della mente per analizzare la possibilità di assegnare una soggettività a una macchina. Nella quarta parte, concludo sottolineando i più comuni dilemmi etici legati all'intelligenza artificiale, come il facilitare il ricorso all'opzione militare nella risoluzione delle controversie internazionali.

Abstract. Popular article published in Club Theologicum on 12 December 2022. The first section connects the problem of personal identity to the sorites paradox. The second delves into the problem in greater depth, referring to Parfit's thought experiments in Reasons and Persons. It concludes by espousing a Haeceitism consistent with Christian theology. The third part presents some typical arguments of philosophy of mind to analyze the possibility of assigning subjectivity to a machine. In the fourth part, I conclude by underlining the most common ethical dilemmas related to artificial intelligence, such as facilitating the use of the military option in the resolution of international disputes.

Url.
https://clubtheologicum.com/2022/12/12/persona-ed-intelligenza-artificiale-le-macchine-hanno-unanima-teologiadigitale-cyborg-lanternadelcercatore/

1. Esponiamo il problema con degli esempi…

Io sono una persona, e ho un certo corpo. Il mio corpo può cambiare. Potrei perdere un braccio e sostituirlo con una protesi. Sarei però sempre io.

Potrei aver bisogno di un trapianto di polmoni, di cuore, di reni, di retine, e poi di una gastrectomia, addirittura di un'emisferectomia (rimozione di un emisfero cerebrale: incredibilmente, si sopravvive e si recuperano molte funzionalità). Anche il mio cervello ne uscirebbe trasformato, ma rimarrei sempre io.

Potrei modificare il mio corpo con impianti tecnologici. Un gruppo di ricercatori dell'università di Stanford nel 2016 ha modificato geneticamente un topo in modo che sviluppasse dei neuroni sensibili agli impulsi luminosi. Quindi gli hanno impiantato nel cervello un led wireless, posizionato in modo tale da stimolare certi neuroni e un nervo della gamba. I ricercatori, controllando l'attivazione di questo led, sono riusciti a far girare il topo nella gabbietta come volevano loro. Una tecnologia molto utile: in caso di terremoto si può mandare il topo con una telecamera tra le macerie e stimolarlo a girare nel modo voluto in cerca di superstiti. Ecco, immaginiamo che riceva anche io un impianto del genere nel cervello. E potrei avere anche organi artificiali, protesi robotiche al posto degli arti e quant'altro. Eppure, sarei sempre io.

Vi è forse una parte del mio corpo che mi è essenziale? Pare di no. Potrei sostituire un qualsiasi mio atomo con un altro simile: nessun atomo mi renderebbe un'altra persona. E anche se ne sostituisco due rimangono sempre io. Anche tre. E via così all'infinito. È il classico problema delle sorite, o della nave di Teseo. Tutte le varie soluzioni costruttiviste, epistemiciste, fuzzy, a soglia mobile, si sono rivelate insoddisfacenti e ad hoc. Per me, la proposta migliore rimane il *relativismo* di Geach: se cambio tutte le mie cellule, sono la stessa

persona ma un diverso insieme di cellule (identico relativamente alla personalità, non relativamente alla "cellulosità"). È logico che una persona possa subire serie di cambiamenti sequenziali nel corpo rimanendo la stessa persona: anche se ogni sua cellula viene sostituita con un'altra, in virtù di questa Storia che unisce i vari stati del corpo quella persona rimane la stessa, anche qualora, ad un certo punto, dovesse diventare un cyborg composto da parti artificiali.

Immaginiamo ora che uno scienziato crei un androide. Io posso essere un cyborg nel senso che *ero* composto da parti naturali, le quali sono poi state via via sostituite con parti artificiali. Questo androide invece è creato subito con parti artificiali. L'io-cyborg e l'androide, quindi, hanno Storie diverse. Epperò fisicamente l'io-cyborg e l'androide possono essere del tutto identici, possono essere composti da uguali parti. Ora, io, anche se divento un cyborg, rimango una persona. Allora lo è anche l'androide a me uguale?

Immaginiamo che lo scienziato prenda un utero artificiale, prenda i gameti maschili e femminili, li unisca e li faccia sviluppare nell'utero artificiale. L'individuo prodotto da questo processo rimane un uomo, una persona: gli uomini *normalmente* non nascono così, ma, anche se si tratta di un processo non-etico, *possono* nascere anche così.

Immaginiamo ora che lo scienziato pazzo, invece di prendere gameti naturali, ne produca lui artificialmente assemblando le molecole una a una esattamente così come sono nei gameti naturali. L'operazione è sempre più corrotta moralmente, ma questi gameti, seppur prodotti da strumenti anziché da un organo, continuano a produrre un uomo con i nostri sentimenti, la nostra spiritualità, insomma una persona. È uno dei modi in cui una persona può venire al mondo.

Al che, se una persona può nascere anche tramite lo sviluppo di gameti artificiali, è possibile che possa nascere anche semplicemente tramite l'assemblaggio di parti artificiali? In altre parole: se io produco un computer intelligentissimo, capace di un comportamento che appare in tutto razionale, emotivo e spirituale, posso definirlo una persona?

2. Che cos'è una persona

Presumibilmente, *l'identità personale transtemporale*, cioè il mio essere sempre la stessa persona nei vari istanti della mia vita, superviene sulla continuità del fatto che mi rende persona. Posso quindi capire cos'è che mi rende persona capendo cos'è che mi rende *la stessa* persona.

I filosofi hanno sviluppato due teorie principali in merito al fatto che rende vera l'identità personale: la *continuità corporea* e la *continuità psicologica.* Per la prima teoria, ciò che costituisce l'identità tra l'io-presente e l'io-passato è il fatto che condividiamo lo stesso corpo, l'identità personale superviene sulla continuità del corpo. Per la seconda teoria, invece, ciò che costituisce l'identità personale è il fatto che l'io-presente abbia accesso ai ricordi dell'io-passato, cioè l'esistenza di caratteristiche psichiche che da uno rimandano all'altro.

Ciò che abbiamo detto sopra rappresenta già un'obiezione convincente contro la continuità corporale. Si tratta, infatti, di una confusione tra *identità* e *composizione*. Il David di Michelangelo è identico al suo marmo? Immaginiamo di ridurre in polvere il David e di usare quella polvere per costruire un gabinetto: quel gabinetto sarebbe il David? No, è al massimo *il marmo del David*. Se invece, di restauro in restauro, sostituissi ogni pezzetto di marmo del David con altro marmo, quello non sarebbe più il David? Certo che lo sarebbe, anche se è cambiata la sua composizione materiale. Cioè: prima le espressioni "David" e "marmo del David" si riferivano *contingentemente* allo stesso ente, dopo queste trasformazioni si riferiscono a enti distinti. È chiaro insomma che non si può ridurre l'identità nel tempo di un ente alla continuità di ciò che lo compone, e quindi non si capisce neanche che fatto corporale usare.

Anche contro la continuità psicologica i filosofi hanno sviluppato varie argomentazioni. L'esempio più famoso si trova in Derek Parfit, in *Ragioni e Persone*. Immaginiamo un teletrasportatore così funzionante:

entro nella cabina di partenza, essa analizza la posizione di ogni mia particella, il mio corpo viene quindi disintegrato mentre la cabina di destinazione assembla le particelle secondo l'analisi fatta. Immaginiamo però che il teletrasportatore abbia un malfunzionamento: nella cabina di destinazione viene creato un corpo, solo che la cabina di partenza non mi ha disintegrato. Uno scienziato viene a rassicurarmi: il corpo nella cabina di partenza si autodistruggerà nel giro di qualche ora. Domanda: dovrei essere preoccupato? Il corpo nella cabina d'arrivo ha la mia stessa esatta struttura e condivide tutte le mie caratteristiche psichiche, tutti miei ricordi. Vi è quindi piena continuità sia corporale che psicologica tra me e l'essere nella cabina di destinazione. Continuerò quindi ad esistere lì? Diremmo proprio di no: io in realtà sono rimasto nella cabina di partenza, e quello dall'altra parte, anche se ha con me continuità psicologica, non sono io. Quindi neanche la continuità psicologica è un criterio accettabile.

Lasciando da parte altre teorie minoritarie ugualmente problematiche (rapporto causale, essenzialità dell'origine, ecc), tutto pare dare ragione a Hume, che negava direttamente l'esistenza di un fatto costituente l'identità. Dato che posso immaginare dei cambiamenti in tutte le mie qualità senza alterare l'identità devo concludere che nessuna qualità mi è essenziale. Ciò ci costringe, se non si vuole affermare l'inesistenza dell'identità, a immaginare una *ecceità*, cioè, nel senso contemporaneo, una *proprietà non-qualitativa impura* (i.e. esprimibile solo riferendosi ad un altro ente) che costituisce l'identità.

Parlando tra credenti, per noi ciò che svolge tale ruolo è l'anima incarnata. Parfit, sostenendo la posizione di Hume, paragona una persona a un Paese: un Paese non è un territorio, non è una popolazione, eppure è null'altro che un territorio e una popolazione. Noi possiamo dire la stessa cosa: una persona non è un corpo/mente, non è un'anima, eppure è null'altro che un corpo/mente e un'anima.

3. Si può sapere se una macchina ha un'anima?

Quindi è una persona ciò che ha un'anima. Ovviamente non esiste un modo scientifico, empirico, per capire se qualcosa ha o non ha un'anima razionale, immortale. Ciò nonostante, possiamo cercare di capire se sia così sulla base di altri fatti derivabili dall'anima razionale, come certe caratteristiche mentali.

Attenzione: l'anima non è la mente. Il nostro concetto di mente come *res cogitans* (e quindi anche i concetti di coscienza, esperienza soggettiva, qualia…) nasce con la filosofia moderna, e tracciare paralleli con la filosofia antica o medievale è sviante. Ciò nonostante, esiste un collegamento tra i due campi. L'anima è la forma del corpo. La forma è ciò che "etichetta" gli enti: la *materia prima* non è un *qualcosa* ma è un indefinito, è *ciò da cui vengono tutti gli enti*, e qualsiasi ente possa essere identificato (anche i quark) è già composto da forma e materia. L'anima è ciò che costituisce l'identità perché è ciò che "etichetta" il corpo rendendolo me, è ciò che individua me in quello che altrimenti sarebbe solo un ammasso indefinito. Ora, il corpo ha l'*intelletto passivo*, che di fatto è la vera *res cogitans* e che alla luce delle neuroscienze moderne può essere ridotto ad una serie di eventi neurali (potenzialmente sufficienti all'essere *senziente*). L'anima, però, fa da *intelletto attivo*, ed è questo che "etichetta" nel passivo i pensieri, è quest'anima che prende quelli che sarebbero solo movimenti nella materia e li rende i miei ragionamenti e le mie esperienze soggettive (più che senziente, *cosciente*). (Aristotele non condividerebbe questo uso dei suoi termini? Colpa sua!)

Ora, di sicuro un computer ha una forma nel senso che è un oggetto definito. Facendo un paragone con il cervello, ha anche un intelletto passivo nel senso detto sopra: i suoi circuiti possono essere simili ai miei neuroni e può fare cose simili a quelle che faccio io. Ma ha una coscienza e una razionalità, e quindi un intelletto attivo, e quindi un'anima, o è piuttosto uno *zombie*, un corpo che simula un comportamento?

Oggi la posizione più comune in filosofia della mente è quella del f*unzionalismo*. Cos'è un cuore? È ciò che svolge la funzione di pompare sangue, non importa come sia fatto. Cos'è il dolore? È ciò che svolge una certa funzione, ciò che porta ad un certo comportamento, non importa come. Tale posizione vuole rendere merito di come sia possibile che organi diversi possano portare a uguali esperienze coscienti: tutti gli uomini

hanno cervelli diversi, ma è intuitivo che abbiano esperienze simili. Ora, in linea di principio, non c'è limite a quale possa essere il supporto in cui si rappresenta una certa funzione. Certo il mio sistema nervoso può avere una serie di eventi, nello specifico l'attivazione delle fibre-C, che costituiscono il mio dolore. Ma anche nei chip di un computer si può simulare lo stesso evento. Dovrei dire, in quel caso, che il mio computer sta provando dolore?

La risposta intuitiva è no, un computer non è neppure senziente. Il filosofo Ned Block fa l'esempio del *China Brain*. Prendiamo tutta la popolazione cinese, facciamo svolgere ad ogni uomo il ruolo di un neurone, quindi li facciamo comunicare tra loro in modo tale da mimare gli eventi neurali corrispondenti al dolore. Quando ciò avviene vi è forse una super-mente costituita dall'attività delle persone? Pare un'assurdità. Il funzionalismo rende merito della mente intesa, appunto, come intelletto passivo, solo come serie di movimenti di un certo tipo. Non parla, invece, della mente intesa come intelletto attivo, e quindi come esperienza soggettiva.

Alla fine, la coscienza risulta tanto fuggente quanto l'anima stessa: la sua natura soggettiva rende impossibile l'esistenza di un fatto empirico in grado di dirci se un certo evento corrisponda ad un'esperienza. I sassi potrebbero essere senzienti a mia insaputa, oppure io potrei essere l'unico uomo senziente sul pianeta Terra. Anche qualora un robot si comportasse *esattamente* come me, e addirittura pregasse, sarebbe per me legittimo trattarlo come uno zombie. Ma sarebbe ugualmente legittimo, allora, trattare un sasso come una persona. Cercare una soluzione è importante. Se infatti ritenessimo non coscienti delle macchine che invece lo sono rischieremmo di essere involontariamente crudeli con loro. Di contro, se ritenessimo coscienti delle cose che non lo sono rischiamo di sminuire la posizione dell'uomo.

I dati del sondaggio di Philpapers (la più grande comunità di filosofi professionisti), condotto da Chalmers (il più influente filosofo della mente vivente) conferma i dubbi. I filosofi sono piuttosto convinti della senzienza di umani adulti (95% sì, 0,2% no), gatti (89% sì, 4% no), neonati (84% sì, 5% no), hanno dubbi su pesci (65% sì, 15% no) e mosche (35% sì, 38% no), negano la senzienza di vermi (24% sì, 47% no), piante (7% sì, 80% no) e particelle (2% sì, 89% no). E le intelligenze artificiali? Se le risposte sono negative per i sistemi *correnti* (3% sì, 82% no), i filosofi sono ottimisti sulle IA future: il 39% crede che saranno coscienti, il 27% no.

Già Turing, in *Computating Machinery and Intelligence*, dice che un credente non dovrebbe porre limiti a Dio: è in linea di principio possibile che anche una macchina riceva un'anima. A tale idea Giovanni Amendola, teologo e matematico specializzato in intelligenza artificiale, obietta che "l'uomo è corpo animato o anima incarnata sin dal concepimento", e senza anima non esiste neppure il corpo in senso proprio, mentre una macchina dovrebbe, per Turing, ricevere l'anima con l'apprendimento, quindi la macchina non è e*ssenzialmente* razionale e quindi neanche animata. Non pare però un'obiezione sufficiente. In primo luogo perché l'animazione immediata, seppur diventata negli ultimi anni una posizione comune, rimane problematica e tutt'altro che scontata. E in secondo luogo perché nulla vieta di immaginare un'animazione immediata anche per una macchina: anche se una macchina, a differenza di uno zigote, non è autonomamente diretta verso la formazione di un corpo capace di razionalità, comunque è possibile assegnare un'anima sin dal primo istante in cui possiamo dire esistente un corpo dotato essenzialmente di certe potenzialità o di una certa teleologia. Il raggiungimento di un certo punto nell'assemblaggio e nella programmazione della macchina diventerebbe equivalente all'unione dei gameti nella fecondazione in vitro, cioè uno dei modi nei quali il corpo di una persona può venire al mondo.

Resta però che nessun fatto oggettivamente individuabile possa dirci se una macchina sia una persona oppure no. E anche la rivelazione pare del tutto neutrale. Credo che solo aspettando l'avvento di queste macchine intelligenti e vivendo insieme a loro si potrà capire, anche grazie alle nostre facoltà intuitive ma sempre evitando di cadere in facili bias d'attribuzione, se siamo davanti a persone oppure no.

4. Pericoli e opportunità: il rapporto tra IA e religione

Ciò che distingue un'IA da qualsiasi algoritmo è la sua capacità di svolgere lavori senza aver ricevuto istruzioni specifiche su come farli.

Pensiamo ai computer scacchistici. Perché Deep Blue, il motore che sconfisse il campione del mondo Kasparov, era un così abile giocatore? Perché programmatori e Grandi Maestri hanno collaborato per creare degli algoritmi che incarnassero i principi della teoria scacchistica. Deep Blue è più forte di un giocatore umano perché ha molta potenza di calcolo in più. Per il resto, gioca come gli uomini gli hanno insegnato: fa le aperture "a memoria" (segue "libri" con mosse standard), cerca di controllare il centro, sviluppa i pezzi, evita i bordi, arrocca il prima possibile, eccetera.

Poi nel 2017 è uscito AlphaZero, di Google. Nessuno ha spiegato ad AlphaZero come si gioca a scacchi: ha iniziato facendo partite a caso contro sé stesso e cercando pattern. Il risultato è stupefacente: nel giro di poche ore di auto-apprendimento AlphaZero è diventato immensamente più potente di qualsiasi altro computer scacchistico. Ma soprattutto, è diventato un computer con personalità, un computer creativo: il suo stile di gioco è alieno eppure intelligibile, aggressivo come una bestia nel valutare più l'iniziativa che il materiale, ha buttato nella spazzatura secoli di teoria scacchistica eppure vince sempre.

Cartesio credeva che due caratteristiche avrebbero distinto sempre l'uomo dalle macchine: la capacità di parlare, e la capacità di svolgere un gran numero di task diverse. In effetti, i programmi tradizionali come Deep Blue possono svolgere solo ciò per cui sono programmati. Nel caso di IA avanzate come AlphaZero non è così: ha imparato a giocare anche a Shogi e a Go, raggiungendo sempre un livello sovraumano. Suo "figlio", MuZero, è ancora più abile: ha imparato a giocare a 57 videogiochi diversi senza conoscenze pregresse. E suo "nipote", Gato, può imparare contemporaneamente come svolgere anche 600 task diverse: sa rispondere a domande, comporre poesie, giocare a videogiochi, impilare blocchi, controllare macchinari…

Qual è il limite? Fin dove ci si può spingere? Ecco, un limite c'è…

Se c'è una cosa che mancherà a queste macchine, è l'etica. Questo perché l'etica non dipende da fatti naturali conoscibili, per un credente essa dipende semplicemente dalla volontà di Dio e dai suoi mitzvot. Un'IA può scoprire un nuovo teorema matematico o sviluppare una nuova teoria fisica, e saremmo disposti ad accettarli anche se non fossimo in grado di dimostrarli in prima persona; non può però scoprire che l'omicidio sia bene, questa affermazione non la accetteremo. È sempre l'uomo a dover dare un'etica all'IA, a svilupparla in modo tale che non faccia ciò che non vogliamo, addirittura a educarla con l'esempio. E questo è un primo incontro tra IA e religione.

Infatti lo sviluppo di IA estremamente avanzate pone nuovi problemi etici anche in merito al comportamento umano. Le IA possono essere usate dalle grandi industrie per manipolare, sfruttare bias, creare dipendenze affinché potenziali clienti passino sempre più tempo sui loro servizi e rivelino sempre più dati su di sé. Le IA possono essere usate per creare armi autonome, in grado di scendere sul campo di battaglia e prendere decisioni: questo toglie la responsabilità delle uccisioni dagli uomini, rendendo più probabile e semplice la scelta della guerra come modo di gestire le relazioni internazionali. Alcuni hanno ipotizzato lo sviluppo di IA capaci di simulare relazioni sessuali con degli umani, cosa che pone evidenti problematiche etiche. Le IA, secondo alcuni, potrebbero portare ad una vera e propria fine del lavoro, un mondo dove quasi tutte le mansioni saranno svolte da IA e gli unici posti di lavoro a sopravvivere saranno ultra-specializzati e accessibili solo a pochissime persone: come si dovrebbe gestire uno scenario del genere? E poi il problema del potenziamento umano: la IA possono aiutare a sviluppare tecnologie in grado di aumentare le capacità umane, come impianti cerebrali che donano una capacità di calcolo superiore a quella di qualsiasi uomo passato; ma sarebbe etico espandere a tal punto in modo artificiale le proprie capacità? C'è il rischio di creare una élite di individui "potenziati" mentre alla grande maggioranza della popolazione potrebbe essere escluso l'accesso a tale tecnologia. Anche in questi campi la religione deve offrire la sua prospettiva etica.

Ciò nonostante, credo che non si debba sovrastimare il rischio che può porre una IA malvagia o, questo sì più probabilmente, un uso malvagio delle IA da parte degli uomini. Possiamo sperare che le IA verranno sviluppate dall'uomo per il bene degli uomini, come un servizio, e non come un gioco per creare "persone artificiali". IA testuali come Open AI possono produrre testi su qualsiasi tema venga loro richiesto, anche su questioni teologiche, e così possono offrirci nuove idee e mostrarci nuovi modi di affrontare le questioni. IA text-to-image come Midjourney possono immaginare nuovi stili artistici e creare immagini potenti, ci aiutano già a creare una nuova arte cristiana, a trovare nuove ispirazioni, a visualizzare realtà spirituali che credevamo inimmaginabili.

Concludiamo con un aneddoto riportato da Linda Kinstler sul NYT. Rob Barrett era un ingegnere all'IBM negli anni '90. Stava lavorando sulla gestione della privacy per un browser. Il suo capo gli aveva detto solo "fa' la cosa giusta". Dopo un po' Barrett è tornato dicendogli "non conosco abbastanza teologia per essere un buon ingegnere!" Ha chiesto un periodo di vacanza per studiare l'Antico Testamento. Pare che non sia più tornato.