

IL PROBLEMA DELLA COSCIENZA NELLA PROSPETTIVA DELLA INTELLIGENZA ARTIFICIALE

Gaetano Licata

Abstract *The problem of consciousness from the perspective of Artificial Intelligence.* Both in philosophical thought and in common sense, consciousness has always been considered the identifying feature of human nature. The question of the mind-body relationship, which has engaged philosophical reflection from Descartes and Suárez onwards, today obtains a renewed centrality thanks to the debate on Artificial Intelligence, and in particular regarding the project of an artificial consciousness. I will then consider some strategies aimed at solving the problem of artificial consciousness. I will discuss the relevance of consciousness for behavior, the relationship between the unconscious and consciousness, and I will criticize the irrelevance of consciousness proposed by some neuroscientists within the framework of the eliminativist theories; I will discuss some emergentist theses. Finally, I will support a non-reductionist thesis on consciousness based on the works of Maturana and Varela, and a monist position on the mind-body problem.

Keywords: Consciousness, Artificial Intelligence, Engineering of consciousness, Functionalism, Emergentism, Dualism of matter and spirit.

Abstract Sia nel pensiero filosofico che nel senso comune, la coscienza è stata sempre considerata il tratto identificante della natura umana. La questione del rapporto mente-corpo, che da Cartesio e Suárez in poi ha impegnato la riflessione filosofica, ottiene oggi una rinnovata centralità grazie al dibattito sulla Intelligenza Artificiale, ed in particolare riguardo al progetto di una coscienza artificiale. Prenderò quindi in considerazione alcune strategie volte a risolvere il problema della coscienza artificiale. Tratterò della rilevanza della coscienza per il comportamento, del rapporto fra inconscio e coscienza, e criticherò la irrilevanza della coscienza proposta da alcuni neuroscienziati entro la cornice delle teorie eliminativiste; discuterò alcune tesi emergentiste. Sosterrò infine una tesi non riduzionista sulla coscienza basata sui lavori di Maturana e Varela, ed una posizione monista sul problema mente-corpo.

Parole chiave: Coscienza, Intelligenza artificiale, Ingegnerizzazione della coscienza, Funzionalismo, Emergentismo, Dualismo di materia e spirito.

1. Introduzione

Riusciremo mai a capire cosa è la coscienza? Le attuali teorie fisicaliste, emergentiste o eliminativiste, risponderanno mai a questo enigma? La domanda lascia implicita una singolare difficoltà. Solo una coscienza può fare questa domanda e la fa su sé stessa; sotto questo riguardo la risposta è quindi no: non riusciremo mai a comprendere cosa sia la coscienza nella sua interezza. La coscienza è sì un fenomeno naturale come gli altri ma, dato che essa è proprio quel fenomeno che si pone la domanda sui fenomeni, sarà sempre un passo più in là rispetto alla possibilità di inquadrarla in una teoria. Il modello operativo a cui portano le teorie sul linguaggio, le teorie sull'inconscio e quelle sul comportamento può essere un surrogato, sempre perfettibile, che imita la coscienza. Tale surrogato potrà forse in futuro distinguersi da una coscienza naturale in modi che le coscienze naturali non noteranno, ma non si tratterà mai di una coscienza come la nostra.

Persiste nel dibattito odierno sulla IA un equivoco fra i termini “intelligenza” e “coscienza”. La prima intelligenza che percepiamo è la nostra, e questa è connessa ad una coscienza. Sembra quindi – per il fatto che la “coscienza” artificiale costituisce un problema insormontabile – che l'espressione “Intelligenza Artificiale” sia grandemente pretenziosa. Tale denominazione è una allusione antropomorfizzante ingiustificata al fatto che macchine deterministiche complesse (ma a causalità lineare) sarebbero un buon

modello per spiegare la mente. Gli ottimisti della IA debole sostengono poi che l'intelligenza artificiale possa raggiungere qualsiasi obiettivo senza alcun bisogno di coscienza. Quanti credono nell'intelligenza della IA a prescindere dalla coscienza guardano da un punto di vista evolutivo a forme di vita, abbastanza semplici o unicellulari, che sarebbero in grado di svolgere le funzioni vitali (nutrirsi, riprodursi, muoversi) senza che si postulino in essi alcuna forma di soggettività "cosciente". Chi crede che l' "intelligenza" artificiale possa funzionare e svilupparsi senza risolvere il problema della coscienza dovrebbe rinunciare al nome "Intelligenza Artificiale", dal momento il fenomeno biologico della "intelligenza", per come lo conosciamo, è sempre connesso a forme (anche embrionali) di coscienza. Inoltre tale intelligenza è tanto più sviluppata quanto la coscienza è più presente. Non esiste poi alcuna prova che le forme di vita unicellulari siano esenti da qualche forma embrionale di sentire soggettivo. Anche se coscienza ed intelligenza non si identificano del tutto, solo un essere cosciente può essere capace del comprendere intellettivo: non possono esserne capaci dispositivi atti ad imitare malamente poche funzioni dell'agire umano, assemblati dall'uomo come strumenti di elaborazione algoritmica di *inputs* e privi di autonomia. L'intelligenza cresce col complicarsi filogenetico della coscienza, ma intelligenza e coscienza non indicano lo stesso fenomeno: l'intelligenza è una attitudine più indeterminata della coscienza, ma separarla da questa è problematico. Trascurare queste evidenze ha portato a ipotesi assurde, quali l'idea eliminativista di Chalmers (1996: 153-6), secondo cui possono esistere *zombie* intelligenti, esseri artificiali che riescono a fare tutto ciò che fanno gli umani sebbene privi di coscienza.

Secondo gli eliminativisti, le odierne neuroscienze ci mostrerebbero quanta intelligenza impieghiamo nell'espletare le funzioni della nostra vita senza impiegare la coscienza. Questa strana tesi è una narrazione finalizzata a ratificare come "intelligenti" macchine che non possono essere dotate di coscienza. È la solita strategia dei post-funzionalisti: limitano le pretese della IA, per far sì che le macchine prodotte dall'industria informatica soddisfino le promesse commerciali. Il ridimensionamento della coscienza avrebbe la conseguenza di vanificare uno dei più insidiosi argomenti rivolti contro la IA: l'esperimento mentale della "stanza cinese" di Searle (1980). Gli ottimisti di IA sostengono che questo esperimento sia una confutazione contro la coscienza dei calcolatori, ma non certo contro la loro intelligenza. In base al test di Turing, l'intelligenza di un calcolatore avrebbe potuto essere dimostrata tramite una conversazione fra un computer nascosto ed un umano: nel caso in cui l'umano non fosse riuscito a capire che stava conversando con un calcolatore e non con un altro umano, tale calcolatore avrebbe dovuto essere considerato "intelligente". Searle confuta questa idea immaginando una persona cinese che conversa tramite messaggi in cinese introdotti in una stanza chiusa. Dentro la stanza c'è un uomo che non comprende il cinese e che utilizza un manuale per processare il messaggio ricevuto e produrre la risposta in cinese. Searle dimostra così che il sistema, pur superando il test di Turing, non è per niente intelligente: sebbene l'uomo nella stanza non capisca affatto il cinese, la conversazione procede. Gli eliminativisti commentano l'esperimento sostenendo che "dal lato della stanza non c'è coscienza di quel che viene detto nei messaggi", ma questo non impedisce che la macchina sia intelligente. L'esperimento della stanza cinese però non punta affatto, ritengo, a negare la (inesistente) coscienza delle macchine, bensì a negare nelle macchine la "comprensione semantica e operativa" che gli umani hanno grazie al linguaggio mentre altre specie animali grazie alle loro semiotiche. Gli uomini (e gli animali in genere) vivono in un ambiente di rimandi semiotici, le macchine invece non sono "soggetti" che distinguono fra sé stessi ed il loro ambiente: non hanno intenzionalità, quindi non si può ascrivere loro alcun "intelligere", nel senso del latino "intus legere", cioè "interpretare eventi dell'ambiente come fatti decisivi per la loro esistenza". È praticamente scontato che certe specie animali, pur avendo forme di coscienza embrionali, riescano a vivere la loro vita semiotica tramite comprensione delle modifiche che si verificano nel loro ambiente, rapportandole al loro vivere concreto. Quello che manca alla scatola cinese non è quindi la coscienza ma,

semplicemente, la rete dei rimandi semiotici con l'ambiente in cui è immerso il vivente. In base a questa definizione di "intelligenza", è chiaro che le IA di oggi non hanno alcuna "intelligenza": la coscienza (di sé) è su un livello ulteriore che non viene affatto toccato dall'esperienza mentale di Searle. Eppure la diagnosi clinica automatica (cfr. Licata, 2007), i *software* di dialogo e i sistemi di guida autonoma, sono considerati da alcuni teorici "pienamente intelligenti" e privi di coscienza. In realtà, in questi sistemi non c'è alcuna intelligenza, perché essi non vivono di rimandi semiotici fra l'ambiente ed il sé, non sono quindi nelle condizioni di *intelligere*.

2. Mente, coscienza e IA

Una delle strategie seguite da quanti sostengono questa forma debole di IA è la negazione della importanza della coscienza e, conseguentemente, la sua eliminabilità dai progetti di IA. Si dice ad esempio che, in base alla psicoanalisi, sarebbe chiaro che a) il comportamento umano è dovuto più all'inconscio che alla coscienza, e che b) siamo così limitati nell'accedere alla nostra personalità tramite la nostra coscienza che un analista ha molto più successo di noi nel comprendere i meccanismi latenti che producono il nostro comportamento (cfr. Plebe, 2021). Questo modo di impostare il problema – che viene impiegato da alcuni autori anche per negare il libero arbitrio – è insostenibile, giacché anche l'analista opera tramite coscienza, ed anche a lui rimangono nascosti ambiti della personalità del soggetto analizzato (che al soggetto sono invece noti). La personalità è una dinamica risultante dal contrasto fra forze latenti e vettori coscienti che è troppo riduttivo liquidare come sottoposta al dominio dell'inconscio. Un terzo genere di argomentazioni impiegato dagli eliminativisti è quello secondo cui c) la coscienza sarebbe una attività ingannevole del cervello. Secondo Metzinger (2003, 2007), Gazzaniga (2018), e Dennett (2005), esisterebbe nel nostro cervello un insieme di processi paralleli fra i quali, ogni tanto, si costituirebbe un'unificazione provvisoria: l'illusione di un "io". Secondo un'ottica storica l'eliminativismo è una adesione acritica alle tesi di desostanzializzazione dell'io proposte da Hume contro le posizioni di Cartesio e di Leibniz.

Posto che l'attività del cervello si sviluppa in moduli di attivazione indipendenti, il problema sarà spiegare come mai la nostra coscienza ci presenti la realtà come un'entità unitaria. Metzinger (2003) individua questo "principio unificatore" in un meccanismo di sincronizzazione formato da gruppi di neuroni posti in diversi distretti cerebrali ma collegati tra loro. La loro attivazione, ponendo in relazione processi che si svolgono in zone del cervello separate, darebbe luogo all'esperienza di percezione di totalità che sperimentiamo. Gazzaniga (2018) propone una congettura insieme emergentista ed antidualistica. Il neuroscienziato recupera, in un'ottica aristotelica, il finalismo: il cervello opera una serie immensa di "computazioni" senza che ce ne avvediamo, e tutto ciò ha il proprio fine nella realizzazione della vita cosciente. Il suo monismo si sostanzia nel controllo che i livelli più alti del sistema, che sono il risultato non deterministicamente prevedibile delle computazioni dei livelli inferiori, esercitano su questi ultimi: nessuno spirito disincarnato di tipo cartesiano viene postulato. La coscienza sarebbe una proprietà pluri-locale e distribuita del cervello, emergente come prodotto finale dell'elaborazione di ogni modulo. Come in una pentola d'acqua in ebollizione, dall'attività cerebrale emergerebbero di continuo alla superficie della coscienza alcune "bolle", imponendosi su altre "bolle" possibili, allineandosi le une dopo le altre secondo l'asse temporale. La concezione proposta da Dennett della coscienza come momentanea "fama nel cervello" di particolari processi rispetto ad altri, si avvicina molto alla similitudine delle "bolle" di Gazzaniga. L'ipotesi di Dennett (2005) si basa sui cosiddetti *qualia*: i colori, i suoni, i gusti ecc. così come ci appaiono nella nostra esperienza privata. Secondo il filosofo americano tali proprietà epifenomeniche dell'esperienza soggettiva semplicemente "non esistono". Nessun ente può soddisfare gli speciali requisiti che i *qualia* dovrebbero avere: l'accessibilità esclusiva all'introspezione, il carattere fenomenico (avere un *quale* fa un

certo effetto in prima persona) e il carattere intrinseco (i *qualia* non sono analizzabili in termini di relazioni con altro). Queste caratteristiche rendono i *qualia* degli elementi primitivi refrattari a qualsiasi analisi oggettiva, e la loro incoerenza è un punto fondamentale dell'argomentazione con cui Dennett destituisce le nozioni di coscienza e di io. L'io è un "comandante virtuale", il personaggio di una narrazione creata dal nostro stesso cervello per dare un senso all'esperienza. Anche nel caso della obiezione eliminativista c), secondo cui la coscienza sarebbe un costrutto ingannevole del cervello, siamo di fronte ad un tentativo di sgonfiare, convenientemente, un problema irrisolvibile. Ma è sufficiente leggere il brillante saggio di Nagel (1974) *What is it like to be a bat?*, i lavori di Searle (1983, 1990) e di tanti altri (cfr. Revonsuo, 2010: 30) per avere chiaro che dalle descrizioni oggettive dell'attività neuronale, non siamo quasi mai in grado di trarre informazioni sulla controparte "mentale" e soggettiva dell'attività stessa. Non c'è modo di mettere in parallelo dati neurofisiologici e aspetti soggettivi, figuriamoci negare qualcosa di manifesto come l'io sulla base delle deboli ipotesi dei neuroscienziati sulle attivazioni neurali.

Uno dei primi modelli di coscienza ingegnerizzabile, basato sulla integrazione di processi paralleli proposta da Metzinger e Gazzaniga, è stato lo spazio globale di Bernard Baars (1988; 1997). I pochissimi elementi presenti in ogni istante nella coscienza corrispondono ai processi posti al centro del "palcoscenico" e illuminati da un riflettore (*spotlight*): solo una minima parte di ciò che avviene sul palco viene illuminato. Persuaso della irriducibilità della coscienza ma anche della sua misurabilità, Tononi (2012), seguendo Baars, descrive il modo in cui la coscienza viene generata, e come sia possibile misurarla. La teoria dell'informazione integrata prende le mosse dall'esperienza di prima persona e ne identifica le proprietà essenziali; da questa descrizione trae le caratteristiche necessarie affinché un sistema fisico possa generare la coscienza. Tononi propone un calcolo riguardante due caratteristiche di base alle quali attribuisce un valore matematico: lo stato di coscienza e l'attività neurale del sistema talamo-corticale. Giunge così ad una scala di riferimento del "livello di coscienza": l'indice di complessità perturbativa (PCI). Nulla impedirebbe, in linea di principio, di definire cosciente un robot, qualora risultasse avere il giusto valore di PCI: la coscienza è una struttura irriducibile, che però può essere replicata da qualsiasi sistema fisico *organizzato* in un certo modo specifico.

Le comunità di diverse specie viventi esigono dagli individui una capacità peculiare: prevedere le azioni degli altri. La mente umana gestisce il *mindreading*, un circuito di elaborazione grazie al quale comprende le intenzioni altrui e che, secondo Carruthers (2011), costituisce la matrice generativa della coscienza. La coscienza che ho dei miei stati mentali è evidente in un grado molto superiore rispetto a quella che ho degli stati mentali altrui, ma è comunque interpretativa. Il fatto che io sia a contatto coi miei stati mentali non implica che questi stati siano realmente come li sento. L'auto-percezione per Carruthers non esiste: al suo posto abbiamo un'auto-interpretazione, una riflessione di tipo confabulatorio che il soggetto produce a partire da dati percettivi. Se non si danno mai pensieri puramente consci, nel senso di pensieri che producono direttamente azioni e deliberazioni in virtù del loro essere evidenti, non esistono di fatto azioni cosce e libere. L'intenzione non sarebbe altro che una rappresentazione di ciò che sta per avvenire sulla base di dati percettivi. In realtà è sufficiente fare una analisi semantica della espressione del linguaggio quotidiano "fare un esame di coscienza", o produrre analisi controfattuali sulle azioni che abbiamo compiuto in passato, per renderci conto di quanto la negazione del libero arbitrio, proposta da Carruthers e da altri scienziati cognitivi, sia del tutto insostenibile.

Sarà mai possibile dotare un calcolatore di coscienza? A questa domanda si risponde solitamente in due modi: 1) gli scettici sostengono che la cosa è impossibile per principio; 2) gli ottimisti sono convinti che il progresso tecnico porterà a tale risultato. Molti ottimisti credono nel funzionalismo, cioè nel fatto che gli stati mentali (credenza, desiderio, ecc.) sarebbero qualificati da funzioni simili a quelle matematiche, e non conterebbe nulla la costituzione materiale del supporto nel quale si realizzano. Gli scettici

sono spesso naturalisti biologici: pensano che la coscienza sia un fenomeno biologico realizzabile esclusivamente grazie ad un sistema nervoso. I primi sostengono che gli stati mentali possono essere ridotti ad algoritmi, i secondi affermano che il nostro sistema nervoso è enormemente più complesso ed indeterministico degli algoritmi messi a punto finora. Il dibattito fra Searle e Dennett riguarda proprio questo punto. Il naturalismo pone questioni esiziali per il funzionalismo: può una macchina “pensare” semplicemente per il fatto che esegue un programma? (Searle, 1998); è sufficiente progettare il giusto *software* per replicare una mente? Searle (1998) risponde in modo negativo ad entrambi gli interrogativi, dal momento che, a suo modo di vedere, un programma eseguito da un calcolatore è soltanto un’elaborazione sintattica di informazioni, mentre la mente svolge funzioni semantiche.

Molti esperti di cognizione ritengono che la coscienza sia una proprietà “emergente”. La coscienza sarebbe un fenomeno risultante dalla connessione olistica delle attività distribuite in tutti i distretti del corpo, aventi come centro l’encefalo. Su questo tema vi sono due linee di pensiero. I) Una ritiene che le condizioni standard di emergenza, quelle che riguardano le nostre attuali conoscenze sui neuroni, sono sufficienti per rendere conto dell’emergere degli stati di coscienza. È questa la teoria dell’emergenza debole o quella che viene chiamata da Searle “emergenza 1”. II) L’altra posizione è chiamata emergenza forte, “emergenza 2” (Searle, 1992) o “emergenza radicale”. Questa tesi sostiene che nessuna proprietà riscontrabile nei neuroni sarà mai in grado di spiegare in modo oggettivo le caratteristiche dell’esperienza soggettiva (*qualia*), e ricondurle a ciò che viene osservato oggettivamente nel sistema nervoso. Possiamo però optare per una via intermedia: pur ammettendo che dotare un sistema artificiale di veri e propri *qualia* è impossibile, non possiamo escludere a priori che una macchina possa essere costruita in modo tale da avere una qualche forma di “intenzionalità” artificiale.

3. La teoria di Santiago e la natura della coscienza

Un sistema “autopoietico”, nel lavoro di Maturana e Varela (1985; 1987), è un sistema che ricostruisce continuamente sé stesso, si sostiene e si riproduce dal proprio interno. Un criterio distintivo della vita è il mantenimento della sua stessa organizzazione attraverso il mutamento. Tale mantenimento, però, non è sufficiente a qualificare un sistema come “autopoietico”; il mantenimento si trova infatti anche in sistemi auto-organizzanti semplici come le strutture dissipative (i mulinelli degli scarichi, gli uragani, le stelle). Ciò che caratterizza il mantenimento di organizzazione dei viventi – intesi sempre come strutture dissipative – è “la produzione continua di elementi di base che sostengono la produzione stessa”. Una tale “chiusura operativa” non ha né input né output, è un sistema in sé omeostatico (sebbene in disequilibrio energetico) identificato dall’autonomia entro l’ambiente. Il vivente non ha in sé stesso tutte le sue proprie cause; si sviluppa in un ambiente con il quale è in “accoppiamento strutturale”. L’individuo è in grado di condizionare le cause esterne lasciandone filtrare alcune, in modo che queste possano dare effetto entro il proprio organismo, secondo le proprie esigenze. I confini del sistema lo separano dal suo ambiente e lo mettono in relazione con esso. Questa chiusura organizzativa del vivente rispetto all’ambiente è il segreto primario della coscienza, perché essere coscienti significa “creare/crearsi un mondo”. Maturana (1987: 62) riprende l’etimologia greca della parola organizzazione: il termine indica la distribuzione strumentale dei componenti di un sistema nella costituzione di un’unità meccanica finalizzata. L’organizzazione di un sistema definisce la sua identità, ma l’“istanziamento” delle relazioni definite per un’organizzazione si produce tramite componenti fisiche. Viene quindi definito il concetto di “struttura” come il plesso di componenti che costituiscono una unità nella realizzazione concreta della sua organizzazione. Una unità può modificarsi nella struttura senza perdere l’identità, se mantiene la sua organizzazione (cfr. Maturana e Varela, 1985: 129-130). Questa tesi è molto importante per l’ingegnerizzazione della

coscienza. In base ad essa non si può escludere che una macchina, che implementi nella sua struttura la organizzazione giusta, possa sviluppare una forma di coscienza (cfr. Tononi, 2012).

L'autopoiesi cellulare produce un contorno entro cui contenere la rete delle trasformazioni metaboliche: la membrana è condizione del metabolismo ed il metabolismo è condizione della membrana. È questa la chiusura del sistema che realizza un dentro e un fuori. È questa la base dell'intenzionalità. Il punto radicale dell'emergere della coscienza, nella filogenesi, ritengo sia quello coincidente con l'emergere del vivente nel suo ecosistema. Ovviamente si tratta di uno stadio assolutamente embrionale di soggettività, ma qualsiasi attribuzione di un punto di realizzazione della coscienza, nella filogenesi, che non sia l'inizio della vita sarebbe arbitraria e ingiustificata. Un sistema cognitivo che opera distinzioni che scindono alcuni oggetti da uno sfondo è un "osservatore" (Maturana e Varela, 1985: 53). La questione dell'osservatore inquadra il fatto che la coscienza può anche essere considerata, come oggetto di studio, un "fenomeno" (anche se la coscienza, come soggetto, è proprio il luogo del formarsi dei fenomeni). Abbracciamo così una possibilità esplicativa che rinuncia per principio alla esaustività. L'introspezione e lo scambio di informazioni fra umani può fornire un buon resoconto del lato soggettivo della coscienza, ma su tali basi non è possibile fondare una scienza osservativa: il fenomeno che noi stessi sempre siamo sfugge per principio a misurazioni ed esperimenti ripetibili; la descrizione della coscienza riesce molto meglio alla letteratura che alla scienza. Quando trasformiamo la coscienza, che è la base del darsi dei fenomeni in noi, in un fenomeno a sua volta, ci troviamo alle prese, come nel caso della informalizzabilità del linguaggio naturale, col punto di vista di "impossibile" descrizione totale dell'occhio che guarda sé stesso. Proprio come il linguaggio naturale non ha le risorse semantiche per formalizzare sé stesso (Tarski, 1933), allo stesso modo la coscienza non ha le risorse per la descrizione esaustiva dei processi che si svolgono in sé stessa (o in una coscienza altrui).

Interagendo coi conspecifici si creerà per il vivente un accoppiamento strutturale di ordine superiore: la socialità. La socialità è una coordinazione reciproca fra i viventi, la comunicazione un coordinamento comportamentale peculiare fra conspecifici. Ci realizziamo nell'accoppiamento linguistico, non perché il linguaggio sia il nostro mezzo di comunicazione, ma perché esistiamo nel linguaggio, nella *logosfera* (cfr. Maturana e Varela, 1985: 174). Il linguaggio può apparire come il campo degli scambi dei "significati mentali" veicolati dalle espressioni. Per comprendere la questione del significato è necessario tenere ferma la distinzione fra osservatore ed entità osservata, distinguendo la fisiologia di un'attività dalla sua descrizione. In questo modo si può venire a capo dell'apparente contraddizione tra ciò che il linguaggio sembra essere – la facoltà di produrre "espressioni significative" – e ciò che in realtà è: un dominio di comportamenti linguistici coordinati (in linea con la teoria dei giochi linguistici di Wittgenstein). La confusione è dovuta al fenomeno della metalinguisticità: il fatto che vi può essere un comportamento comunicativo linguistico che viene indicato tramite un altro comportamento comunicativo linguistico. L'accoppiamento comportamentale fra due individui (fisiologia) può contenere come termini espressioni di un altro comportamento, generando riferimenti metalinguistici: questi riferimenti generano l'illusione del rapporto fra espressione e contenuto (descrizione). Tale processo sarebbe alla base della ricorsività descrittiva di prima persona che chiamiamo "io" e che ci permette di effettuare il riferimento al nostro sé (*ibidem*: 195). Non vi sarebbe pertanto un io-sostanza che agisce, ma solo l'agire che, in seconda istanza, genera l'illusione di un io (cfr. Maturana, 1993).

Ritengo che parlare di "illusione di un io" sia però fuorviante. Probabilmente non esistono strutture neuronali stabili che costituiscono l'io, ma nulla ci autorizza ad eliminare dal campo di ciò che esiste realtà che non abbiano un supporto materiale evidente; inoltre con la parola "io" noi indichiamo, nel linguaggio ordinario, la nostra persona nella sua interezza. Il linguaggio permette, a chi lo usa, di descrivere sé stesso

riferendosi agli elementi del comune dominio culturale; ma quel “sé stesso”, che l’individuo descrive e a cui si riferisce tramite il linguaggio, ci sarebbe anche senza il linguaggio che dice “io”, così come il nostro cane ha una sua “individualità”, il suo carattere specifico, la sua memoria, ed il senso del sé pur non avendo il linguaggio verbale.

4. Identità di mente e corpo

La coscienza è un ricreare un mondo-ambiente da parte di un individuo vivente, sulla base di una realtà oggettiva. Ho negato che i sistemi non viventi possano avere una coscienza analoga alla nostra, e allo stesso tempo ho prospettato, in tali sistemi, la possibilità di una approssimazione sempre più soddisfacente della coscienza. I neuroscienziati cercano negli schemi di attivazione neuronale la soluzione ai problemi della coscienza, ma prima di essere divisa in neuroni la rete cerebrale esprime qualcosa di unitario, qualcosa che riplasma e ripropone, ad ogni grado di complessità evolutiva, la cognizione del vivente unicellulare. La coscienza è come un’apertura procedente nello spazio e nel tempo: un’apertura che sente sé stessa ed il proprio mondo interno come incorporato. Essa è il centro di *irradiazione* del mondo-ambiente. Lo stesso corpo che la incorpora è “parte” del mondo esterno, pur essendo proprio questo corpo senziente l’entità da cui si irradia l’apertura creatrice. È questa la base del dualismo storico mente-corpo, il quale, da un certo punto di vista, costituisce un’evidenza inoppugnabile. La teoria di Santiago ci mostra come superare il dualismo: autopoiesi e cognizione sono le chiavi per comprendere la vita e la coscienza; non è necessario postulare nessuno spirito disincarnato che vivificherebbe il corpo materiale per spiegare il mistero dello spirito. Tale mistero è tutto nella complessità dell’ordine superiore che la biochimica apporta nella *materia vivente* con l’avvento della vita. Sostengo in questo modo il monismo ma non come materialismo: materia, energia, spirito sono categorie esplicative storiche alle quali non possono essere mai ridotti interamente misteri come la vita e l’anima.

Senza tenere conto della straordinaria complessità dei meccanismi biochimici che producono il pensiero e, in generale, quel che chiamiamo “spirito”, sono state recentemente sollevate da Sesta (2022) difficoltà come la seguente: secondo il fisicalismo, sia che a) “il pensiero sia causato da processi cerebrali” sia che b) “si identifichi con essi”, rimane il problema di un effetto (il pensiero) per il quale è necessario trovare una causa adeguata: questa, argomenta Sesta, non può essere rinvenuta in meccanismi materiali le cui caratteristiche sono del tutto inconciliabili col loro effetto spirituale. Non si tratta del solo *gap* fra descrizione scientifica in terza persona ed esperienza soggettiva, ma di una vera sproporzione fra causa ed effetto, una volta che si faccia degli stati cerebrali la matrice esclusiva degli eventi psichici, o c) la materia stessa di cui tali eventi psichici rappresentano una semplice descrizione funzionale (Sesta, 2022: 201). Rispondo che se b) il pensiero si identifica coi processi cerebrali, e se c) la fisiologia cerebrale è la materia di cui gli eventi mentali sono solo una descrizione (tesi che condivido entrambe), cade qualsiasi *gap* e qualsiasi sproporzione fra causa ed effetto. Non è questo un materialismo che elimina la componente spirituale: la teoria di Santiago spiega bene come quel che chiamiamo “materia” può generare il pensiero. Semplicemente i corpi viventi sono dotati di *perceptio*, *appetitus* e pensiero: il mistero non è nella distinzione fra materia e spirito, ma nella distinzione fra materia vivente e materia inanimata.

Il monismo che mi pare più coerente è allora quello espresso da Nagel (2015: 61): «la coscienza non è un effetto dei processi del cervello che costituiscono le sue condizioni fisiche; piuttosto, quei processi cerebrali sono essi stessi più che fisici». Al fondo di un evidente dualismo fra materia e spirito, figlio di una auto-intuizione fortissima, il monismo proposto dalla teoria di Santiago, coerente con la posizione di Spinoza (*Ethica*, II, prop. VII), può essere inteso anche nel senso di Nagel. In questa ottica superare il dualismo significa non porre più il problema come se rimanesse aperta la questione dell’origine degli stati mentali dagli stati corporei, per il fatto che oggi persiste un vuoto

esplicativo fra la descrizione dell'eccitazione neurofisiologica e il contenuto psichico corrispondente (Noë e Thompson, 2004). Sesta (2022: 202-203) sostiene che qualunque soluzione dovrebbe rimanere fedele al "doppio aspetto" fisico e spirituale, senza dissolvere l'uno nell'altro, ma anche senza farli interagire secondo rapporti di causa ed effetto. Se il dualismo delle sostanze è troppo – "esistono due cose in una ma non si riesce a spiegare la loro unità" –, il dualismo delle proprietà è troppo poco: "esiste una sola cosa ma non si spiega la diversità descrittiva fra gli aspetti fisici e quelli mentali" (Sesta, 2022: 203). Ritengo che il solo dualismo accettabile sia quello fra le proprietà descrittive degli eventi cerebrali, che presentano aspetti materiali e aspetti che la tradizione chiama "spirituali", ma che altro non sono che la vita cosciente dell'individuo: il vero mistero qui rimane la vita, non il fatto che i corpi viventi siano materia mentre la loro attività di pensiero sarebbe immateriale o spirituale. Il ragionamento di Sesta (2022: 203), più che dimostrare il dualismo fra materia e spirito, in realtà lo presuppone. La cartesiana distinzione fra mente e corpo esclude che stati mentali e stati corporei possano essere identificati. Gli stati mentali sono soggettivi e inosservabili, non localizzabili e non divisibili, privi di peso e dimensione; gli stati corporei sono osservabili, sono divisibili e occupano uno spazio, hanno peso e dimensione. Dal mio punto di vista, gli stati mentali *sono* gli stati corporei: non che i primi siano spirituali immateriali e ideali, mentre i secondi estesi e materiali. Semplicemente i corpi viventi hanno percezione, pensiero e desiderio, e questo è dovuto al fatto che sono vivi. Non ci sono operazioni mentali "che sbucano" in un soggetto fisico che non è una sostanza mentale (cfr. Sesta, 2022: 203). La mia posizione non è quindi un eliminativismo materialista, sebbene al pari di questo identifichi stati mentali ed eventi cerebrali, ma un monismo della sostanza sul quale si determina poi un dualismo delle proprietà descrittive degli eventi cerebrali (cfr. Pennisi, 2021).

Una coscienza ha sempre un mondo esterno ed un mondo interno: essa è la soglia fra questi due mondi; il mondo interno è anche un mondo di memoria, una storia dell'io che ci raccontiamo (cfr. Kahneman, 2011). La coscienza è come la superficie di "emersione dal" e di "immersione nel" mondo dei processi che la mente opera inconsciamente: questo flusso è la vita della mente di cui la coscienza è il limine "divisivo", per lo più involontario. La coscienza è il "fra" che scinde mondo interiore e mondo esteriore, che in modo incomparabile li commistiona e li mette in relazione. La coscienza è, in senso husserliano, il nostro essere: è su questa base che Nagel può condurre la sua argomentazione sulla difficoltà di immaginare cosa si prova ad "essere" un pipistrello. La *res cogitans* non è quindi una *res*, non certo nel senso di un oggetto come gli altri: essa è lo "spazio" della possibilità delle cose, del ricostruire la realtà esterna e dell'avere a che fare con una realtà interna.

Riferimenti bibliografici

Aristotelis *De Anima*, recognovit brevisque adnotatione critica instruxit W.D. Ross, Oxford University Press, Oxford, 1956.

Baars, B., 1988, *A Cognitive Theory of Consciousness*, Cambridge University Press, Cambridge and New York.

Baars, B., 1997, *In the Theater of Consciousness: The Workspace of the Mind*, Oxford University Press, Oxford and New York.

Bedau, M.A., 1997, *Weak Emergence*, in "Noûs" 31, 11, pp. 375-399.

Bengio, Y., 2009, *Learning Deep Architectures for AI*, in "Foundations and Trends in Machine Learning", 2 (1), pp. 1-127, Now, Boston-Delft.

Bengio, Y., 2019, *The Consciousness Prior*, in “ArXiv”, abs/1709.08568v2.

Brentano, F., 1997, *Psicologia dal punto di vista empirico*, L. Albertazzi (ed.), Laterza, Roma - Bari.

Carruthers, P., 2011, *The Opacity of Mind. An Integrative Theory of Self-Knowledge*, Oxford University Press, Oxford.

Chalmers, D., 1996, *The Conscious Mind: In Search of a Fundamental Theory*, Oxford University Press, Oxford.

Chalmers, D., 2006, *Strong and Weak Emergence*, in P. Davies, P. Clayton (eds.), *The Re-Emergence of Emergence: The Emergentist Hypothesis from Science to Religion*, Oxford University Press, Oxford, pp. 244-254.

Chalmers, D., 2011, *A computational foundation for the study of cognition*, in “Journal of Cognitive Science” 12, (4), pp. 323-357.

Clayton P., Davies P., 2006, *The Re-Emergence of Emergence: The Emergentist Hypothesis from Science to Religion*, Oxford University Press, Oxford.

Dennett, D.C., 2005, *Sweet Dreams. Philosophical Obstacles to a Science of Consciousness*, MIT Press, Cambridge (Ma).

Descartes, R., *Meditationes de prima philosophia, in qua Dei existentia et animae immortalitatis demonstrantur*, M. Soly, 1641; trad. it. di S. Landucci, *Meditazioni metafisiche*, Laterza, Roma - Bari, 1997.

Feinberg, T.E., 2001, *Why the mind is not a radically emergent feature of the brain*, in “Journal of Consciousness Studies” 8, 9-10, pp. 123-145.

Feinberg, T.E., Mallatt, J., 2020, *Phenomenal Consciousness and Emergence: Eliminating the explanatory gap*, in “Frontiers in Psychology”, 11, p. 1041.

Franklin S., Patterson F.G., 2006, *The LIDA Architecture: Adding New Modes of Learning to an Intelligent, Autonomous, Software Agent. Integrated Design and Process Technology*, Society for Design and Process Science, San Diego CA.

Gazzaniga, M., 2018, *The Consciousness Instinct: Unraveling the Mystery of How the Brain Makes the Mind*, Farrar Straus & Giroux, New York City; trad. it. di F. Peri, *La coscienza è un istinto. Il legame misterioso tra il cervello e la mente*, Raffaello Cortina Editore, Milano, 2019.

Hume, D., 1739-1740, *A Treatise of Human Nature*; trad. it. A. Carlini, E. Mistretta in D. Hume, *Opere filosofiche*, Laterza, Roma-Bari, 2010.

Kahneman, D., 2011, *Thinking, Fast and Slow*, Ferrar, Straus and Giroux, New York.

Licata, G., 2002, *Teoria dei nomi e teoria delle idee in Platone*, in “Dialegesthai”, 4, 1, 2002, in: <<https://mondodomani.org/dialegesthai/articoli/gaetano-licata-01>>

Licata, G., 2005, *Categorie logico-ontologiche, categorie sintattico-semantiche e unità mente-corpo*, in C. Rosciglione e G. Zanet (eds.), *Mente e corpo. Aspetti filosofici e scientifici*, in “Studium Philosophicum”, 4 (2005), pp. 113-115.

Licata, G., 2007, *Precision and Uncertainty. Fuzzy Logic in Clinical Diagnosis*, Selecta Medica, Pavia.

Licata, G., 2013, *Many-valued Logics and Quantum Mechanics*, in G. Licata, L. Sesta (eds.), *Philosophical essays on language, ontology and science*, Franco Angeli, Milano.

Maturana H.R., Varela F.J., 1980, *Autopoiesis and Cognition. The Realization of the Living*, Boston Studies in the Philosophy and History of Science (Vol. 42); trad. it. *Autopoiesi e cognizione. La realizzazione del vivente*, Marsilio, Venezia, 1985.

Maturana H.R., Varela F.J., 1987, *The Tree of Knowledge: The Biological Roots of Human Understanding*, Shambhala, Boston.

Maturana, H.R., 1993, *Autocoscienza e Realtà*, Raffaello Cortina, Milano.

Metzinger, T., 2003, *Being No One. The Self-Model Theory of Subjectivity*, MIT Press, Cambridge (Ma).

Metzinger, T., 2007, *The Ego Tunnel: The Science of the Mind and the Myth of the Self*, Basic Books, New York City.

Nagel, T., 2012, *Mind & Cosmos. Why the Materialist Neo-Darwinian Conception of Nature is Almost Certainly False*, Oxford University Press, Oxford; trad. it. di M. Di Francesco, *Mente e cosmo. Perché la concezione neodarwiniana della natura è quasi certamente falsa*, Raffaello Cortina, Milano, 2015.

Nagel, T., 1974, *What is it like to be a bat?*, in "The Philosophical Review", 83, 4, pp. 435-450.

Noë A., Thompson E., *Are There Neural Correlates of Consciousness?*, in "Journal of Consciousness Studies", 1 (2004), pp. 3-28.

Pennisi, A., 2021, *Che ne sarà dei corpi? Spinoza e i misteri della cognizione incarnata*, Il Mulino, Bologna.

Plebe, A., 2021, *Una coscienza per l'intelligenza artificiale: il punto su teorie e sviluppi*, in: <<https://www.agendadigitale.eu/cultura-digitale/intelligenza-artificiale-coscienza/>>

Revonsuo, A., 2010, *Consciousness. The Science of Subjectivity*, Psychology Press Ltd, London.

Searle, J.R., 1980, *La mente è un programma?*, in "Le Scienze", 259, Marzo 1990.

Searle, J.R., 1983, *Intentionality: An Essay in the Philosophy of Mind*, Cambridge University Press, Cambridge.

Searle, J.R., 1992, *The Rediscovery of the Mind*, MIT Press, Cambridge.

Searle J.R., 1990, *The Mystery of Consciousness*, New York Review Books, New York City; trad. it. di E. Carli, *Il mistero della coscienza*, Raffaello Cortina, Milano, 1998.

Sesta, L., 2022, *Una certa dualità. Il problema mente-corpo e il paradosso della thinking matter*, in "Giornale di Metafisica", 1 (2022), Anno XLIV, pp. 193-207.

Spinoza, B., 1677, *Ethica Ordine Geometrico Demonstrata*, in Spinoza, *Opere*, F. Mignini (ed.), Mondadori, Milano, 2007.

Tarski, A., 1933, *Pojęcie prawdy w językach nauk dedukcyjnych*, Warsaw: Nakładem Towarzystwa Naukowego Warszawskiego; English Translation by J.H. Woodger, *The Concept of Truth in Formalized Languages*, in A. Tarski, *Logic, Semantics, Metamathematics*, second edition by J. Corcoran, Hackett, Indianapolis, 1983, pp. 152-278.

Tononi, G., 2012, *PHI: A Voyage from the Brain to the Soul*, Pantheon, New York City; trad. it. di S. Ferraresi, *PHI. Un viaggio dal cervello all'anima*, Codice, Torino, 2017.

Wittgenstein, L., 1953, *Philosophical Investigations*, edited by G.E.M. Anscombe, R. Rhees, Blackwell, Oxford, 1953; trad. it. di M. Trinchero, *Ricerche filosofiche*, Einaudi, Torino, 2009.

