Intelligenza e Coscienza nell'Intelligenza Artificiale: un'analisi filosofico-scientifica

Corrado Baccheschi

20 giugno 2024

Indice

1 Introduzione

Negli ultimi anni, si è assistito a un enorme sviluppo dell'Intelligenza Artificiale (IA) e, conseguentemente, al dibattito filosofico-scientifico sull'intelligenza e la coscienza delle macchine (Butlin et al. 2023). L'interesse per l'IA non si limita alle sue applicazioni pratiche, ma si estende anche alla comprensione dei concetti di *intelligenza* e *coscienza*, sia in contesti biologici che artificiali. Questa ricerca mira a esplorare tali concetti, cercando di definire cosa significhi realmente "intelligenza" e "coscienza" nel contesto delle macchine intelligenti.
L'analisi parte nel Capitolo 2 e dalla definizione di intelligenza, un termine complesso che ha visto numerose interpretazioni nel corso della storia, coinvolgendo discipline come la psicologia, la biologia, le neuroscienze e la filosofia. Distinguere tra intelligenza umana e non umana è essenziale per comprendere le diverse sfaccettature del termine e per affrontare il tema dell'intelligenza artificiale.
Successivamente, nel Capitolo 3 viene approfondito il concetto di coscienza, esaminando le teorie filosofiche e scientifiche che cercano di spiegare questo fenomeno (Sez. 3.1). La discussione include il problema difficile della coscienza, le teorie materialiste e dualiste, e approcci singolari come il panpsichismo. Si passa poi alla Sezione 3.2 e alla discussione delle teorie scientifiche sulla coscienza, analizzando le ben note Recurrent Processing Theory, Global Workspace Theory e Higher Order Theories, per finire con approcci più innovativi di tipo predittivo come la teoria di Seth (Seth 2023).
Il Capitolo 4 si concentra poi sulla definizione e possibilità della coscienza artificiale. Vengono esaminati alcuni dei criteri necessari in letteratura(Sez. 4.2) per attribuire coscienza a una macchina, come la consapevolezza, la memoria, l'apprendimento, le esperienze soggettive e vengono proposte anche le abitudini (Noë 2010). In particolare, viene posta attenzione tra i processi consci e inconsci di memoria e apprendimento.
Nel Capitolo 5, infine, vengono presentati e discussi vari casi di studio, che illustrano come diverse teorie sulla coscienza possano essere implementate in sistemi di IA attuali. Discutendo tali approcci viene evidenziato come, nonostante la suddivisione stati consci e inconsci sia cruciale per una coscienza artificiale, spesso per questi ultimi non venga data una formalizzazione né un'interazione con quelli consci. Questo studio non solo intende approfondire la comprensione dei concetti di intelligenza e coscienza, ma anche stimolare un dibattito più ampio su cosa significhi realmente per una macchina essere intelligente e consapevole, sfidando il nostro antropocentrismo e aprendo nuove prospettive su come percepiamo le entità artificiali nella nostra società. Si vedrà nelle conclusioni che per una coscienza artificiale è necessario anche sviluppare i processi inconsci e che forse piuttosto che discutere di coscienza artificiale, sarebbe più idoneo trattare de *la coscienza che una macchina può avere*, distaccandosi dall'antropocentrismo che la pretende uguale alla nostra.

2 Cos'è l'intelligenza

Definire l'*intelligenza* è indubbiamente arduo. Per una ottima comprensione e definizione di tale concetto numerose discipline concorrono nel caratterizzarla come la psicologia, la biologia, le neuroscienze e pure la filosofia. Può essere utile innanzitutto distinguere l'intelligenza umana e l'intelligenza *non* umana. L'intelligenza umana può essere definita come "il potere intellettuale degli esseri umani, caratterizzato da complesse sfide cognitive e alti livelli di motivazione e autoconsapevolezza" (Tirri 2011). Essa per alcuni autori come Gardner sarebbe *multipla* e sarebbero presenti quella logico-matematica, linguistica, spaziale, musicale, cinestesica, interpersonale, intrapersonale, naturalistica ed esistenziale/teoretica. (Gardner 1999) Per quanto riguarda l'intelligenza *non* umana, gli scienziati hanno anche tentato di indagare l'intelligenza animale o, più in generale, la *cognizione animale*, gli esperimenti di Wolfgang Köhler sull'intelligenza delle scimmie sono un esempio di ricerca in quest'area, così come lo è il libro di Stanley Coren, *The Intelligence of Dogs* (Coren 1995). Nonostante ciò, la ricerca sul campo dell'intelligenza è stata fino agli anni 2000 concentrata in quella umana, con numerose teorie, per altro controverse, come nel libro *The Bell Curve: Intelligence and Class Structure in American Life* (Herrnstein e Murray 1994) del 1994 in cui gli autori sostengono che l'intelligenza umana è sostanzialmente influenzata sia da fattori ereditari che ambientali e che sarebbe un ottimo predittore di molti risultati personali e lavorativi, inclusi reddito, status socioeconomico e probabilità di essere coinvolti in crimini. L'accusa nei confronti del libro è che gli autori avrebbero affermato che le differenze nel quoziente di intelligenza (QI) tra gruppi razziali ed etnici sarebbero almeno in parte di origine genetica, una visione che ora è considerata screditata dalla scienza tradizionale (Panofsky et al. 2021). Nel 1994, viene pubblicato all'interno del *Mainstream Science on Intelligence* (Gottfredson 1997) una risposta alla controversia sul libro *The Bell Curve* che, oltre che essere firmata da cinquantadue ricercatori, su

un totale di 131 invitati, descriveva l'intelligenza come: "una capacità mentale molto generale che, tra le altre cose, implica la capacità di ragionare, pianificare, risolvere problemi, pensare in modo astratto, comprendere idee complesse, apprendere rapidamente e imparare dall'esperienza. Non si tratta semplicemente di imparare dai libri, di una ristretta abilità accademica o di intelligenza nel sostenere i test. Piuttosto, riflette una capacità più ampia e profonda di comprendere ciò che ci circonda: "dare un senso" alle cose o "capire" cosa fare." (Gottfredson 1997) Sempre durante gli anni 90 un'altra definizione puramente ristretta al campo umano appartiene a Reuven Feuerstein che con la sua teoria della Structural Cognitive Modifiability descrive l'intelligenza come "la propensione *unica* degli *esseri umani* a cambiare o modificare la struttura del loro funzionamento cognitivo per adattarsi alle mutevoli esigenze di situazioni di vita" (Feuerstein et al. 2002; Feuerstein 1990).

2.1 L'intelligenza cognitivista ed il ruolo del *problem solving*

Il problem solving è un processo mentale mirato a trovare un percorso che porti dal punto di partenza a un risultato desiderato. Questa abilità viene spesso utilizzata come una misura pratica dell'intelligenza, in quanto coinvolge l'uso del pensiero logico, valutato attraverso il quoziente intellettivo, per risolvere problemi specifici. Questo concetto rappresenta l'approccio cognitivista[1] nello studio dell'intelligenza (Newell e Simon 1972). Definire l'intelligenza attraverso il problem solving è stato un passo fondamentale per gli psicologi, permettendo loro di passare da una visione scolastica dell'intelligenza a concetti più diversificati. Esempi di tali concetti includono l'intelligenza fluida e cristallizzata di Raymond Cattell (Cattell 1963), l'intelligenza logica e creativa, e le teorie più recenti delle già citate intelligenze multiple di Gardner (Gardner 1999), come pure dell'intelligenza emotiva di Daniel Goleman (Goleman 1996).

2.2 L'intelligenza come comportamento *teleologico* di un agente

Secondo alcuni autori, per definire l'*intelligenza* in modo soddisfacente si deve invece espandere tale concetto a tutti gli esseri biologici e *non*. Come fa notare Nello Cristianini infatti, dimostrare intelligenza e/o possederla non significherebbe essere simili agli esseri umani (Cristianini 2023). In altre parole, non bisogna pensare all'intelligenza come legata solo al genere umano e misurata con le nostre strumentazioni (vedasi i test IQ) ma un'insieme di abilità utili a raggiungere un obbiettivo per un certo *agente*[2] in un dato ambiente. Inoltre, nella visione di Cristianini, possiede intelligenza un qualsiasi agente capace di comportarsi in modo efficace in situazioni nuove (Cristianini 2023). Cristianini definisce pertanto l'intelligenza "in termini di comportamento di un agente, ovvero di qualsiasi sistema in grado di agire nel suo ambiente, usando informazioni sensoriali per prendere decisioni" (Cristianini 2023). Questa definizione, dice inoltre Cristianini, nasce nel 1991 quando James Albus definiva appunto l'intelligenza come "la capacità di un sistema di agire in modo appropriato in un ambiente incerto, dove le azioni appropriate sono quelle che aumentano le probabilità di successo" (Cristianini 2023; Albus 1991) una definizione peraltro coerente con quella cognitivista del problem solving (vedi Sez. 2.1). Gli agenti con tali capacità, in grado di massimizzare le proprie probabilità di successo, mostrerebbero un comportamento "teleologico", dalla parola greca *telos*, che Aristotele usò per definire lo scopo ultimo o la direzione spontanea di moto di un'entità (Cristianini 2023). Un comportamento teleologico è quindi inteso a portare a termine un obbiettivo e pertanto qualsiasi agente (anche artificiale e non biologico) dotato di tale abilità sarebbe quindi *intelligente*.

2.3 L'intelligenza artificiale

Se si assume l'intelligenza come non esclusiva umana e generalizzata ad un comportamento teleologico di un qualsiasi agente anche non biologico (per esempio *artificiale*), come discusso precedentemente, la conseguenza naturale è poter quantomeno concepire l'esistenza dell'intelligenza cosiddetta *artificiale*. Questo processo non è banale, come fa notare Cristianini, si tratterebbe di un vero e proprio *passo copernicano* volto ad abbandonare l'antropocentrismo (Cristianini 2023). L'intelligenza artificiale (AI), nel suo senso più ampio, è l'intelligenza esibita dalle macchine, in particolare dai sistemi informatici (Russel e Norvig 2021). Fu Alan Turing la prima persona a condurre ricerche nel campo che chiamò appunto *intelligenza artificiale* (Copeland 2004). Tuttavia,

[1] La psicologia cognitiva, anche detta cognitivismo, è una branca della psicologia applicata allo studio dei processi cognitivi, teorizzata a partire dagli anni 1960, che ha come obiettivo lo studio dei processi mentali mediante i quali le informazioni vengono acquisite dal sistema cognitivo, elaborate, memorizzate e recuperate. Rimpiazzò l'orientamento teorico allora prevalente, il *comportamentismo*, che invece teorizzava la non indagabilità dei processi mentali, e l'associazione diretta tra stimolo e risposta.

[2] Si noti che con questo termine l'autore intende anche una forma non biologica

la nascita dell'intelligenza artificiale come disciplina accademica risale al 1956 (Russel e Norvig 2021). Negli ultimi anni, grazie all'enorme ricerca sul campo e la nascita del Deep Learning si è cominciato a distinguere tra la cosiddetta intelligenza artificiale debole e quella forte.

2.3.1 L'intelligenza artificiale forte e debole

La ricerca sull'intelligenza artificiale forte (IA forte o strong AI) ha come obiettivo la creazione di una IA capace di replicare completamente l'intelligenza umana, e viene sovente chiamata Intelligenza Artificiale Generale ("Artificial General Intelligence", AGI) per distinguerla da progetti di IA comuni e meno ambiziosi. Sebbene in origine l'AGI fosse proprio l'obiettivo originale dei ricercatori nel campo dell'IA, le controversie sulla non emulabilità del cervello umano da parte di una macchina, avrebbe ridotto di molto le aspettative in merito. Tuttavia, le ricerche continuano e si osserva che, se si riuscisse a replicare i nostri processi mentali, si potrebbe creare una macchina senziente. Questa macchina sarebbe cosciente, almeno in una misura paragonabile alla nostra (Goertzel 2014). Ciò solleverebbe complesse sfide etiche, poiché dovremmo confrontarci con questioni legate ai diritti e al trattamento di tali macchine, alla responsabilità delle loro azioni e all'impatto sociale di avere entità senzienti artificiali nella nostra società (Goertzel 2014). In contrasto con l'intelligenza artificiale forte, l'intelligenza artificiale debole si riferisce all'uso generico dell'IA per problem-solving. Un esempio di programma di intelligenza artificiale debole è un algoritmo per il gioco a GO. Diversamente dall'intelligenza artificiale forte, quella debole non realizza un'auto-consapevolezza, ma è esclusivamente un agente intelligente che può tuttavia arrivare ad avere una conoscenza irraggiungibile da un essere umano. Si pensi per esempio ai modelli come ChatGPT che riescono a *leggere* e *rispondere*[3] anche testi lunghi in molteplici lingue in tempo rapidissimo, circostanza impossibile per un solo umano.

3 Cos'è la coscienza

E' oggettivamente complesso definire in modo esauriente la coscienza. Tuttavia, è utile innanzitutto comprendere cosa significhi averla e, dal processo inverso, ricavare quantomeno alcune definizioni più rilevanti nella storia della scienza della coscienza. Secondo il neuroscienziato Anil Seth, essere una creatura cosciente significherebbe che "c'è qualcosa che si *prova* a essere tale creatura" (Seth 2023). Questa definizione della coscienza come uno stato esperenziale dell'essere trae le sue radici da un articolo del filosofo Thomas Nagel "What Is It Like to Be a Bat?" il quale sostiene che mentre un essere umano potrebbe essere in grado di immaginare cosa significhi essere un pipistrello prendendo "il punto di vista del pipistrello", sarebbe comunque impossibile "sapere cosa significhi per un pipistrello essere un pipistrello" (Nagel 1974). A tal proposito, Nagel sostiene che "un organismo ha stati mentali coscienti se e solo se vi è qualcosa che si prova a *essere* quell'organismo – qualcosa che si prova *per* quell'organismo" (Nagel 1974) e questa asserzione ha posto le basi e le fondamenta per gli studi successivi sulla coscienza. Un'altra definizione arriva da John Searle che la definisce come un processo *continuo*: "stati soggettivi di consapevolezza o sensibilità che iniziano quando ci si sveglia al mattino e continuano per tutto il periodo in cui si è svegli fino a quando si cade in un sonno senza sogni, in coma, o si muore o si è in altri modi tali da essere inconsci" (Chalmers 2002). Prima di addentrarsi nelle differenti dottrine filosofiche di pensiero sulla coscienza, può essere utile operare la suddivisione del neuroscienziato Anil Seth rispetto alla distinzione tra le proprietà *fenomenologiche* della coscienza e le sue proprietà *funzionali* e comportamentali (Seth 2023). Le prime sono relazionate a ciò di cui un sistema come la coscienza *è fatto*. Pertanto sarebbero legate a come emergano le esperienze soggettive da stati fisici (Seth 2023). Le seconde, invece, sono maggiormente legate a cosa un sistema *fa* ed alla relazione che esiste tra le funzioni che svolge, in altre parole come esso *funzioni* (Seth 2023) (vedasi il *funzionalismo* nella Sezione 3.1.2)

3.1 La coscienza nella filosofia

3.1.1 Il problema *difficile* della coscienza

Oltre al problema di definire in modo esauriente la coscienza, un altro argomento molto controverso in filosofia e filosofia della mente riguarda il descrivere *per quale motivo* siamo soggetti ad esperienze

[3]Si noti che i termini *leggere* e *rispondere* sono qui usati al posto del più adatto *capire*. Questo poichè vi è un lungo dibattito sulla capacità reale dell'IA di *comprendere* e di *capire* realmente il significato, la semantica delle lingue (si veda Sez. 4.1.1: Searle e l'esperimento della stanza cinese) in merito si veda Searle 1990; Searle 1980 ed anche Cole 2023.

coscienti. Il neuroscienziato Anil Seth riassume tale questione come una serie di domande su come la coscienza accada, come le esperienze coscienti siano collegate al funzionamento biofisico del cervello e del nostro corpo, e come spiegare la loro relazione con gli atomi o con qualsiasi altra cosa di cui è fatto l'Universo (Seth 2023). La formulazione classica di tale questione è nota come il *problema difficile* della coscienza ed è stata coniata dal filosofo David Chalmers all'inizio degli anni Novanta (Seth 2023). Sostanzialmente, Chalmers distingue il problema difficile da quello cosiddetto *facile*, in realtà insieme di problemi facili, che invece avrebbero a che fare con la spiegazione di come sistemi fisici, quali il nostro cervello, possano dare origine a un numero qualsiasi di proprietà funzionali e comportamentali (vedesi inizio del capitolo) (Seth 2023; Chalmers 1995). In altre parole, i problemi facili riguardano l'analisi meccanicistica dei processi neurali che accompagnano il comportamento. Esempi di questi includerebbero il funzionamento dei sistemi sensoriali, il modo in cui i dati sensoriali vengono elaborati dal cervello ed il modo in cui tali dati influenzano il comportamento (Chalmers 1996). Il problema difficile, al contrario, è il problema del *perché* e del *come* questi processi siano accompagnati dall'esperienza (Chalmers 1995). Nelle prossime sezioni vengono esposte e introdotte le principali correnti di pensiero filosofiche che mirano sia a definire la coscienza, sia ad affrontare il problema difficile.

3.1.2 La coscienza come fenomeno fisico: il materialismo e il funzionalismo

Le definizioni a inizio capitolo di Nagel e Searle sull "essere coscienti" e "avere coscienza" assumono che l'Universo sia composto da entità fisiche e che gli stati di coscienza siano, oppure emergano da, particolari composizioni di tali entità [4] e rientrano in uno dei differenti quadri concettuali per pensare alla coscienza: il fisicalismo/materialismo.[5] Tale posizione filosofica sarebbe quella ad oggi assunta dalla maggior parte dei neuroscienziati (Seth 2023). Una prospettiva più computazionale appartiene a filosofi come Daniel Dennet che propone una teoria fisicalista della coscienza basata sul cognitivismo, che vede la mente in termini di elaborazione delle informazioni.[6] In tal senso, per Dennet il cervello sarebbe un'insieme di "agenzie indipendenti" che "revisionano" in continuazione l'informazione in *input* e danno luogo a un'esperienza cosciente. Arriva a dire che: "tutte le varietà di percezione – anzi tutte le varietà di pensiero o di attività mentale – si realizzano nel cervello attraverso processi paralleli e multitraccia di interpretazione ed elaborazione degli input sensoriali" (Dennet 1991). Teorie come quella di Dennet rientrano nella varietà fisicalista del *funzionalismo* o *mente-computer* che vede, appunto, il cervello come un calcolatore e l'esperienza cosciente come il risultato di calcoli, di processi svolti da tale calcolatore. Si noti che questo celerebbe delle forti assunzioni come fa notare Anil Seth, significherebbe infatti che basterebbe assicurarsi che un sistema "abbia l'opportuno tipo di "corrispondenza input-output" e ciò sarà sufficiente per dare origine alla coscienza"; la coscienza sarebbe quindi *simulabile* da un computer (Seth 2023). In altre parole, l'idea del funzionalismo è che la coscienza non dipenda da ciò di cui sia fatto un sistema ovvero le sue proprietà fenomenologiche, ma dalle sue proprietà funzionali, da come trasforma gli *input* in *output* (Seth 2023). D'altro canto, non sorprendentemente, il *funzionalismo* sarebbe la teoria più coerente con la possibilità dell'esistenza di un'AI cosciente. Tuttavia, la teoria materialista non risolverebbe il problema difficile, sostiene infatti Chalmers che se si accettasse che l'esperienza cosciente scaturisca da una base fisica, resterebbe tuttavia da capire *perchè* e *come* essa scaturisca (Chalmers 1995). Nonostante la visione di Chalmers, l'evidenze neuroscientifiche sosterrebbero la visione materialista poiché mostrerebbero come la coscienza e i processi mentali siano correlati con l'attività neurale nel cervello. Diversi studi hanno infatti identificato i *correlati neurali della coscienza* (NCC), ovverosia modelli di attività cerebrale strettamente associati alle esperienze soggettive e, pertanto, coscienti (Friedman et al. 2023).

3.1.3 *Res cogitans* e *res extensa*: il dualismo cartesiano

Una corrente diametralmente opposta al materialismo/fisicalismo è il *dualismo* che affonda le sue radici nella concezione cartesiana secondo cui la coscienza (in questo contesto la mente, *res cogitans*) e corpo (*res extensa*) siano sostanze o modalità di esistenza distinte (Crane e Patterson

[4] In letteratura viene fatto spesso l'esempio della "rossezza" di un fiore. L'esperienza della "rossezza" di un tulipano per esempio: secondo i fisicalisti l'esperienza è un particolare stato cerebrale fisico e la "rossezza" è una caratteristica di quello stato.

[5] Si noti che i due termini possono essere considerati sinonimi, anche se il termine *materialismo* avrebbe origini meno recenti di *fisicalismo* e quest ultimo consisterebbe piuttosto in una "tesi linguistica" dove gli enunciati corrisponderebbero a enunciati fisici, mentre il materialismo sarebbe una tesi più generale sulla natura delle cose (Seth 2023).

[6] L'information Processing Theory è un approccio di comprensione del pensiero umano che considera la cognizione come essenzialmente di natura computazionale, dove la mente è il software e il cervello è l'hardware (Shannon e Weaver 1963).

2000); come esse interagiscano sarebbe affrontato da due sottocorrenti distinte: interazionismo e epifenomenalismo.[7] In questo senso, la "sostanza" *mente* potrebbe tranquillamente esistere senza corpo, in quanto ne è distaccata. [8] Se il materialismo/fisicalismo ha riscontri scientifici per i NCC dalle neuroscienze, il dualismo avrebbe a suo favore un argomento filosofico, molto controverso, chiamato degli *zombie filosofici*. Il paradosso riguarderebbe la capacità di immaginarsi un individuo totalmente uguale a noi nel corpo, nello spazio e nel tempo ma privo di coscienza e mente. La chiave in tutto ciò sarebbe che se è possibile immaginare un tale individuo, difatti uno *zombie* nostro clone, dunque sarebbe possibile separare il corpo dalla mente e la leicità di una tale concezione autorizzerbe a dire che esse siano distinte e divisibili. In altre parole, sostiene Chalmers, se possiamo concepire un essere fisicamente identico a una persona che non ha coscienza (da qui il nome più specifico di *argomentazione della concepibilità*), l'esistenza di quest'ultima non può essere spiegata a partire da fatti relativi alla fisica. È evidente che necessita di spiegazioni ulteriori (Chalmers 1996). Non mancano ovviamente obiezioni dei fisicalisti come Dennet che sostengono che gli zombi filosofici sarebbero logicamente incoerenti, e pertanto impossibili (Dennet 2009).

3.1.4 La coscienza generatrice del mondo fisico: l'idealismo

Un'altra posizione totalmente opposta alle precedenti è l' *idealismo* che sostiene fondamentalmente il contrario: sono gli stati fisici ad emergere dagli stati di coscienza o mentali. La chiave qui è comprendere come la materia emerga dalla mente, non viceversa come nel materialismo (Seth 2023). Sarebbe quindi la mente a costituire l'ultima fonte della realtà e l'esistenza di tutte le cose dipenderebbe da essa. In contrasto con il materialismo inoltre, l'idealismo afferma il primato della coscienza come origine e prerequisito di tutti i fenomeni, rifiutando sia il materialismo che il dualismo. [9].

3.1.5 L'Universo cosciente: il panpsichismo

La corrente del panpsichismo sostiene che la mente sia una caratteristica fondamentale e onnipresente della realtà (Philip et al. 2017). Viene anche descritta come una teoria secondo la quale "la mente è una caratteristica fondamentale del mondo che esiste in tutto l'universo" (Bruntrup e Jaskolla 2017). I panpsichisti presuppongono che il tipo di mentalità che conosciamo attraverso la nostra esperienza è presente, in qualche forma, in un'ampia gamma di corpi naturali (Clarke 2004). Rispetto alle precedenti teorie, David Chalmers descrive il panpsichismo come un'alternativa sia al materialismo che al dualismo (Chalmers 2015a). In particolare, Chalmers afferma che il panpsichismo rispetterebbe le conclusioni sia a favore come l'argomentazione della concepibilità (o degli zombie filosofici, vedi Sezione 3.1.3) sia contro il dualismo (Chalmers 2015a). Tuttavia, uno dei principali argomenti contro il panpsichismo è il problema cosiddetto della *combinazione* (Chalmers 2015b). Tale argomentazione riguarda il fatto che se la coscienza fosse onnipresente, di conseguenza ogni atomo ne avrebbe un livello minimo. Seppur accettassimo questa visione, resterebbe comunque da spiegare come queste "minuscole coscienze si combinino", ad esempio per creare esperienze coscienti più ampie come "la fitta di dolore" che si sente al ginocchio (Keith 2016). Nonostante questo, la *teoria integrata dell'informazione della coscienza* in inglese Integrated Information Theory of consciousness(IIT)[10], proposta dal neuroscienziato e psichiatra Giulio Tononi nel 2004 (Tononi 2004), si ispira al panpsichismo e postula che la coscienza sarebbe diffusa e possa trovarsi anche in alcuni sistemi semplici (Tononi e Koch 2015); non mancherebbero comunque controversie su tale teoria come l'accusa di essere pseudoscienza non falsificabile (Lenharo 2023).

3.2 La coscienza nella scienza

In questa sezione, vengono elencate le teorie scientifiche sulla coscienza. In particolare, ispirandosi al lavoro di Butlin et al (Butlin et al. 2023), viene discussa anche un'altra teoria scientifica recente, la teoria dell'*allucinazione controllata* del neuroscienziato Anil Seth (Seth 2023).

[7] L'interazionismo sostiene che il mentale e il fisico sono fondamentalmente distinti ma interagiscono in entrambe le direzioni: gli stati fisici influenzano gli stati mentali e viceversa. L'epifenomenalismo sostiene che gli stati fisici influenzano gli stati mentali, ma nega che gli stati mentali influenzino gli stati fisici (Chalmers 2002).

[8] Secondo questa visione, l'esperienza della "rossezza" di un tulipano sarebbe ancora uno stato cerebrale, tuttavia totalmente distinto da qualsiasi proprietà fisica di tale stato.

[9] Wikipedia, voce "Idealism", https://en.wikipedia.org/wiki/Idealism

[10] La teoria dell'informazione integrata (IIT) propone un modello matematico per la coscienza di un sistema. Comprende un quadro in ultimo destinato a spiegare perché alcuni sistemi fisici (come il cervello umano) siano coscienti

3.2.1 Recurrent Processing Theory (RPT)

La Recurrent Processing Theory (RPT), traducibile come Teoria dell'elaborazione ricorrente, è una delle teorie che Butlin et al esplorano e che risulta essere di spicco nel gruppo di teorie neuroscientifiche sull'elaborazione nelle aree percettive del cervello. Per tale motivo, ci si riferisce a tali teorie come teorie *locali* della coscienza, contrapposte a quelle che sono invece *globali*, le prime infatti affermano che attività cerebrali in zone circoscritte siano sufficienti per esperienze coscienti, mentre le seconde sostengono che ci debba essere un'attivazione più generale. Nello specifico caso di RPT, essa assume che l'esperienza (visiva) cosciente non richieda il coinvolgimento di altre aree non deputate alla vista. La RPT è principalmente legata all'elaborazione visiva ed afferma che gli stati inconsci e consci corrispondono a fasi distinte di tale processo. In particolare, RPT ha l'intento principale di spiegare cosa distingue gli stati nei quali gli stimoli visivi sono percepiti consapevolmente a quelli in cui sono invece rappresentati inconsciamente. Un semplice stimolo visivo non sarebbe sufficiente infatti per un'esperienza cosciente, invece necessita di essere forte o saliente a sufficienza, tale da provocare l'innesco di un flusso di informazione ricorrente, in cui i segnali vengono rimandati dalle aree superiori della gerarchia visiva a quelle inferiori. Se ciò avviene, questa elaborazione ricorrente genera una rappresentazione consapevole. Da questo punto di vista, vi sarebbe una distinzione cruciale tra operazioni visive consce e incosce: l'estrazione di caratteristiche da una scena visiva sarebbe ancora inconscia, mentre la sua percezione effettiva (come il distinguere figura e sfondo) richiederebbe una visione cosciente (Butlin et al. 2023). Tuttavia, fa notare Butlin et al che vi sono due aspetti critici della RPT (Butlin et al. 2023). Uno, il primo, è che alcuni ritengono che l'elaborazione ricorrente sia un fenomeno biologico legato a specifici neurotrasmettitori e meccanismi neurali (Lamme 2010). Infatti, come osserva Butlin, da questa prospettiva non sarebbe possibile realizzare una coscienza artificiale, poiché l'elaborazione ricorrente verrebbe vincolata a processi biologici particolari (Butlin et al. 2023). Il secondo aspetto è che, come già introdotto, la RPT può essere concepita solo come una teoria della coscienza visiva, senza affrontare ciò che è necessario o sufficiente per la coscienza in generale. Questa interpretazione lascia aperte due questioni: se le esperienze coscienti non visive richiedano processi simili a quelle visive e se debbano essere soddisfatte ulteriori condizioni per la coscienza visiva. Infine, RPT non è stata estesa oltre la visione e l'assunzione che l'attività nelle aree cerebrali visive sia sufficiente per le esperienze coscienti, non sarebbe esente da dubbi (Butlin et al. 2023; Block 2007).

3.2.2 Global Workspace Theory (GWT)

La Global Workspace Theory (GWT), traducibile come "spazio di lavoro globale" si basa sull'idea che gli esseri umani e altri animali si servano di sistemi specializzati, spesso denominati *moduli*, per eseguire vari compiti cognitivi. Tali moduli sono integrati attraverso un accesso comune a uno "spazio di lavoro globale", un'ulteriore "area" nel sistema in cui le informazioni possono essere rappresentate e condivise. Da questo punto di vista, essa si differenzia dalle teorie cosiddette locali come la RPT precedentemente affrontata, che invece sostengono il coinvolgimento di limitate aree cerebrali dare luogo all'esperienza cosciente. Secondo la GWT, un contenuto mentale—che può includere percezioni, pensieri, emozioni, ecc.—diventerebbe consapevole quando accede a questo "spazio di lavoro globale" che è distribuito tra le regioni frontali e parietali della corteccia cerebrale. Pertanto, diventiamo coscienti di un contenuto mentale quando esso è diffuso all'interno di questo spazio di lavoro corticale; in questo modo, il contenuto può essere utilizzato per guidare il comportamento e prendere decisioni. In altre parole, la teoria dello spazio di lavoro globale suggerisce che la coscienza emerga dalla capacità del cervello di integrare informazioni da vari moduli specializzati all'interno di uno spazio comune, che consente l'accesso e la manipolazione di tali informazioni per guidare il comportamento consapevole (Butlin et al. 2023; Seth 2023)

3.2.3 Higher Order Theories (HOT)

Le teorie di ordine superiore (Higher Order Theories, HOT) si distinguono dalle altre per l'accento posto sull'idea che, affinché uno stato mentale sia cosciente, il soggetto deve essere consapevole di essere in quello stato mentale. Questo viene spiegato attraverso il concetto di *rappresentazione di ordine superiore* che rappresenta qualcosa riguardo ad altre rappresentazioni, mentre le rappresentazioni di *primo ordine* rappresentano qualcosa riguardo al mondo (non rappresentazionale). Ad esempio, una rappresentazione visiva di una poltrona rossa è uno stato mentale di primo ordine, mentre la credenza di avere una rappresentazione di una poltrona rossa è uno stato mentale di ordine superiore. Le teorie di ordine superiore sono sostenute da tempo dai filosofi e una delle principali motivazioni di questa visione è il cosiddetto "argomento semplice": se uno stato mentale è cosciente, il soggetto è consapevole di essere in quello stato; essere consapevoli di qualcosa implica

rappresentarlo; quindi, la coscienza richiede una rappresentazione di ordine superiore dei propri stati mentali. In altre, parole la teoria sostiene che un contenuto mentale diventi cosciente quando un processo cognitivo di "livello superiore" è orientato verso di esso, rendendolo cosciente. Secondo questa teoria, la coscienza è strettamente legata ai processi metacognitivi, che implicano una "cognizione sulla cognizione". Come la GWT, infine, anche HOT sostiene che le regioni cerebrali frontali abbiano un ruolo chiave per la coscienza (Seth 2023; Butlin et al. 2023).

3.2.4 Predictive Processing Theories (PPT)

Le Predictive Processing Theories (PP) rappresentano un framework teorico che sostiene che l'essenza della cognizione risiede nella minimizzazione degli errori nelle previsioni della stimolazione sensoriale (Butlin et al. 2023). All'interno di questo contesto teorico si colloca la teoria dell'"allucinazione controllata" (Seth 2023) di Anil Seth. Seth propone un approccio innovativo alla coscienza, suggerendo di liberarsi dai vincoli imposti dal cosiddetto "problema difficile" (vedi Sez. 3.1.1). Seth sostiene che insistere su questo problema sia controproducente e possa scoraggiare lo studio della coscienza. Al contrario, egli definisce ciò che ritiene debba essere l'obiettivo di una buona scienza della coscienza: il "vero problema" (Seth 2023). Secondo Seth, il vero problema consiste nell'affrontare lo studio della coscienza attraverso tre aspetti fondamentali: spiegare, predire e controllare le proprietà fenomenologiche della coscienza. Egli critica teorie come la GWT, HOT così come altre simili, perché si concentrano principalmente sulle proprietà funzionali della coscienza. Seth argomenta che queste teorie non spiegano perché una particolare configurazione dell'attività cerebrale – o altri processi fisici – si traducano in un particolare tipo di esperienza cosciente; si limitano a stabilire che ciò avviene senza fornire una vera comprensione del fenomeno (Seth 2023). Tuttavia, la sua teoria ha ancora una matrice fisicalista e sostiene che gli stati di coscienza emergano da stati fisici e cerebrali. In particolare, Seth vede il cervello come una macchina *predittiva* ed afferma che le nostre percezioni sono il risultato della "migliore ipotesi" formulata dal cervello (Seth 2023). In altre parole, noi percepiamo il mondo non come esso *è* realmente, ma come il cervello crede che sia, ovvero come ce lo fa *sembrare*. Secondo Seth, non esisterebbe la "rossezza" di una poltrona nell'Universo fisico. Non esisterebbe proprio il concetto di "colore", essendo solo uno strumento utile che l'evoluzione ha approntato affinché il cervello possa riconoscere e tenere traccia degli oggetti nelle diverse condizioni di illuminazione. Il colore è solo un'*allucinazione controllata*, una fantasia neuronale creata dal nostro cervello (Seth 2023). Ciò non implica che la poltrona non esista come oggetto fisico; essa è effettivamente un corpo fisico con le sue qualità primarie (occupa uno spazio, è solido, può muoversi). Tuttavia, il modo in cui percepiamo le sue qualità secondarie, come il colore – che dipendono dall'osservatore e derivano dalle onde luminose – è un'allucinazione generata dal cervello Seth 2023. Questa prospettiva implica che la nostra esperienza del mondo è sempre mediata da interpretazioni cerebrali. L'esperienza della "rossezza" di una poltrona, quindi, non deriva da una proprietà intrinseca dell'oggetto, ma è una costruzione della nostra mente basata su processi neuronali predittivi. Di conseguenza, dato che le nostre percezioni non corrispondono necessariamente a cose che hanno un'esistenza indipendente dalla mente, il problema difficile appare meno complesso, fino a diventare addirittura semplice (Seth 2023). Non c'è più la necessità di spiegare perché il colore rosso della poltrona provoca l'esperienza della "rossezza" e come siano legati, poiché l'esperienza stessa è un'allucinazione controllata del cervello. È importante notare che con il termine "allucinazione" non si intende la visione o la percezione di qualcosa che non esiste. Piuttosto, si tratta, appunto, di un'allucinazione *controllata*, ovvero guidata dalla realtà del mondo fisico, a differenza delle allucinazioni vere e proprie, causate ad esempio da malattie psichiatriche. Seth riassume così questo concetto: "alluciniamo sempre. È solo quando siamo d'accordo sulle nostre allucinazioni che parliamo di realtà" (Seth 2023).

4 Coscienza artificiale

Una volta esposti i concetti di intelligenza, intelligenza artificiale e coscienza e le loro differenti definizioni e proprietà, è possibile addentrarsi nel campo controverso della *Artificial Consciousness* (AC) (Thaler 1998). La coscienza artificiale, nota anche come coscienza della macchina (Machine Consciousness, MC) (Reggia 2013), o coscienza digitale (Elvidge 2018), è la coscienza che si ipotizza sia possibile nell'intelligenza artificiale (Ron 2008). Tale campo di studio incrocia e interseca la filosofia della mente, la filosofia dell'intelligenza artificiale, le scienze cognitive e le neuroscienze. Normalmente, si assume che se è possibile l'Intelligenza Artificiale Generale (cioè un'IA forte) (vedi Sez. 2.3.1) dunque essa sarebbe anche dotata di coscienza Goertzel 2014. Come già visto

nel Cap. 3, sez. 3.1.2, alcuni studiosi di base materialisti, ritengono che la coscienza sia generata dall'interazione di varie parti del cervello; questi meccanismi sono anche provati dalle neuroscienze ed etichettati come NCC. I materialisti che sostengono *anche* il funzionalismo, credono inoltre che la costruzione di un sistema (ad esempio, un sistema informatico) in grado di emulare ciò che avviene nei NCC, si tradurrebbe in un sistema totalmente consapevole (Graziano 2013).

4.1 Coscienza artificiale e filosofia

4.1.1 Plausibilità della coscienza artificiale ed il problema dell'*intenzionalità*

Nel suo articolo "Coscienza artificiale: utopia o possibilità reale", Giorgio Buttazzo afferma che un'obiezione comune alla coscienza artificiale è che i computer non possono mostrare creatività, emozioni o libero arbitrio. Lo scienziato afferma infatti che che: "un computer, come una lavatrice, è uno schiavo gestito dai suoi componenti" (Buttazzo 2001). Dell'opinione opposta è Chalmers, da cui per altro viene uno degli argomenti più espliciti a favore della plausibilità della coscienza artificiale, basato sul *computazionalismo* (Hauser 2001) [11]. La sua proposta è che i giusti tipi di calcoli siano sufficienti per possedere una mente cosciente e che qualsiasi sistema che implementa determinati calcoli sia perciò senziente (Chalmers 2011). Tuttavia, l'argomento pilastro contro la possibilità di una coscienza artificiale riguarda il fatto che una macchina non possa mai davvero possedere *intenzionalità*. Con intenzionalità, in questo contesto della coscienza o della mente, si intende l'idea che la coscienza sia sempre diretta ad un oggetto, che abbia sempre un contenuto. Franz Brentano nella sua opera *Psychologie vom Empirischen Standpunkte* (Psicologia dal punto di vista empirico) definisce l'intenzionalità come la caratteristica principale dei fenomeni psichici (o mentali), tramite cui essi possono essere distinti dai fenomeni fisici. Ogni fenomeno mentale, ogni atto psicologico ha un contenuto, è diretto a qualche cosa (l'oggetto intenzionale). In altre parole, ogni atto di *credere*, *desiderare* ha un oggetto: il *creduto*, il *desiderato*. Per Searle è proprio l'intenzionalità la componente principale della mente umana ed è strettamente legata all'evento di coscienza (Cole 2023; Searle 1980). La sua argomentazione si basa sull'esperimento mentale della Stanza Cinese. In questo esperimento, Searle, madrelingua inglese, si immagina all'interno di una stanza con delle istruzioni che traducono i simboli dall'inglese al cinese. Queste istruzioni spiegano dettagliatamente come combinare i simboli per formare frasi, domande e altre costruzioni linguistiche. All'esterno della stanza, Searle immagina un madrelingua cinese con cui comunicare. Questa comunicazione diventerebbe progressivamente più complessa man mano che le istruzioni si allungano e si complicano (Cole 2023; Searle 1980) e dal punto di vista esterno, potrebbe sembrare che Searle sia diventato fluente in cinese. Tuttavia, la sua conoscenza si limiterebbe alle istruzioni fornite (il *programma*), riguardanti solo la dimensione *sintattica* senza alcuna comprensione della lingua cinese cioè della *semantica* (Cole 2023; Searle 1980). In altre parole, in ultimo Searle conoscerebbe solo la corrispondenza *input-output* tra i simboli inglesi (*in input*) e quelli cinesi (in *output*), senza avere una reale *comprensione* del significato dei simboli cinesi (Cole 2023 e Searle 1980). In questo modo Searle rivolge una critica al funzionalismo che pretende di poter astrarre l'intelligenza dal suo sostrato biologico e che la coscienza sia simulabile da un computer. Searle critica anche il test di Turing, ritenendolo inadatto a stabilire se veramente una macchina sia in grado di *pensare* (Hauser 2001). Tuttavia, non rifiuta direttamente il materialismo, mantenendo un nesso imprescindibile tra mente e corpo, proponendo la soluzione del *naturalismo biologico*, ovverosia l'idea che le proprietà biologico-chimiche del cervello producano gli eventi mentali (Hauser 2001). Il suo rifiuto è limitato al funzionalismo *mente-computer* e alla concezione di una coscienza artificiale. Per Searle infatti sarebbe impossibile una coscienza artificiale/un'IA forte, a meno che si riproduca totalmente il tessuto biologico del nostro cervello ad un livello tale da permettere di far emergere un'intelligenza dotata d'intenzionalità (Searle 1990).

4.1.2 Test della coscienza artificiale

Sebbene il test di Turing possa essere valido per determinare se una macchina sia intelligente o meno, esso presenterebbe diversi problemi se esteso alla coscienza (Hauser 2001; Searle 1990). Nel 2014, Victor Argonov ha suggerito un test non di Turing per determinare la coscienza delle macchine basandosi sulla capacità di produrre giudizi filosofici (Argonov 2014). Lo scienziato sostiene che una macchina debba essere considerata cosciente se è in grado di produrre giudizi su tutte le proprietà problematiche della coscienza non avendo alcuna conoscenza filosofica innata su questi temi e nessuna discussione filosofica durante l'apprendimento (Argonov 2014). Sebbene

[11] Secondo il computazionalismo i processi mentali consistono in processi di calcolo che operano su simboli e tali processi sono equivalenti a quelli effettuati da un computer (Hauser 2001).

questo test possa essere utilizzato per *rilevare la coscienza*, non può tuttavia confutarne l'esistenza. In altre parole, un risultato positivo dimostrerebbe che la macchina è cosciente, ma un risultato negativo non proverebbe nulla. Ad esempio, l'assenza di giudizi filosofici potrebbe essere causata dalla mancanza di intelletto della macchina, non dall'assenza di coscienza. La ricerca di Butlin et al (Butlin et al. 2023) invece rappresenta un passo in avanti in questo senso poichè teorizza differenti indicatori/test di coscienza artificiale che se soddisfatti aumentano la probabilità che un sistema possa essere senziente. Tali indicatori si basano sulle teorie più famose della coscienza, alcune discusse nella Sezione 3.2, e verranno analizzati nel Capitolo 5 dedicato ai Casi di studio.

4.2 Aspetti necessari per una coscienza artificiale

Prima di esporre i tentativi condotti per implementare coscienza artificiale, in letteratura si trovano vari aspetti ritenuti più o meno necessari per rendere possibile la realizzazione di una coscienza artificiale, di seguito vengono analizzati quelli che potrebbero essere i principali.

4.2.1 Consapevolezza

La consapevolezza (*awareness*) è presente nei 12 principi che Aleksander teorizza come fondamentali per realizzare una AC, egli la chiama *The Awareness of Self* (consapevolezza di sè stessi) (Aleksander 1995). Inoltre, secondo Della Santina et al, almeno nella robotica (un campo dell'intelligenza artificiale), la consapevolezza sarebbe già diffusa (Della Santina et al. 2024). Questa conclusione viene tratta dalla definizione di "awareness" dal dizionario inglese, ovverosia "la conoscenza che qualcosa esista, o la comprensione di una situazione o di un argomento in un dato momento sulla base di informazioni o esperienze" (liberamente tradotto dall'inglese) (Della Santina et al. 2024). Un drone che si crea una rappresentazione interna del suo ambiente, dicono gli autori, sarebbe coerente con questa definizione. Tuttavia, fanno notare che il concetto di "consapevolezza" contiene molte altre sfumature che, evidentemente, la definizione piuttosto riduttiva del dizionario non è in grado di cogliere (Della Santina et al. 2024). A supporto di questo, si noti che in letteratura awareness e consciousness sono spesso considerati sinonimi (Hussain et al. 2009) e come già visto in precedenza la coscienza è qualcosa di arduo da definire.

4.2.2 Memoria

Stando ad alcuni autori (Dainton 2000; Phillips 2018), le nostre esperienze coscienti sembrano essere profondamente radicate nel tempo ed influenzate dai nostri ricordi (Butlin et al. 2023). Già Tulving, negli anni Novanta, sosteneva che gli eventi coscienti interagirebbero con sistemi di memoria diversi rispetto il tipo di contenuto immagazzinato. In particolare, Tulving identitifca tre tipi distinti di memorie: procedurale, semantica ed episodica (Tulving 1972; Tulving 1985). La memoria procedurale riguarda il modo in cui vengono eseguiti determinati compiti e azioni. Quella semantica immagazzina fatti, concetti, nomi e conoscenza generale simbolicamente rappresentabile. Infine, la memoria episodica giocherebbe un ruolo importante nel ricordare eventi vissuti in prima persona (Tulving 1972). Tuttavia, è controverso il ruolo della memoria nella coscienza. Butlin et al, dando per scontato il funzionalismo (vedi Sez. 3.1.2) e computazionalismo (vedi Sez. 4.1.1), fanno ad esempio notare che, nel caso particolare della memoria episodica, essere in grado ricordare un dato legato alla tua infanzia, come l'indirizzo della casa dove sei cresciuto, il più delle volte non avrebbe impatto sulla coscienza (Butlin et al. 2023). Gli psicologi cognitivi considerano inoltre possibile avere memoria senza esserne consapevoli, si parlerebbe della cosiddetta *memoria implicita* inconscia contrapposta a quella *esplicita* e conscia anche detta *declarative memory* (memoria dichiarativa) (Schachter 1997; Ullman 2004). Seguendo questo ragionamento, sorgerebbe un altro problema ovvero come dotare una macchina anche di questo tipo di memoria incoscia. La questione viene discussa da Piletsky che sostiene che proprio lo sviluppo dell'inconscio della macchina ci porterà infine a cambiamenti radicali nella filosofia della coscienza e nella risoluzione del problema difficile (vedi Sez. 3.1.1) (Piletsky 2019).

4.2.3 Apprendimento

Per Bernard Baars e Cleeremans, l'apprendimento e l'adattamento a nuove situazioni dipendono e son collegate all'esperienza cosciente (Baars 1995; Cleeremans e French 2002). Cleeremans in particolare definisce l'apprendimento (learning) come un insieme di processi avanzati che dipendono dalla sensibilità evoluta grazie all'esperienza soggettiva (Cleeremans e French 2002). Un aspetto fondamentale è l'apprendimento dichiarativo, o esplicito, che è rivolto a immagazzinare tutte le informazioni di cui si può parlare, in contrasto con l'apprendimento implicito, che è non

consapevole. Questo tipo di apprendimento è direttamente collegato alla coscienza poiché i ricordi dichiarativi si formano sulla base di ciò di cui siamo coscienti (Butlin et al. 2023). L'apprendimento implicito, invece, si differenzia dall'apprendimento esplicito proprio per l'assenza di consapevolezza su ciò che viene appreso (Sun 2008), come nel caso di imparare ad andare in bicicletta. Seguendo questo argomento, Aksyuk propone un quadro teorico di studio della coscienza totalmente basato sull'apprendimento dichiarativo come misura di coscienza, riconoscendolo inoltre come causa determinante dell'esperienza soggettiva, in quanto non accessibile ai processi o ai sistemi inconsci (Aksyuk 2023). Si spinge a dire che se l'apprendimento dichiarativo è presente in un dato sistema, quest'ultimo sarà pertanto cosciente (Aksyuk 2023). Tuttavia, sebbene in generale l'apprendimento sia stato emulato (in modo incompleto), si pensi infatti all'apprendimento automatico, (Machine Learning) essi presentano comunque differenze critiche tra di loro; una macchina deve essere esposta a molti più esempi di noi per apprendere associazioni input output relativamente semplici.

4.2.4 Esperienze soggettive e *qualia*

Nella filosofia della mente, i qualia sono definiti come esempi di esperienza soggettiva e cosciente. In particolare, possono essere descritti in termini di cosa sentiamo e di cosa facciamo esperienza, come una fitta di dolore o l'esperienza del colore rosso (Searle 2005). Non è tuttavia chiaro il ruolo dei qualia nella coscienza, Chalmers assume la loro esistenza e come conseguenza anche l'argomento degli zombie filosofici (vedi Sez. 3.1.3), il problema difficile (vedi Sez. 3.1.1) e la spiegazione insufficiente del materialismo (Chalmers 1995) . Dennet d'altro canto, sostiene che il concetto di qualia è confuso e non analizzabile senza risultare in contraddizioni, pertanto non costituirebbe un valido argomento contro il fisicalismo (Dennet 1991). Tuttavia, per le neuroscienze i qualia esisterebbero e alcuni filosofi intenti a volerli eliminare, interpreterebbero erroneamente ciò che costituisce scienza (Koch 2020). Recentemente, una ricerca fa notare che esistono effettivamente eccessive definizioni e punti di vista diversi sui qualia, rendendoli concetti poco chiari e costruttivi (come per altro Dennet sosteneva). Per tale motivo, gli autori ridefiniscono i qualia in modo più generale e sostengono che questo porterebbe a una serie di vantaggi nell'affrontare il problema mente-corpo, il problema difficile ed il problema dell'intenzionalità di Searle [12]. Infine, la proposta finale degli autori è fortemente ambiziosa: i robot con i qualia, potrebbero essere coscienti (Haikonen 2022).

4.2.5 Abitudini

Per abitudine (inglese *habit*) si intende una routine di comportamento che si ripete regolarmente e che tende a manifestarsi a livello subconscio (Butler et al. 2018). Questo concetto è peraltro strettamente collegato all'apprendimento e alla memoria procedurale precedentemente discussi. In particolare, l'apprendimento associativo alla base delle abitudini è caratterizzato da un accumulo incrementale di informazioni nel tempo nella memoria procedurale (Wood e Neal 2007). Secondo il filosofo Alva Noë, collega di Searle, la letteratura spesso trascura l'importanza delle abitudini, nonostante queste definiscano al meglio la nostra natura come esseri viventi in relazione *con* e *per* il nostro ambiente (Noë 2010). Il suo libro intitolato "Perchè non siamo il nostro cervello", è a partire dal titolo una critica alle teorie delle neuroscienze (come quella dell'allucinazione controllata di Seth, vedi Sez. 3.2.4) che ritengono il cervello come creatore della coscienza e che vede noi esseri umani come "cervelli in una vasca" (Noë 2010). Inoltre, Noë afferma che "quello che serve per avere una mente come la nostra è avere abitudini come le nostre", e sostiene che "solo un essere dotato di abitudini può avere una mente come la nostra" (Noë 2010). Con queste affermazioni, Noë critica le (neuro)scienze cognitive che tendono a vedere l'essere umano come privo di abitudini, e accusa l'intellettualismo di essere riduttivo rispetto alla nostra natura di organismi abitudinari (Noë 2010). L'intellettualismo concepirebbe infatti l'essere umano principalmente come un agente preposto a valutare, decidere e pianificare. Noë ritiene che anche gli scienziati nel campo della robotica dovrebbero abbandonare questa visione intellettualistica. Egli sostiene che essi "hanno speso le loro energie per costruire robot sempre più intelligenti — capaci di giocare a scacchi o di evitare ostacoli. Meglio sarebbe fabbricare robot dotati di abitudini" (Noë 2010). In altre parole, fintanto che seguiamo la visione intellettualistica, secondo la quale l'intelligenza è strettamente collegata alla capacità di giudizio, classificazione, pianificazione e organizzazione, non potremo dotare le macchine di coscienza. Ciò che ci distingue realmente sono le abitudini, peraltro sviluppate a livello inconscio (Noë 2010). Esistono scienziati che si concentrano sul dotare i robot di abitudini, curando l'interazione con il loro ambiente piuttosto che focalizzarsi esclusivamente sulla razionalità (Noë 2010). Per Noë, dunque, la nostra coscienza si sviluppa a partire dall'interazione con

[12] Nel paper originale ci si riferisce al *Symbol Grounding Problem*, qui si adotta una delle definizioni di Harnad che lo assimila al problema dell'intenzionalità della stanza cinese di Searle (Harnad 1990).

l'ambiente (il Mondo) lungo *sentieri cognitivi abitudinari* e non deriva esclusivamente dai processi meccanistici neurali (Noë 2010).

5 Casi di studio

In questo capitolo vengono discusse alcune delle differenti implementazioni di una possibile Coscienza Artificiale. Vengono presentati e discussi vari casi di studio che illustrano come diverse teorie sulla coscienza possano essere implementate in sistemi di IA, evidenziandone i vantaggi ed i limiti. Le implementazioni si riferiscono alle teorie scientifiche sulla coscienza già discusse nel Cap. 3, Sez. 3.2. ed estratte dalla ricerca di Butlin et al (Butlin et al. 2023). Butlin intende fornire un insieme di indicatori, ognuno per una teoria specifica della coscienza, così da facilitare lo studio dell'argomento e la valutazione di un sistema artificiale come consapevole. Butlin sostiene che la probabilità che un sistema sia cosciente aumenta con il numero di indicatori soddisfatti, anche se questi appartengono a teorie diverse (Butlin et al. 2023). In altre parole, se un sistema rispetta numerosi indicatori provenienti da differenti teorie della coscienza, è più probabile che esso possieda una qualche forma di consapevolezza. Questo approccio permette una valutazione più sfumata e completa della coscienza nei sistemi artificiali, che questa ricerca sostiene fortemente. Tuttavia si intende anche proporre l'implementazione CLARION, un'architettura cognitiva sviluppa da Sun ed il suo team (Sun et al. 2001) che rispetto alle teorie standard, definisce e modella una relazione dei processi inconsci (impliciti) con quelli consci (espliciti).

5.1 Implementazioni della RPT e PP

La Recurrent Processing Theory (RPT) sostiene che l'elaborazione dei segnali nel cervello avviene in modo ricorrente, un concetto implementato nelle Recurrent Neural Networks (RNNs). Queste reti neurali artificiali hanno neuroni che ricevono input influenzati dai propri output passati attraverso dei cicli, simulando così l'elaborazione ricorrente proposta dalla RPT, particolarmente nel campo della visione (Butlin et al. 2023). Le Predictive Processing theories (PP), invece, affermano che la coscienza richiede un fenomeno di *inferenza attiva*, dove le nostre percezioni sensoriali emergono dalla "migliore ipotesi" costruita dal cervello. Questo approccio viene definito *elaborazione predittiva*. Molti ricercatori considerano l'elaborazione predittiva una condizione necessaria per la coscienza, motivo per cui Butlin include insieme a RPT la presenza di "moduli di input che utilizzano la codifica predittiva (PP-1)" (Butlin et al. 2023). Recenti studi dimostrano che la codifica predittiva può migliorare l'elaborazione complessiva nella visione artificiale, come evidenziato dal sistema PredNet, che predice il frame successivo di un video in input. Per quanto riguarda RPT, tuttavia, fa notare Butlin che è cruciale stabilire se l'elaborazione ricorrente sia o meno necessaria per la coscienza. In particolare, le reti neurali ricorrenti possiedono una ricorrenza *fisica*, che Butlin chiama *ricorrenza implementativa* per distinguerla da quella cosiddetta *algoritmica* posseduta da reti non ricorrenti (feedforward) con sufficienti strati e pesi condivisi tali da poter imitare un'elaborazione ricorrente (Butlin et al. 2023). Da un lato, se si considera l'aspetto della ricorrenza implementativa come condizione necessaria della coscienza, essa non è realizzabile, almeno al momento, poichè significherebbe creare un sistema artificiale in cui i neuroni sono *fisicamente* realizzati da componenti hardware specifici, un approccio totalmente diverso dagli strumenti di IA odierni che piuttosto *simulano* le reti neurali cerebrali, tralasciando l'hardware sottostante (Butlin et al. 2023). D'altro canto, ignorare la costituzione *fisica* e quindi la ricorrenza implementativa, parrebbe molto di convenienza e pertanto la RPT sarebbe criticata da Doerig et al (Doerig et al. 2019). Per comodità, Butlin considera come indicatore di coscienza artificiale la ricorrenza algoritmica (RPT-1), già presente in molti metodi di IA. Tuttavia, RPT-1 non è l'unico indicatore. Butlin suggerisce anche un secondo indicatore, RPT-2, che prevede la ricorrenza algoritmica (RPT-1) mirata a generare rappresentazioni percettive integrate di scene e oggetti in relazioni spaziali. RPT-2 consiste in "moduli di input che creano rappresentazioni percettive organizzate e integrate" (Butlin et al. 2023). Qui è cruciale la distinzione tra elaborazione inconscia e conscia: mentre l'estrazione di caratteristiche da una scena visiva può avvenire in modo inconscio, operazioni percettive come la separazione figura-sfondo possono richiedere la consapevolezza cosciente. Pertanto, RPT-2 pone l'accento sull'importanza delle rappresentazioni percettive organizzate e integrate. La ricorrenza algoritmica (RPT-1) è infine una caratteristica riscontrata in molte architettetture Deeplearning, Recurrent Neural Networks(RNNs), long short-term memory networks (LSTMs) e gated recurrent unit networks (GRUs) (LeCun et al. 2015). Per la RPT-2 invece sistemi come le Convolutional Neural Networks sono spesso considerati buoni simulatori del sistema visivo, sebbene tale affermazione sia stata recentemente criticata (Butlin et al. 2023).

5.2 Implementazioni della GWT

La Global Workspace Theory (GWT) propone che si utilizzino moduli specifici per eseguire vari compiti cognitivi. Un contenuto mentale, che può includere percezioni, pensieri, emozioni, e altro, diventa consapevole quando accede a uno "spazio di lavoro globale". Questo spazio costituirebbe l' "area" in cui i diversi moduli cognitivi si integrano, permettendo una condivisione e interazione delle informazioni. Butlin teorizza quattro indicatori, sostenendo che maggiori sono quelli rispettati da un sistema artificiale, maggiore è la probabilità che tale sistema implementi correttamente GWT e sia pertanto cosciente secondo tale teoria. Il primo indicatore pone l'accento sul possedere sistemi specializzati (moduli) in grado di svolgere compiti in parallelo. Tale indicatore ha senso poichè per la GWT ciò che conta per la coscienza è il processo che integra i moduli, piuttosto che le loro esatte caratteristiche. GWT-1 è pertanto descritto come "GWT-1: sistemi specializzati multipli in grado di funzionare in parallelo (moduli)" (Butlin et al. 2023). Il secondo è una conseguenza del primo, poichè deve esistere una restrizione sul flusso delle informazioni dove i moduli hanno una maggiore capacità collettiva dello spazio di lavoro globale che alimentano. Inoltre, questo presuppone l'esistenza di un meccanismo di controllo che quindi seleziona le informazioni maggiormente rilevanti da far fluire dai moduli allo spazio globale. Pertanto il secondo indicatore è descritto come "GWT-2: spazio di lavoro con capacità limitata, che comporta un collo di bottiglia nel flusso di informazioni e un meccanismo di attenzione selettiva" (Butlin et al. 2023). Il terzo indicatore racchiude l'essenza dello spazio di lavoro globale, tutti i moduli accedono all'informazione di tale spazio e la condividono, pertanto GWT-3 viene descritto come "Broadcast: disponibilità delle informazioni nello spazio di lavoro per tutti i moduli" (Butlin et al. 2023). Fa notare Butlin che GWT-1 e GWT-2 possono essere soddisfatti da sistemi aventi più moduli di input, che alimentano uno spazio di lavoro limitato, da cui le informazioni fluiscono poi verso uno o più moduli di output. Tuttavia GWT-3 implica che le informazioni debbano rifluire *anche* dallo spazio di lavoro ai moduli di input, influenzandone l'elaborazione. Pertanto ciò significa che i moduli di input debbano rispettare anche ricorrenza algoritmica – fornendo un'ulteriore giustificazione per l'indicatore RPT-1 (Butlin et al. 2023), discusso precedentemente. Per garantire che lo spazio di lavoro globale faciliti interazioni controllate tra i moduli cognitivi, è essenziale che il meccanismo di selezione delle informazioni sia sensibile sia allo stato del sistema sia ai nuovi input. Tale meccanismo permette alle rappresentazioni interne di influenzare la selezione delle informazioni da parte dei moduli e questo requisito porta al quarto indicatore: "GWT-4: Attenzione dipendente dallo stato, che consente di utilizzare lo spazio di lavoro per interrogare i moduli in modo sequenziale al fine di eseguire compiti complessi" (Butlin et al. 2023). Concentrandosi da GWT-1 a GWT-4 e e sui Large Language Models (LLM) basati sull'architettura Transformer come GPT-3 (Brown et al. 2020) e GPT-4 (OpenAI 2023) è possibile sostenere che ciascuno di essi possieda alcune delle proprietà degli indicatori GWT. In un Transformer, un'operazione chiamata "auto-attenzione" viene utilizzata per selezionare le informazioni più rilevanti provenienti da diverse parti di una sequenza di input (Vaswani et al. 2023). I Transformers sono costituti da strati alternati di di teste di attenzione (che eseguono l'autoattenzione) e da strati che ricevono l'output delle teste. Se consideriamo gli elementi che elaborano le informazioni da ciascuna posizione della sequenza di input (teste di attenzione) come moduli, esiste una somiglianza di base tra l'architettura Transformer e lo spazio di lavoro globale: entrambi integrano informazioni da più moduli (Butlin et al. 2023). Da questa prospettiva, i Transformers soddisfano GWT-1 poichè possiedono dei moduli (le teste di attenzione), GWT-2 che implica appunto un meccanismo di controllo (auto-attenzione) per selezionare solo le informazioni rilevanti da far fluire nello spazio limitato, ed anche GWT-3 in quanto esiste una "trasmissione globale" nel senso che l'informazione del sistema può essere utilizzata dalle teste dell'attenzione per influenzare un'ulteriore elaborazione. Da questo punto di vista anche GWT-4 sarebbe rispettato perchè l'informazione in uno strato dipenderebbe dallo stato del sistema nello strato precedente. Tuttavia, un problema è che i Transformer non sono ricorrenti, è controverso che essi rispettino GWT-3 in quanto non vi sono moduli che ricevono indietro delle informazioni (Butlin et al. 2023). In altre parole, la ricorrenza algoritmica (RPT-1) giustificata per altro da GWT-3, non sarebbe soddisfatta.

5.3 Implementazioni della HOT

La Higher Order Theory (HOT) sostiene un modello in cui la coscienza richiede sia rappresentazioni percettive di primo ordine dei dati sensoriali, sia rappresentazioni di ordine superiore che assegnano una misura di "realtà" a queste rappresentazioni di primo ordine. Una delle teorie computazionali basate su HOT è la Perceptual Reality Monitoring theory (PRM) (Lau 2019). La PRM afferma che la coscienza dipende da un meccanismo capace di distinguere l'attività nei sistemi percettivi

dal semplice rumore. Dal punto di vista computazionale, questo processo è simile a quello delle Generative Adversarial Networks (GAN), reti neurali composte da *generatori* e *discriminatori*. Il generatore tenta di produrre dati sintetici e il discriminatore cerca di distinguere i dati prodotti dai dati "reali". Il discriminatore nelle GAN è paragonabile al meccanismo di ordine superiore ipotizzato dalla PRM (Lau 2022). In generale, Butlin costruisce tre indicatori per la coscienza artificiale emergenti dalle teorie HOT. I primi due, HOT-1 e HOT-2, stabiliscono che il modello dovrebbe integrare rappresentazioni percettive di primo ordine e rappresentazioni di ordine superiore mirate a stabilire la "realtà" delle prime, ispirandosi alla PRM (Butlin et al. 2023). Per esempio, la credenza di avere una rappresentazione di una poltrona rossa è di ordine superiore rispetto alla semplice rappresentazione dell'oggetto (primo ordine). Seguendo il ragionamento della PRM, il meccanismo di monitoraggio etichetta alcuni contenuti percettivi come "reali" quando il sistema tende a considerarli tali. Questo implica l'esistenza di un agente che si occupa del monitoraggio e dell'etichettatura. Pertanto, Butlin propone un terzo indicatore, HOT-3, che rappresenta "un'agenzia guidata da un sistema di formazione delle credenze e di selezione delle azioni, con una forte disposizione ad aggiornare le credenze in base ai risultati del monitoraggio" (Butlin et al. 2023). Le teorie computazionali HOT, come la PRM, tentano di rispondere alla domanda "perché gli stati coscienti sono esperiti in quel modo?" e per farlo, utilizzano la Quality Space Theory, che postula che le qualità fenomeniche siano ridotte a discriminanti e somiglianze (Lau 2022). Per esempio, per avere l'esperienza del colore rosso di un tulipano, è necessario comprendere implicitamente la sua somiglianza con il colore di una mela rossa e la sua differenza rispetto al verde di una foglia (Butlin et al. 2023). Questa conoscenza implicita potrebbe dipendere da una codifica sparsa e continua nei sistemi percettivi, piuttosto che in categorie discrete (Lau 2022. Pertanto, il quarto indicatore, HOT-4, richiede un sistema dotato di "codifica sparsa e fluida" che genera uno "spazio di qualità" (Butlin et al. 2023). Questo spazio sarebbe presente in tutte le reti neurali di tipo deep (DNN), che utilizzano spazi di rappresentazione continui e sparsi. Infatti, una delle caratteristiche più importanti di tutti i modelli DNN è che ogni strato possiede uno spazio di rappresentazione sparso e fluido. Anche i metodi standard di machine learning possono essere modificati per realizzare rappresentazioni sparse, ad esempio, tramite tecniche di regolarizzazione applicate ai DNN (Butlin et al. 2023). In sintesi, dalla prospettiva delle HOT e degli indici di coscienza artificiale basati su di esse, implementare una coscienza artificiale sembra relativamente semplice.

5.4 CLARION

CLARION (Connectionist Learning with Adaptive Rule Induction On-line) è una sofisticata architettura cognitiva sviluppata da Ron Sun (Sun et al. 2001) e il suo team, progettata per simulare una vasta gamma di fenomeni psicologici integrando processi cognitivi sia impliciti (consci) che espliciti(inconsci). Questa architettura è suddivisa in quattro principali sottosistemi: azione-centrato, non-azione-centrato, motivazionale e meta-cognitivo (Sun et al. 2001). Il suo punto di forza risiede nella capacità di modellare l'apprendimento e il ragionamento combinando l'uso di reti neurali per i processi bottom-up e sistemi basati su regole per quelli top-down. CLARION si distingue per la sua teoria dei due sistemi rappresentazionali interagenti: la conoscenza esplicita, rappresentata localmente attraverso regole, e la conoscenza implicita, rappresentata distributamente tramite reti neurali (Sun et al. 2001). Nel modello CLARION, la memoria esplicita è dominata da processi di livello superiore, come quelli associati al lobo frontale del cervello, che gestiscono il richiamo delle regole e delle conoscenze esplicite. D'altra parte, la memoria implicita è più influenzata da processi di livello inferiore, riflettendo una gerarchia simile a quella del cervello umano (Sun et al. 2001). Un aspetto cruciale di CLARION è il suo meccanismo di trasformazione dell'apprendimento implicito in esplicito. Questo avviene attraverso un processo bottom-up in cui le regole esplicite vengono estratte dalle rappresentazioni distribuite nelle reti neurali. In altre parole, CLARION permette di cogliere la complessità dei processi cognitivi umani, contribuendo significativamente alla comprensione e alla simulazione dei meccanismi sottostanti i processi sia che consapevoli che non. Infine, CLARION è stato usato anche per spiegare il degrado delle prestazioni dovuto all'ansia, proponendo che questo dipenda da uno sbilanciamento tra il contributo dei processi impliciti ed espliciti (Wilson et al. 2009; Coward e Sun 2004).

5.5 Discussione finale

In questo capitolo sono state discusse varie implementazioni di coscienza artificiale, ispirate al lavoro di Butlin et al (Butlin et al. 2023). e alle teorie esposte nella Sezione 3.2. In particolare, per quanto riguarda tutte queste implementazioni, basandosi sui requisiti per una coscienza artificiale, analizzati nella letteratura (Sezione 4.2), si ritiene che memoria (Sezione 4.2.2) e apprendimento

(Sezione 4.2.3) siano certamente presenti in modo nativo (nonostante per l'apprendimento valga ancora il discorso che non sia identico a quello umano). Tuttavia, è curioso che, nonostante l'importanza attribuita alla distinzione tra processi consci e inconsci, soprattutto nel caso di RPT, non si ponga enfasi sulla formalizzazione di questi ultimi. In questa relazione, infatti, non si sostiene che i processi inconsci siano passivi, ma che siano collegati ai processi consci e che quindi dovrebbero essere implementati o almeno modellati. Da questo punto di vista, la memoria implicita inconscia e l'apprendimento implicito e non dichiarativo non sono adeguatamente considerati. Inoltre, non vi è alcuna traccia di abitudini (Sez. 4.2.5) manifeste. Per quanto riguarda esperienze soggettive, qualia e consapevolezza, il discorso si sposta su un piano più filosofico che scientifico, quindi tali aspetti vengono lasciati da parte sebbene rimangano concetti rilevanti. Tutte le implementazioni qui discusse presuppongono che si diventi coscienti quando certe condizioni sono soddisfatte e implementate, ma non si occupano di definire i processi inconsci e la relazione che intrattengono con quelli consci. Da questa prospettiva, solo il modello CLARION affronta, formalizza e implementa i processi impliciti ed espliciti, ponendo inoltre l'attenzione sull'interazione tra questi due tipi di processi.

6 Conclusioni

In questa relazione si sono analizzati da diverse prospettive filosofiche-scientifiche i concetti di intelligenza e coscienza biologiche ed artificiali e le teorie ed implementazioni dell'AC (Artificial Consciousness). Inizialmente, nel Capitolo 2, un breve excursus sul concetto di intelligenza ha permesso di compiere un passo copernicano (Cristianini 2023) uscendo dall'antropocentrismo (vedi Sez. 2.3) che restringeva tale abilità ad essere esclusiva umana. Tale processo ci ha consentito di accettare e parlare di intelligenza cosiddetta *artificiale* e di discuterne le implicazioni rispetto a quella debole e la cosiddetta forte (vedi Sez. 2.3.1), quest'ultima considerata avente coscienza e perciò senziente come noi umani.
Nel Capitolo 3, si è affrontato il concetto di coscienza, evidenziando quanto sia ancora più difficile da definire rispetto all'intelligenza. Sono stati esaminati vari quadri teorici e filosofici che tentano di descrivere la coscienza, mettendo in luce la complessità del tema. In breve, il materialismo/fisicalismo assume una relazione causale tra Mondo fisico e stati di coscienza (Sez. 3.1.2), il dualismo sostiene invece che coscienza (mente) e corpo siano divisi (Sez. 3.1.3), l'idealismo presuppone che la coscienza generi tutta la realtà (Sez. 3.1.4) ed infine il panpsichismo assume che la coscienza sia qualcosa presente in tutto l'Universo (Sez. 3.1.5).
Nel Capitolo 4, si è introdotto il concetto di Coscienza Artificiale (AC - Artificial Consciousness), discutendo la sua plausibilità e le controversie ad essa associate, come il *problema dell'intenzionalità* nella Sezione 4.1.1. Nella sezione 4.2 sono stati esplorati alcuni degli aspetti ritenuti necessari dalla letteratura per realizzare una coscienza artificiale. Tra questi, le abitudini rappresentano un approccio originale che sottolinea l'importanza dell'inconscio nella definizione di una mente simile alla nostra, come precedentemente esposto dal filosofo Alva Noë (Noë 2010).
Nel Capitolo 5 sono stati esaminati diversi tentativi e approcci per implementare la coscienza nelle macchine. Questo capitolo ha evidenziato come, nelle implementazioni di AC, manchi spesso una modellazione e formalizzazione adeguata delle relazioni e del funzionamento tra processi inconsci (impliciti) e consci (espliciti).
L'opinione generale di questa relazione è che, come affermato da Piletsky (Piletsky 2019), coscienza e inconscio non siano equivalenti nei processi mentali. In particolare, la presente ricerca sottolinea che, anche se si riuscisse a emulare completamente la coscienza umana, rimarrebbe da chiarire come implementare l'inconscio, ossia i processi inconsci come l'apprendimento e memoria impliciti e le abitudini. In questo contesto, l'implementazione CLARION (Sun et al. 2001), che distingue tra processi impliciti ed espliciti e spiega la loro relazione, è considerata una strada promettente da seguire. Tuttavia, queste affermazioni valgono assumendo che sistemi dotati di coscienza artificiale e IA forte abbiano/possano avere una coscienza umana come la nostra. È infatti cruciale riconoscere che la coscienza artificiale potrebbe invece non replicare esattamente la coscienza umana, e pertanto non necessitare dei processi inconsci. Da questo punto di vista la presente ricerca condivide la visione di Searle del naturalismo biologico per quanto riguarda noi umani e che gli eventi mentali emergano da proprietà biologico-chimiche del cervello (Hauser 2001), tuttavia la estende sostenendo che ci possano essere anche altre definizioni di coscienza. Non siamo veramente così speciali. Da ciò ne consegue che potrebbe essere controintuitivo tentare di emulare il nostro funzionamento e che la coscienza artificiale, o per meglio dire *la coscienza che una macchina è in grado di avere*, potrebbe manifestarsi in modi, modalità che nemmeno conosciamo od ancora potrebbe essersi già manifestata. Pertanto, questo richiede un allontanamento dall'antropocentrismo, accettando che

la coscienza possa manifestarsi in modi diversi nelle macchine così come precedentemente discusso per l'intelligenza. Tuttavia, si avverte la necessità di stabilire criteri per valutare la presenza de *la coscienza che una macchina è in grado di avere*: sebbene il test di Argonov (Argonov 2014) infatti non può essere valido in quanto intriso di antropocentrismo (vedi Sez. 4.1.2), lo studio di Butlin et al (Butlin et al. 2023) rappresenta invece un discreto passo in questo senso, fornendo diversi indicatori sulla base delle teorie scientifiche sulla coscienza, alcuni dei quali discussi nel Capitolo 5 di questa relazione. Inoltre, come già discusso, accettare i concetti di intelligenza artificiale e di *agenti intelligenti* implica adottare una definizione non antropocentrica di *intelligenza* quale "la capacità di un agente di portare a termine un compito in un dato ambiente mostrando un comportamento teleologico" (Cristianini 2023). Un simile approccio di decentralizzazione dell'umano, applicato alla coscienza artificiale, ci porta ad accettare la possibilità dell'esistenza anche di *agenti coscienti*. Nondimeno, se accettiamo che gli agenti artificiali possano essere coscienti in una maniera diversa da quella umana, dobbiamo esplorare e comprendere queste forme di coscienza. Questo potrebbe portare a nuove modalità di interazione con le macchine e a una maggiore comprensione della natura della coscienza stessa. Per quanto riguarda la coscienza (in questo contesto solo umana), la presente ricerca condivide le definizioni del materialismo e che essa sia pertanto legata a fenomeni fisici; esclude tuttavia che un computer possa avere una mente *identica* alla nostra, poichè tale concezione racchiude tutta l'essenza dell'antropocentrismo dal quale si intende liberare. Un'altra "dottrina" da decostruire è l'intellettualismo. Si sostiene, infatti, che la visione di Noe sia coerente e che per emulare l'umano sia necessario dotare robot e IA di abitudini, emozioni e capacità di interazione con l'ambiente, piuttosto che concentrarsi esclusivamente sull'aspetto razionale (Noë 2010). Dal punto di vista della scienza della coscienza, sebbene la presente ricerca non si schieri per una o per l'altra teorie affrontate nella Sezione 3.2, ritiene utile ed innovativo l'approccio di Anil Seth (Seth 2023) che pone l'accento sulle proprietà fenomenologiche della coscienza, piuttosto che su quelle funzionali . Di conseguenza, si sostiene che un dialogo tra scienza e filosofia sia di cruciale importanza per argomenti multidisciplinari come la coscienza. In conclusione, si sostiene come Butlin (Butlin et al. 2023) che nel panorama attuale IA coscienti non ce ne siano, tuttavia si crede che distaccarsi dall'antropocentrismo e sviluppare nuovi criteri per definire e riconoscere la coscienza nelle macchine potrebbe essere la chiave per avanzare in questo campo. La comprensione e l'accettazione di forme diverse di coscienza potrebbe aprire inoltre nuove possibilità nel campo dell'intelligenza artificiale e della filosofia della mente.

Bibliografia

Aksyuk, V. A. (2023). *Consciousness is learning: predictive processing systems that learn by binding may perceive themselves as conscious.* arXiv: `2301.07016 [q-bio.NC]`.

Albus, J. S. (1991). "Outline for a theory of intelligence". In: *≪IEEE Transactions on Systems, Man, and Cybernetics≫* 21.3, pp. 473–509.

Aleksander, Igor (1995). "Artificial neuroconsciousness an update". In: *IWANN '96: Proceedings of the International Workshop on Artificial Neural Networks: From Natural to Artificial Neural Computation.* A cura di Mira José e Sandoval Francisco. Berlin, Heidelberg: Springer., pp. 566–583.

Argonov, Victor (2014). "Experimental Methods for Unraveling the Mind-body Problem: The Phenomenal Judgment Approach". In: *≪Journal of Mind and Behavior≫* 35.1-2, pp. 51–70.

Baars, Bernard J. (1995). *A cognitive theory of consciousness (reprint).* Cambridge: Cambridge University Press, ISBN 978-0-521-30133-6.

Block, N. (2007). "Consciousness, accessibility, and the mesh between psychology and neuroscience." In: *Behavioral and Brain Sciences* 30, pp. 481–499.

Brown, T. et al. (2020). "Language models are few-shot learners. Advances in neural information processing systems". In: *Advances in neural information processing systems.*

Bruntrup, Godehard e Ludwig Jaskolla (2017). *Panpsychism: Contemporary Perspectives.* New York, NY: Oxford University Press, p. 365.

Butler, Gillian et al. (2018). *Managing Your Mind: The Mental Fitness Guide (3rd ed.)* Oxford University Press.

Butlin, Patrick et al. (2023). *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness.* arXiv: `2308.08708 [cs.AI]`.

Buttazzo, Giorgio (2001). "Artificial consciousness: Utopia or real possibility?" In: ≪*Computer.*≫

Cattell, R. B. (1963). "Theory of fluid and crystallized intelligence: A critical experiment". In: ≪*Journal of Educational Psychology*≫ 54, pp. 1–22.

Chalmers, David J. (1995). "Facing up to the problem of consciousness". In: ≪*Journal of Consciousness Studies*≫ 2.3, pp. 200–219.

— (1996). *The Conscious Mind: in Search of a Fundamental Theory*. Oxford University Press.

— (2002). *Philosophy of mind: classical and contemporary readings*. Oxford University Press, pp. 16–17. ISBN: 9780195145816.

— (2011). "A Computational Foundation for the Study of Cognition". In: ≪*Journal of Cognitive Science*≫ 12.4, pp. 325–359.

— (2015a). *Panpsychism and Panprotopsychism*. In Alter, Torin; Nagasawa, Yugin (A cura di.). Consciousness in the Physical World: Perspectives on Russellian Monism. Oxford: Oxford University Press.

— (2015b). *The Combination Problem for Panpsychism*. In Brüntrup, Godehard; Jaskolla, Ludwig (eds.). Panpsychism: Contemporary Perspectives. New York: Oxford University Press.

Clarke, D.S (2004). *Panpsychism: Past and Recent Selected Reading*. State University of New York Press, p. 1.

Cleeremans, Axel e Robert French (2002). *Implicit Learning and Consciousness An Empirical, Philosophical and Computational Consensus in the Making*. Routledge ISBN 9781138877412.

Cole, David (2023). *The Chinese Room Argument*. In Zalta, Edward N. (ed.). Stanford Encyclopedia of Philosophy. URL: `https://plato.stanford.edu/archives/sum2023/entries/chinese-room/`.

Copeland, J. (2004). *The Essential Turing: the ideas that gave birth to the computer age*. Oxford, England: Clarendon Press.

Coren, Stanley (1995). *The Intelligence of Dogs*. Bantam Books.

Coward, L.Andrew e Ron Sun (2004). "Criteria for an effective theory of consciousness and some preliminary attempts". In: *Consciousness and Cognition* 13.2, pp. 268–301. ISSN: 1053-8100. DOI: `https://doi.org/10.1016/j.concog.2003.09.002`. URL: `https://www.sciencedirect.com/science/article/pii/S1053810003001016`.

Crane, Tim e Sarah Patterson (2000). *History of the Mind-Body Problem*. Routledge, pp. 1–3.

Cristianini, Nello (2023). *La Scorciatoia. Come le macchine sono diventate intelligenti senza pensare in modo umano*. il Mulino.

Dainton, B. (2000). *Stream of Consciousness: Unity and Continuity in Conscious Experience*. Routledge.

Della Santina, Cosimo et al. (2024). *Awareness in robotics: An early perspective from the viewpoint of the EIC Pathfinder Challenge "Awareness Inside"*. arXiv: `2402.09030 [cs.RO]`.

Dennet, Daniel (1991). *Consciousness Explained*. The Penguin Press.

— (2009). *Coscienza: Che cosa è*. Editori Laterza.

Doerig, A. et al. (2019). "The unfolding argument: Why IIT and other causal structure theories cannot explain consciousness." In: *Consciousness and Cognition* 72, pp. 49–59.

Elvidge, Jim (2018). *Digital Consciousness: A Transformative Visio*. John Hunt Publishing Limited.

Feuerstein, R. (1990). "The theory of structural modifiability. B. Presseisen (A cura di.), Learning and thinking styles: Classroom interaction". In: ≪*National Education Associations*≫.

Feuerstein, R. et al. (2002). *Dynamic assessments of cognitive modifiability*. ICELP Press.

Friedman, Garret et al. (2023). "The Current of Consciousness: Neural Correlates and Clinical Aspects". In: ≪*Current Neurology and Neuroscience Reports*≫ 23.7, pp. 345–352.

Gardner, Howard (1999). *Intelligence Reframed: Multiple Intelligences for the 21st Century*. New York: Basic Books.

Goertzel, Ben (2014). "Artificial General Intelligence: Concept, State of the Art, and Future Prospects". In: *Journal of Artificial General Intelligence* 5.1, pp. 1–46. DOI: https://doi.org/10.2478/jagi-2014-0001. URL: https://www.sciencedirect.com/science/article/pii/S0364021301000350.

Goleman, Daniel (1996). *Intelligenza emotiva (Emotional Intelligence: Why It Can Matter More Than IQ, 1995)*. traduzione di Isabella Blum e Brunello Lotti, Collana Saggi stranieri, Milano, Rizzoli.

Gottfredson, Linda S. (1997). "Intelligence". In: ≪*Mainstream Science on Intelligence*≫ 24, pp. 13–23.

Graziano, Michael (2013). *Consciousness and the Social Brain*. Oxford University Press.

Haikonen, Pentti O. (2022). "Qualia, Consciousness and Artificial Intelligence". In: *Journal of Artificial Intelligence and Consciousness* 09.03, pp. 409–418. DOI: 10.1142/S2705078522500126. eprint: https://doi.org/10.1142/S2705078522500126. URL: https://doi.org/10.1142/S2705078522500126.

Harnad, Stevan (giu. 1990). "The symbol grounding problem". In: *Physica D: Nonlinear Phenomena* 42.1–3, pp. 335–346. ISSN: 0167-2789. DOI: 10.1016/0167-2789(90)90087-6. URL: http://dx.doi.org/10.1016/0167-2789(90)90087-6.

Hauser, Larry (2001). "Chinese Room Argument". In: *Internet Encyclopedia of Philosophy*.

Herrnstein, Richard J. e Charles Murray (1994). *The Bell Curve: Intelligence and Class Structure in American Life*. Free Press.

Hussain, Amir et al. (2009). *Brain Inspired Cognitive Systems*. New York: Springer Science+Business Media, p. 298.

Keith, Frankish (2016). *Why Panpsychism Is Probably Wrong*. The Atlantic.

Koch, Christof (2020). *The feeling of life itself: why consciousness is widespread but can't be computed*. (First MIT Press paperback edition 2020 ed.), Cambridge, MA London: The MIT Press. ISBN 978-0-262-53955-5.)

Lamme, V. A. F. (2010). "How neuroscience will change our view on consciousness." In: *Cognitive Neuroscience* 1.3, pp. 204–220.

Lau, H. (2019). "Consciousness, metacognition, perceptual reality monitoring." In: *PsyArXiv*.

— (2022). *In Consciousness we Trust: The Cognitive Neuroscience of Subjective Experience*. Oxford University Press.

LeCun, Y. et al. (2015). "Deep learning." In: *Nature* 521, pp. 436–444.

Lenharo, Mariana (2023). "Consciousness theory slammed as 'pseudoscience' — sparking uproar". In: ≪*Nature*≫.

Nagel, Thomas (1974). "What Is It Like to Be a Bat?" In: *The Philosophical Review* 83.4, 435–450 (Cosa si prova a essere un pipistrello? Tr. it. Castelvecchi, Roma 2020.)

Newell, A. e H. A. Simon (1972). *Human problem solving*. Prentice-Hall.

Noë, Alva (2010). *Perché non siamo il nostro cervello: Una teoria radicale della coscienza*. Scienza e idee 202. Milano: Raffaello Cortina Editore. ISBN: 9788860303455.

Panofsky, A. et al. (2021). "How White nationalists mobilize genetics: From genetic ancestry and human biodiversity to counterscience and metapolitics". In: ≪*American Journal of Physical Anthropology*≫ 175.2, pp. 387–398.

Philip, Goff et al. (2017). *Panpsychism*. Zalta, Edward N. (A cura di.). Stanford Encyclopedia of Philosophy. URL: https://plato.stanford.edu/archives/win2017/entries/panpsychism/.

Phillips, I. (2018). *Consciousness, time, and memory*. in The Routledge Handbook of Consciousness, Routledge, pp. 286–297.

Piletsky, Eugene (2019). "Consciousness and Unconsciousness of Artificial Intelligence". In: ≪*Future Human Image*≫ 11, pp. 66–71.

Reggia, James (2013). "The rise of machine consciousness: Studying consciousness with computational models". In: ≪*Neural Networks*≫ 44, pp. 112–131.

Ron, Chrisley (2008). "Philosophical foundations of artificial consciousness". In: ≪*Artificial Intelligence in Medicine.*≫ 44.2, pp. 119–137.

Russel, Stuart J. e Peter Norvig (2021). *Artificial Intelligence: A Modern Approach (4th ed.).* Hoboken: Pearson.

Schachter, D. L. (1997). "Implicit memory: history and current status". In: ≪*Journal of Experimental Psychology: Learning, Memory, and Cognition*≫ 13.3, pp. 501–518.

Searle, John R. (1980). "Minds, brains, and programs". In: ≪*Behavioral and Brain Sciences*≫ 3.

— (1990). "Is the Brain's Mind a Computer Program?" In: ≪*Scientific American*≫ 262.

— (2005). *Mind. A Brief Introduction.* Raffaello Cortina Editore (trad. it. La mente).

Seth, Anil (2023). *Come il cervello crea la nostra coscienza.* Raffaello Cortina Editore.

Shannon, C. e W. Weaver (1963). *The mathematical theory of communication.* University of Illinois Press.

Sun, Ron (2008). *The Cambridge handbook of computational psychology.* Cambridge: Cambridge University Press. ISBN: 9780521857413.

Sun, Ron et al. (2001). "From implicit skills to explicit knowledge: a bottom-up model of skill learning". In: *Cognitive Science* 25.2, pp. 203–244. ISSN: 0364-0213. DOI: https://doi.org/10.1016/S0364-0213(01)00035-0. URL: https://www.sciencedirect.com/science/article/pii/S0364021301000350.

Thaler, S. L. (1998). "The emerging intelligence and its critical look at us". In: ≪*Journal of Near-Death Studies*≫ 17.1, pp. 21–29.

Tirri, Kirsi (2011). *Measuring Multiple Intelligences and Moral Sensitivities in Education. Moral Development and Citizenship Education.* Springer.

Tononi, Giulio (2004). "An information integration theory of consciousness". In: ≪*BMC Neuroscience*≫ 5.1.

Tononi, Giulio e Christof Koch (2015). "Consciousness: here, there and everywhere?" In: ≪*Philosophical Transactions of the Royal Society B: Biological Sciences*≫ 370.1668.

Tulving, E. (1972). "Episodic and Semantic Memory." In: ≪*Organization of Memory*≫. A cura di E. Tulving e W. Donaldson. Cambridge, MA: Academic Press., pp. 381–403.

— (1985). "Memory and consciousness." In: ≪*Organization of Memory*≫ 26.1.

Ullman, Michael T. (2004). "Contributions of memory circuits to language: the declarative/procedural model". In: ≪*Cognition*≫ 92.1-2, pp. 231–70.

Vaswani, A. et al. (2023). *Attention is all you need.* arXiv: 1706.03762.

Wilson, Nicholas R. et al. (2009). "A motivationally-based simulation of performance degradation under pressure". In: *Neural Networks* 22.5. Advances in Neural Networks Research: IJCNN2009, pp. 502–508. ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2009.06.022. URL: https://www.sciencedirect.com/science/article/pii/S0893608009001324.

Wood, Wendy e David T. Neal (2007). "A new look at habits and the habit-goal interface". In: *Psychological Review.* 114.4, pp. 43–863.