THIS IS A PRE-PRINT, WEB VERSION OF THE
PH.D. DISSERTATION:

"I^{+}"

FOR THE FULL, FINAL VERSION, PLEASE CONTACT ME.

UNIVERSITA' VITA-SALUTE SAN RAFFAELE

PROGRAMMA DI DOTTORATO IN
FILOSOFIA E SCIENZE COGNITIVE

I^+

L'accelerazione esponenziale dell'Intelligenza

Director Of Studies: Prof. Michele Di Francesco

Tesi di DOTTORATO di RICERCA di: Marta Rossi

Matr. 004942

Ciclo di Dottorato XXV

SSD: M-FIL/03

Anno Accademico 2011/2012

UNIVERSITA' VITA-SALUTE SAN RAFFAELE
PROGRAMMA DI DOTTORATO IN
FILOSOFIA E SCIENZE COGNITIVE

Tesi di Dottorato di: Marta Rossi DoS: Prof. Michele di Francesco

I^+. L'accelerazione esponenziale dell'Intelligenza

L'evoluzione tecnologica degli ultimi decenni ha cambiato repentinamente ogni aspetto della vita quotidiana di individui e società. Per alcuni studiosi, questo non è che l'inizio: il progresso sarà sempre più veloce, fino a raggiungere, nell'arco di qualche decennio, la *Singolarità Tecnologica*, un punto di non ritorno nello sviluppo della civiltà umana in cui sistemi naturali e artificiali convivranno in una società del futuro, profondamente mutata. Lo scopo di questo lavoro è valutare lo scenario della Singolarità con gli strumenti concettuali propri della riflessione filosofica contemporanea: in particolare, la tesi esamina i principali argomenti a favore di questa ipotesi, ne discute possibili obiezioni, svela i presupposti teorici impliciti nella visione e collega il dibattito tecnico-scientifico a svariati argomenti filosofici.

La dissertazione si divide in tre parti principali: dopo aver introdotto la Singolarità ed effettuato una breve ricognizione del territorio, il secondo capitolo affronta l'aspetto dell'accelerazione tecnologica - quali ragioni abbiamo per credere che il progresso tecnologico continuerà ad accelerare esponenzialmente? Il terzo capitolo affronta invece il secondo punto chiave, quello dell'accelerazione di intelligenza: per molti studiosi, infatti, l'avvento della Singolarità non sarà possibile se non attraverso la costruzione di macchine pensanti, grazie alle quali potremo diventare (individualmente e/o collettivamente) sempre più intelligenti. Intelligenza Artificiale e filosofia della mente vengono inquadrate all'interno di una prospettiva tecnologica; in particolare, la discussione evidenzia una sorta di "tesi di indipendenza" tra assunti metafisici e Singolarità: anche se molte tesi portate a sostegno dell'accelerazione di intelligenza fossero parzialmente errate, questo non sarebbe di per sé un motivo sufficiente per rigettare la Singolarità. Infine, il quarto capitolo discute l'aspetto umanamente più coinvolgente del dibattito: la desiderabilità di questa "nuova era" a partire dalle conseguenze che la Singolarità avrà per la nostra vita di persone, agenti morali, cittadini in una comunità.

Indice

1.Singolarità

"*L'uomo è una corda tesa fra l'animale e il superuomo,*
una corda sopra un abisso."

(Friedrich Nietzsche)

1.1 Preliminari

Cominciamo con una semplice domanda: come sarà il mondo nel 2060? Raymond Kurzweil[1] ci dà un assaggio degli scenari futuri quando ci dice che 'riusciremo a reingegnerizzare tutti gli organi e i sistemi dei nostri organismi e cervelli biologici in modo che siano di gran lunga più potenti'[2]. Più in dettaglio, secondo lo scienziato ci troveremo nella situazione in cui 'miliardi di nanorobot nei capillari del cervello estenderanno enormemente l'intelligenza umana'[3] e che, quasi sicuramente, 'la capacità umana di comprendere le emozioni e di rispondere in modo appropriato (...) sarà (...) capita e padroneggiata dalla futura intelligenza delle macchine'[4]. Estendendo il concetto di potenziamento fisico/mentale dell'essere umano, ci troveremo al punto in cui 'vivremo così a lungo da vivere per sempre'[5]. L'insieme di questi sconvolgimenti epocali portati dal repentino sviluppo tecnologico viene generalmente etichettato come "Singolarità tecnologica", in analogia con una singolarità in fisica – ovvero una porzione di spazio-tempo dove le leggi conosciute non si applicano più.

Preso *letteralmente*, lo scenario della Singolarità può apparire incredibile. Al momento di scrivere queste righe (settembre 2012), l'utilizzo dei nanorobot in campo industriale è ancora agli albori: il mio computer, non solo non capisce le emozioni – banale, ma, molto spesso, sbaglia perfino le correzioni ortografiche automatiche; d'altra parte, la vita media in Italia si aggira attorno agli 88 anni – per una donna, come me, nata nel 1986. Se ci si ferma alla lettura delle prime pagine dei testi e degli articoli

[1] Raymond Kurzweil è inventore, scienziato, imprenditore, divulgatore scientifico, nonchè il più autorevole personaggio nel panorama della Singolarità Tecnologica. Il suo Kurzweil (2005) - tradotto in italiano come Kurzweil (2008) - sarà il libro principalmente discusso in *questo* lavoro.
[2] Kurzweil (2008), p. 27.
[3] Kurzweil (2008), p. 28.
[4] Kurzweil (2008), p. 28.
[5] Kurzweil (2004), p. 1.

sull'argomento, nonché ai relativi slogan, non stupisce affatto che l'atteggiamento accademico prevalente (o quantomeno, prevalente fino allo scorso decennio) sia quello di non prendere sul serio *nulla* di ciò che viene detto sull'argomento.

Tuttavia, "spulciando" – con un po' di umiltà – nella storia della scienza e della tecnologia, le cose appaiono molto meno chiare e le nostre certezze potrebbero quantomeno iniziare a vacillare: nel 1989, ad esempio, la prestigiosa multinazionale di consulenza McKingsey, decretava in un report che 'il mercato per i telefoni cellulari [allora più simili a cabine telefoniche in miniatura] non ha alcun futuro'; per non parlare dell'inizio degli anni Novanta, quando scienziati e biochimici, appena riusciti a sequenziare un decimillesimo del genoma umano[6], dichiaravano – senza mezzi termini – che *almeno un secolo* sarebbe stato necessario per il completamento dell'impresa. Oggi, secondo *Wikipedia* – altro miracolo del "futuro" – sul nostro pianeta ci sono quasi 6 miliardi di telefoni cellulari e, neanche a dirlo, giusto nel 2000 è stato raggiunto il primo *draft* di un sequenziamento completo del genoma umano. Facciamoci dunque una seconda domanda e chiediamoci: come facciamo a sapere che le affermazioni "incredibili" di prima non saranno *ovvie* fra venti o trenta anni? Proprio per questo motivo, il nostro atteggiamento proverà ad essere esattamente l'opposto di quello appena visto, ovvero ci chiederemo cosa accadrebbe se prendessimo sul serio tali – apparentemente bizzarri – argomenti.

La Singolarità, come l'Essere, si dice in molti modi (sulle differenze ci soffermeremo più avanti). Per gli scopi di questa prima ricognizione del territorio, la seguente caratterizzazione della tesi "La Singolarità è vicina" (d'ora in avanti **SV**) è sufficientemente generale da coprire lo spirito (se non la lettera) di quello che i maggiori sostenitori propongono e sufficientemente precisa da poter essere criticamente discussa e analizzata. **SV** è quindi la congiunzione di due tesi ben specifiche:

SV$_1$) Lo sviluppo tecnologico ha natura *esponenziale*.

SV$_2$) Siamo sull'orlo di una nuova era nella civiltà umana.

[6] Kurzweil (2008), p. 13.

Parleremo in seguito dettagliatamente di entrambe, ma qualche precisazione preliminare è sicuramente indispensabile. La natura esponenziale e *non* lineare del progresso è al cuore di $\mathbf{SV}_1$: quando consideriamo come incredibili le predizioni accennate nell'introduzione – o quando i biochimici all'inizio degli anni Novanta stimavano la probabile durata del Progetto Genoma Umano in centinaia di anni – è perché adottiamo implicitamente una visione "lineare" del progresso, ovvero se abbiamo impiegato 2,5 anni per completare il 5% di un dato compito, ne serviranno quasi altri cinquanta per terminarlo. La naturale tendenza umana a pensare linearmente sarebbe anche una probabile spiegazione[7] del perché lo scenario dipinto dai sostenitori di **SV** sia da molti automaticamente etichettato come incredibile e del perché, più in generale, l'idea stessa di Singolarità sia piuttosto recente (ci basti pensare al fatto che il termine risale a una storica conferenza di Vernon Vinge del 1993 e che prima degli anni Sessanta[8] non c'è proprio traccia di idee simili a $\mathbf{SV}_1$). D'altro canto, $\mathbf{SV}_2$ è una tesi con una collocazione ben delimitata del tempo: anche assumendo che $\mathbf{SV}_1$ sia vera, è quasi un "caso" che noi oggi siamo nel punto giusto della storia umana per cogliere i frutti della prossima grande rivoluzione tecnologica – non Napoleone, non Newton, non Gauss, *noi* abbiamo la possibilità reale di essere i testimoni di quello che viene spesso definito un "cambio di paradigma" nello sviluppo della civiltà umana. Raymond Kurzweil chiama questa epoca la 'quinta era', un'epoca caratterizzata dalla "fusione di tecnologia e intelligenza umana". Il simbolo di questa nuova era sarà un nuovo tipo di intelligenza – artificiale, naturale o ibrida, – e quel che succederà sarà il superamento delle potenzialità del cervello umano grazie alla tecnologia. E, poiché la natura esponenziale del progresso di $\mathbf{SV}_1$ ovviamente si manterrà inalterata anche nella nuova fase dell'umanità espressa da $\mathbf{SV}_2$, la natura della cosiddetta "accelerazione di intelligenza" sarà essa stessa esponenziale. In altre parole, Kurzweil può affermare, come dicevamo nelle pagine precedenti, che 'vivremo così a lungo da vivere per sempre' perché confida sul fatto che presto saremo, semplicemente, troppo intelligenti per morire.

[7] Vedi ad es. Kurzweil (2008), pp. 10-14.
[8] Cfr. Good (1965).

1.1.1 Questioni empiriche vs questioni concettuali

Possiamo dividere in due categorie gli argomenti a favore della Singolarità: empirici e concettuali. Nella sezione precedente abbiamo ad esempio affermato che lo sviluppo del progresso tecnologico ha un trend *esponenziale* e su questa base abbiamo esposto diverse previsioni (ad es., che presto avremo macchine più intelligenti di qualsiasi essere umano); ovviamente, questo tipo di argomenti, preso in isolamento da altre considerazioni, presta il fianco a due tipi di obiezioni:

1) I dati storici permettono differenti, ugualmente compatibili, interpretazioni.
2) Un trend potrebbe tranquillamente cessare per un qualsiasi fattore, endogeno o esogeno.

Rispetto ad (1) è importante osservare fin da ora che molti studiosi[9] hanno sostenuto che il trend evidenziato da Kurzweil non sia esponenziale; rispetto a (2), è semplice osservare che qualsiasi previsione sullo sviluppo dell'umanità sia da intendersi sotto diverse assunzioni non scontate – ad es., senza che una guerra nucleare ci riporti al medioevo nel prossimo secolo[10].

Accanto dunque ad argomenti puramente *a posteriori* – i.e. 'le cose sono sempre andate così e sempre così andranno' – i sostenitori della Singolarità propongono argomenti *a priori* – i.e. 'le cose sono andate così e vanno così perché *devono* andare così'. In una sezione apposita analizzeremo dunque il modello di sviluppo proposto da Kurzweil e altri per valutarne le assunzioni chiave.

Accanto agli argomenti *pro* Singolarità, tutta una serie di proposte, scenari, suggestioni collegati ad essa si basano su assunzioni (a volte esplicite, a volte no) filosofiche forti – ed è qui che il lavoro del filosofo è forse più utile nel comporre i pezzi del puzzle in un insieme omogeneo e il più possibile coerente. Nella prossima sezione menzioneremo esplicitamente i dibattiti filosofici collegati alla Singolarità che andremo, in un modo o nell'altro, ad analizzare in questo lavoro, ma un esempio può fin d'ora essere utile: probabilmente *tutti* i sostenitori della Singolarità condividono la tesi che presto i nostri computer (o un sistema artificiale equivalente) saranno più

[9] Vedi ad es. Modis (2006), Modis (2002).
[10] Curiosamente, l'avvicinarsi alla Singolarità potrebbe rendere, almeno inizialmente, più e *non* meno probabile, l'insorgere di eventi catastrofici su larga scala. Ritorneremo sull'argomento in seguito.

intelligenti di noi, tesi che ovviamente presuppone la verità di una dottrina metafisica sugli stati mentali ben precisa, nota in letteratura con il nome di *funzionalismo*[11]. Se, *per impossibile*, si "dimostrasse" che il funzionalismo è perlopiù falso, tutta l'impalcatura della Singolarità comincerebbe apparentemente a scricchiolare[12].

1.1.2 Perché occuparsene

Lo scenario della Singolarità ha relazione intima con alcuni tra i principali dibattiti della filosofia contemporanea, non solo per gli argomenti proposti, ma anche per il tipo di tecnologie coinvolte: ad es., grazie al *mind uploading*[13] molti "esperimenti mentali" cesseranno di essere puramente "mentali" e diventeranno *problemi pratici.* Inoltre, come anticipato, molte discussioni sulla Singolarità avvengono *presupponendo* la verità di dottrine metafisiche, psicologiche e morali tutt'altro che scontate. In particolare, i seguenti quesiti filosofici sono tutti di rilevanza primaria quando si discute di Singolarità:

Ontologia del mentale: come si costruisce una mente? Cosa significa possederne una? Quali diritti/doveri si accompagnano al possesso di una mente?

Identità personale: è possibile sopravvivere in una forma di vita non a base di carbonio? Gli esseri umani sono *token* o *type*? Se l'immortalità è desiderabile, di quale parte di me *dovrei* volere la sopravvivenza?

Dibattito morale sul potenziamento umano: è giusto potenziarsi? Ho il diritto di farlo in quanto individuo libero? La società nel suo complesso dovrebbe promuovere il potenziamento umano (come fa, ad esempio, con l'istruzione)?

[11] Se l'attesa è troppa – un intero capitolo di questa tesi sarà dedicata al funzionalismo – si veda Levin (2010) per un'ottima introduzione.
[12] O forse no. Dipanare i legami tra la Singolarità e le varie dottrine metafisiche è, come dicevamo nella nota precedente, uno dei compiti di questo lavoro.
[13] Per una *roadmap* filosoficamente ispirata si veda Bostrom, Sandberg (2010).

Dibattito sociale/politico sullo sviluppo: quali conseguenze sociali avrà lo sviluppo tecnologico? Cosa significa che lo "sviluppo" è un "progresso"? Possiamo disegnare sistemi/istituzioni per governare la transizione nel modo "migliore" possibile?

Oltre a questi motivi "indiretti", ci sono motivi *diretti* per occuparsi di Singolarità da una prospettiva filosofica? Crediamo di sì. Per la maggior parte, le suggestioni di questi scenari di futurologia non sono altro che il risultato di portare a conseguenze estreme fenomeni, attitudini e concetti già ampiamente a disposizione oggi – una pratica con cui la buona filosofia dovrebbe continuamente cimentarsi. In secondo luogo, è proprio nel dipanare i nodi concettuali di queste situazioni, magari improbabili, che l'utilità di un'analisi filosofica emerge prepotentemente. Anche valutando la probabilità della Singolarità come estremamente bassa, il possibile valore/disvalore connesso a tale evento è così grande[14] da renderlo certamente uno degli argomenti principali dell'agenda filosofica contemporanea – cosa che attualmente *non* è, se non in alcuni settori molto specifici. Parafrasando Nick Bostrom, crediamo che le questioni connesse alla Singolarità siano cosi importanti da dover essere investigate con la stessa serietà accademica e risorse riservate per altre questioni cruciali del nostro tempo. Scopo di *questo* lavoro è (anche) convincere il lettore che questo sia vero.

1.1.3 La struttura della tesi

Ora che l'argomento del lavoro dovrebbe essere stato inquadrato a sufficienza, una ricognizione del territorio che andremo ad esplorare è indubbiamente utile. Il capitolo di introduzione si chiude con una breve presentazione dei due concetti chiave di **SV**: l'accelerazione della tecnologia e quella di intelligenza che servono fondamentalmente ad introdurre le problematiche del Capitolo 2 e del Capitolo 3.

Nel secondo capitolo, discuteremo in dettaglio $\mathbf{SV}_1$, presentando argomenti *pro* e *contro* lo sviluppo esponenziale della tecnologia: analizzeremo criticamente il concetto

[14] Nel valutare l'utilità (o il danno) connesso ad un evento, si pondera l'utilità per la probabilità che l'evento si verifichi: ad es., l'utilità di comprare un biglietto della lotteria è bassa perché un premio consistente viene moltiplicato per una probabilità di vittoria molto bassa. Nel caso in discussione, per quanto bassa si stimi la probabilità, i cambiamenti connessi alla Singolarità sono così potenzialmente grandi che plausibilmente il valore atteso dell'evento rimane significativo.

di sviluppo presentato da Kurzweil e valuteremo le obiezioni sollevate sia agli argomenti *a priori*, sia agli argomenti *empirici*.

Nel terzo capitolo, renderemo più precisa l'idea (di per sé piuttosto vaga) di "accelerazione di intelligenza", connettendo l'idea alle scienze cognitive – vecchie e nuove che siano – e alla filosofia della mente: un risultato piuttosto sorprendente della discussione è che, complessivamente, le assunzioni implicite in $\mathbf{SV}_2$, necessitano di una "base metafisica" piuttosto scarsa; in altre parole, anche se molte dottrine filosofiche associate alla Singolarità fossero pesantemente scorrette, la possibilità di uno sviluppo esponenziale dell'intelligenza non sarebbe perciò, *ipso facto*, impossibile.

Dato il bilancio fondamentalmente positivo emerso da questi capitoli, l'ultima sezione del lavoro affronterà la questione *psicologicamente* più ardua: anche ammesso di accettare la Singolarità come un futuro davvero plausibile, abbiamo motivi di desiderarne l'avvento (e non piuttosto di ostacolarne la venuta)? Rispondere a questo interrogativo ci porterà ad avventurarci in altre aree della filosofia e a valutare in contesti diversi il *trade-off* tra vantaggi individuali e conseguenze sociali su larga scala di determinati "salti tecnologici".

1.2 Accelerazione tecnologica

Come molte idee controverse, l'idea della Singolarità parte da un'osservazione relativamente banale e certamente condivisibile: negli ultimi cinquanta anni il progresso tecnologico è stato molto più rapido che nei precedenti cinquanta; non solo, andando un po' indietro con gli anni, è innegabile che grandi cambiamenti di "paradigma tecnologico" siano piuttosto ravvicinati nella nostra epoca rispetto a quanto avveniva in passato – basti pensare che in sessant'anni si è passati da un mondo senza computer ad un mondo con un miliardo di smartphone, ciascuno più potente di qualsiasi PC nei primi decenni della storia dell'informatica[15]. Non solo: se concentriamo la nostra attenzione all'*Information Technology*, la cosiddetta "Legge di Moore" è stata descritta a metà anni Sessanta ed ha predetto in modo spettacolare il trend nello sviluppo delle performance

[15] Alan Turing - nel pioneristico Turing (1950) - prediceva (pensando di fornire una stima ottimistica) computer con una memoria di 128 MB per la fine del millennio. 128 MB è circa 500 volte meno di quanto oggi possieda un comune cellulare.

dei processori[16]. La Legge originale – pubblicata dal co-fondatore Intel Gordon Moore nel 1965[17] – osservava che il numero di componenti presenti nei circuiti integrati raddoppia ogni anno; poiché il numero dei transistor in un circuito è fortemente correlato con altre capacità fondamentali delle macchine digitali (capacità di elaborazione, ma anche memoria e risoluzione dei sensori), non stupisce che la Legge trovi vasta applicazione in tutta l'informatica (una popolare formulazione della Legge dice infatti che la *performance* dei processori raddoppia ogni 18 mesi).

La tesi della Singolarità diventa controversa proprio quando si distacca dalla originale Legge di Moore, *estendendola*: nel *passato*, sostenendo che la legge vale da molto prima dell'invenzione dei transistor, nel *futuro*, sostenendo che gli attuali PC stanno per essere soppiantati da una nuova forma di sistemi computazionali. Kurzweil elenca infatti *quattro* paradigmi computazionali prima dei computer[18] e ipotizza che i computer quantistici o il *DNA computing* soppianteranno presto l'attuale tecnologia. In uno slogan, la Legge di Moore diventa una legge sulla computazione in sistemi fisici in generale, piuttosto che una generalizzazione su una tecnologica specifica (i transistor sui circuiti). In questo modo, Kurzweil può predire che, quando la legge non sarà più valida per i PC, saranno altri dispositivi a portarne avanti il corso e sarà proprio questa esplosione esponenziale continua di potenza computazionale a sostenere l'arrivo della Singolarità.

1.3 Accelerazione di intelligenza

Parassitica rispetto all'accelerazione tecnologica, l'accelerazione di intelligenza è il secondo elemento chiave nell'avvicinamento alla Singolarità. Mentre, come abbiamo visto, esistono evidenze non controverse sul fatto che, quantomeno nell'ultimo secolo, la tecnologia stia accelerando in modo deciso, non sembrano esserci uguali segnali per quanto riguarda l'intelligenza: esiste forse un senso per cui l'umanità, nel *suo complesso*, diventa più intelligente con l'accumularsi di conoscenza e il passaggio di *know-how* tra le generazioni, ma non sembra che gli esseri umani, presi *individualmente*, siano più

[16] In verità, lo ha predetto così accuratamente che essa stessa è usata dai produttori per stabilire *benchmark* e piani di investimento, rendendola simile ora a una "*self-fulfilling prophecy*".
[17] Cfr. Moore (1965), p. 4.
[18] In particolare, Kurzweil elenca elettromeccanica, relè, valvole termoioniche, transistor come tecnologie precedenti ai circuiti integrati odierni (vedi Kurzweil (2008) pp. 63-68).

intelligenti di un secolo fa (di certo, non *esponenzialmente* più intelligenti). Nel caso dell'intelligenza, dunque, non basta che un trend si mantenga in un certo modo; è necessario piuttosto che avvenga un cambiamento *radicale* perché essa possa accelerare esponenzialmente.

Fondamentalmente, gli approcci che discuteremo sono di due tipi: approcci ibridi, in cui il cambiamento avviene su un essere umano già esistente attraverso impianti artificiali o somministrazione di sostanze potenzianti; approcci artificiali, in cui un sistema intelligente viene esplicitamente costruito da un altro sistema intelligente. Nel primo caso, le tecnologie di riferimento sono legate allo sviluppo biomedico e farmacologico e il dibattito filosofico più vicino (spesso di naturale morale) è quello sull'*enhancement*[19]; nel secondo caso, le tecnologie di riferimento sono legate all'Intelligenza Artificiale e alle scienze cognitive e il dibattito filosofico più vicino è quello sulla natura degli stati mentali (il funzionalismo, precedentemente accennato, e la tesi dell'IA forte) e l'identità personale.

[19] Vedi ad esempio Savulescu, Bostrom (2011) e Kahane, Savulescu, ter Meulen (2011).

2. Tecnologia

"Ogni tecnologia sufficientemente avanzata
è indistinguibile dalla magia."

(Arthur C. Clarke)

2.1 La legge dei ritorni accelerati

Come anticipato nel capitolo introduttivo, il punto di partenza in ogni spiegazione della Singolarità è senza dubbio l'evidenza statistica presentata da Kurzweil e colleghi: è virtualmente impossibile leggere una introduzione alla Singolarità senza imbattersi in un grafico logaritmico sulla Legge di Moore o qualche altro trend esponenziale (Fig. 1). Pertanto, non sorprende che una notevole parte del dibattito sull'argomento sia incentrata su tali famigerati grafici: tuttavia, vedremo che la battaglia sulla *forma* dello sviluppo tecnologico si gioca in realtà su tutta una serie di fronti. La più sistematica, famosa e discussa proposta dell'accelerazione tecnologica si deve, ovviamente, a Raymond Kurzweil: il "pacchetto" di evidenza empirica e argomenti *a priori* da lui portati a sostegno dell'arrivo della Singolarità è conosciuto come "Legge dei Ritorni Accelerati" (**LRT** d'ora in poi) – benché non sia l'unica proposta sul mercato, è la visione che principalmente discuteremo in questo capitolo, soffermandoci, dove interessante, su alcune originali proposte alternative:

LRT) Il tasso di progresso di un processo evolutivo (biologico e tecnologico) cresce esponenzialmente nel tempo. Di conseguenza, i "ritorni" di un processo evolutivo (e.g. velocità, tempi di esecuzioni, mantenimento dei costi, potenza, etc.) crescono esponenzialmente nel tempo[20].

Abbiamo già incontrato la Legge di Moore sullo sviluppo esponenziale delle capacità dei transistor, ma per Kurzweil è solo una delle molteplici realizzazioni della **LRT**: dal costo del sequenziamento del DNA alla risoluzione della risonanza magnetica

[20] Vedi Kurzweil (2008), pp. 40-43.

funzionale (fMRI), dai bit spediti in Rete al PIL dell'intera economia, Kurzweil (2008) contiene una impressionante carrellata di graifici che apparentemente supportano la **LRT**. D'altra parte, lo stesso Kurzweil è molto esplicito riguardo la portata stessa della sua idea:

> 'La legge dei ritorni accelerati vale per tutte le tecnologie, anzi, per tutti i processi evolutivi. La si può seguire con precisione nelle tecnologie basate sull'informazione perché abbiamo indici ben definiti (…) per misurarla.'[21]

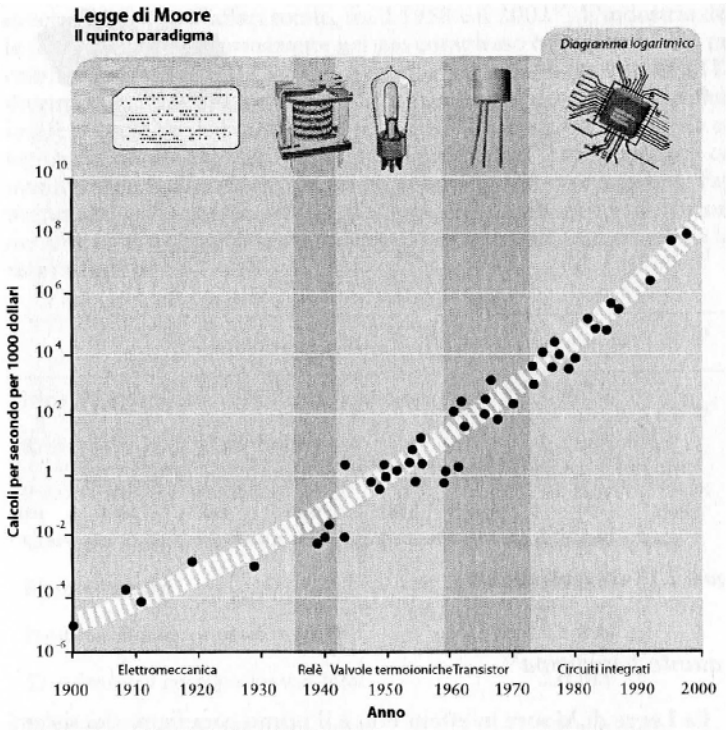

Fig.1: Legge di Moore attraverso diversi paradigmi computazionali (Kurzweil (2008), p. 64).

La **LRT** – dice Kurzweil – 'descrive l'accelerazione del ritmo di un processo evolutivo e la crescita esponenziale dei suoi prodotti'[22]; inoltre, Kurzweil aggiunge che

[21] Kurzweil, (2008), p. 58.
[22] Kurzweil (2008), p. 36.

'la Singolarità è l'inevitabile risultato della legge dei ritorni accelerati'[23]: se quindi l'intera visione si basa sulla **LRT**, argomentare per la legge diviene dunque un punto cruciale della dialettica kurzweliana. Ci sono almeno tre tipi di evidenze che Kurzweil porta a sostegno della verità di **LRT**:

i) Evidenza "empirica".
ii) Evidenza "concettuale".
iii) Evidenza "matematica".

Esaminiamo le tre tipologie una alla volta.

2.1.1 Evidenza empirica

La tipologia (i) è già stata menzionata più volte a proposito dei grafici logaritmici continuamente proposti dai sostenitori della Singolarità: è l'argomento più semplice da esporre e più veloce da capire (un'immagine, si sa, vale più di mille parole): in breve, se disponiamo su un grafico, avente scala logaritmica, lo sviluppo di un processo produttivo nel tempo (memoria nei calcolatori, velocità dei processori, output economici, diffusione di Internet, precisione della fMRI, etc.) troveremo invariabilmente che i dati formeranno una linea retta, indicando dunque un ritmo di sviluppo *esponenziale*[24]. Ovviamente, quando osserviamo una *particolare* tecnologia per un periodo di tempo sufficiente, troveremo che ad un certo punto lo sviluppo rallenterà invece di continuare ad accelerare: ad esempio, ad un certo punto la tecnologia dei relè aveva smesso di migliorare[25]. Significa forse che la **LRT** non vale più? No, ma per capire questa risposta e dunque apprezzare in pieno la strategia argomentativa, occorre anche introdurre l'*analisi concettuale* sui cicli di vita di qualsiasi tecnologia.

[23] Kurzweil (2008), p. 37.
[24] Può essere utile qui ricordare che una linea retta, in un grafico i cui assi abbiano scala logaritmica, equivale a una crescita esponenziale (i.e. una linea "curva") in un grafico disegnato su scala usuale.
[25] Cfr. Kurzweil (2008), p. 64.

2.1.2 Evidenza concettuale

Secondo Kurzweil, ogni tecnologia attraversa tre fasi fondamentali[26]:

1) Crescita *lenta*: la prima fase della crescita esponenziale.
2) Crescita *rapida*: la fase avanzata, esplosiva, della crescita esponenziale.
3) *Livellamento*: la fase matura, quando il paradigma si stabilizza.

Quando si visualizzano queste tre fasi insieme su un grafico, il risultato è una curva "a *S*" (la cosiddetta crescita logistica), quella tipica dei fenomeni di contagio: all'inizio il numero di infetti dall'influenza in una popolazione cresce molto piano fino ad un valore *soglia*; a quel punto, in poco tempo, la gran parte della popolazione viene infettata fino a raggiungere uno stato di equilibrio – poiché, banalmente, non ci sono più individui contagiabili a disposizione. Come si concilia allora la crescita logistica con lo sviluppo esponenziale essenziale per $\mathbf{SV}_1$? La risposta è che la crescita esponenziale è il risultato di *S-curve* successive (Fig. 2): quando la nuova tecnologia rimpiazza quella precedente, il trend complessivo risulta esponenziale anche se, di fatto, è la composizione di diverse crescite logistiche. In particolare, la chiave di ogni processo evolutivo sono i *feedback* positivi tra uno stadio dell'evoluzione e il successivo: ogni miglioramento viene incorporato nello stadio successivo, che a sua volta produce nuovi miglioramenti e così via[27].

Ritornando al caso delle performance dei relè nei primi calcolatori, l'evidenza empirica e l'analisi concettuale forniscono la seguente diagnosi: è assolutamente vero che quella particolare tecnologia era in fase di livellamento, ma questo non ha interrotto il trend di sviluppo più generale, dato che l'arrivo della nuova tecnologia (i transistor) ha continuato il miglioramento esponenziale dei calcolatori. In altre parole, l'evidenza empirica anomala (la crescita logistica e non esponenziale) che possiamo osservare in certi grafici svanisce se facciamo "*zoom out*" sullo stesso grafico e analizziamo un periodo di tempo più ampio – contenente, cioè, la tecnologia successiva a quella che

[26] Cfr. Kurzweil (2008), p. 43.
[27] Nel caso della tecnologia Kurzweil aggiunge anche un secondo ciclo di *feedback* positivi: più una tecnologia diventa efficiente (ad es., migliorano le performance dei transistor) maggiori risorse la società investe nella tecnologia, portando ad un secondo livello di crescita esponenziale (ovvero, il tasso di crescita cresce esso stesso esponenzialmente, vedi Kurzweil (2008) p. 495). Tuttavia, nessuna delle obiezioni che esamineremo in seguito verterà su questo particolare: quando non c'è pericolo di confusione, ometteremo dunque il riferimento esplicito a questo secondo livello di crescita.

stiamo esaminando. Allo stesso modo, qualcuno potrebbe obiettare che la legge di Moore sui transistor potrebbe rallentare e raggiungere una sorta di livellamento molto presto: possiamo già vedere come questa critica alla **LRT** possa essere facilmente confutata da considerazioni analoghe a quelle appena svolte; come in quel caso una nuova tecnologia aveva soppiantato la vecchia, così un nuovo modello di computazione soppianterà quello attuale quando comincerà a rallentare.

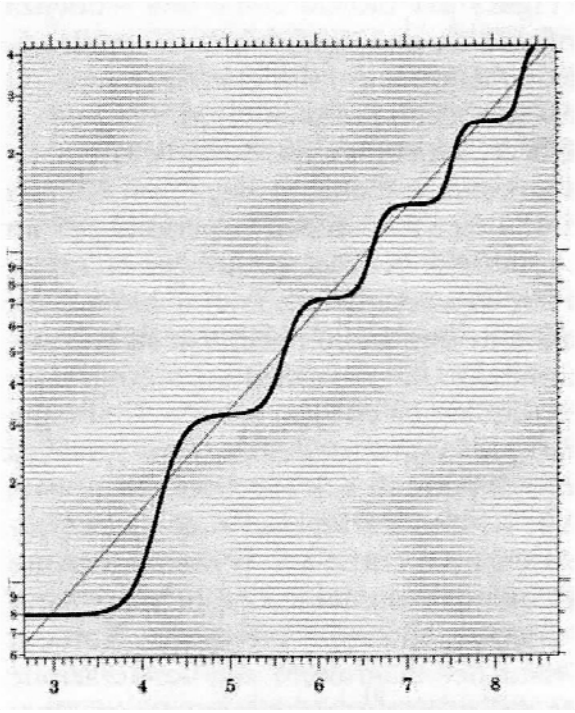

Fig. 2: Crescita esponenziale costituita da una cascata di curve ad S (Kurzweil (2008), p. 44).

2.1.3 Evidenza matematica

Infine, c'è un'ulteriore tipologia di evidenza da esaminare, quella dei modelli matematici di sviluppo. L'appendice a Kurzweil (2008) contiene i dettagli del modello astratto di sviluppo per i ritorni accelerati; come facilmente intuibile, è questo l'argomento più astratto e generale, nonché quello meno conosciuto al grande pubblico: cercheremo di fornirne qui un resoconto non tecnico ma comunque il più possibile esaustivo. Il modello prevede tre variabili fondamentali: nonostante sia formulato per le

tecnologie informatiche maggiormente pertinenti all'accelerazione di intelligenza[28], 'le formule sono simili per altri aspetti dell'evoluzione'[29]:

V = la *capacità di calcolo* (misurata in calcoli al secondo per costo unitario).
W = la *conoscenza dell'umanità* nella costruzione di dispositivi computazionali.
t = il *tempo*.

Come accade in ogni modello, è la relazione funzionale che definiremo tra le variabili ad essere cruciale. La prima assunzione del modello è dunque la seguente: W e V sono collegati in modo *lineare* – ogni volta che apprendiamo qualcosa di nuovo su come costruire un calcolatore migliore, la velocità dei nostri calcolatori aumenta linearmente:

1) $V = c_1 W$

In altre parole, (supponendo c_1=10 , per esemplificare il concetto) se prima la nostra conoscenza era 1 e le prestazioni 10, ora che abbiamo appreso una nuova nozione, la conoscenza è 2 e le prestazioni 20. Come si vede, l'assunzione è molto ragionevole, poiché parte dal presupposto che il miglioramento sia *incrementale*: un modulo dopo l'altro, una scoperta dopo l'altra, i nostri dispositivi artificiali migliorano. La seconda assunzione riguarda il tasso di cambiamento della nostra conoscenza; in particolare, richiediamo che il cambiamento in W sia proporzionale alla velocità dei nostri calcolatori – più sono veloci i nostri calcolatori, più in fretta acquisiamo nuove informazioni:

2) $dW/dt = c_2 V$

Sostituendo (1) in (2) il risultato è dunque

3) $dW/dt = c_1 c_2 W$

[28] L'esplicita giustificazione di Kurzweil su questo punto è una ulteriore evidenza indiretta di quanto $\mathbf{SV}_1$ e $\mathbf{SV}_2$ siano, per l'inventore americano, due facce della stessa medaglia.
[29] Kurzweil (2008), p. 493.

La cui soluzione è:

4) $W = W_0 e^{c_1 c_2 t}$

La conoscenza *W* cresce quindi esponenzialmente[30].

Il lettore con un *background* prevalentemente filosofico può essere comprensibilmente scettico sulla possibilità che tale semplice derivazione possa essere la *dimostrazione* di un fatto profondo e non banale come **LRT** – e, certamente, ci sono ottimi argomenti che si possono sollevare in questo senso. Tuttavia, è importante fin da subito evitare grossolane obiezioni: da una parte, i modelli più utili sono spesso frutto di enormi semplificazioni della realtà sottostante; dall'altra, bisogna considerare i tre tipi di argomento come parti diverse di un'unica visione: l'analisi concettuale spiega intuitivamente i fenomeni, il modello matematico astrae le variabili fondamentali e permette l'analisi dei dati empirici, i quali ne corroborano l'esattezza.

Ricostruendo razionalmente il percorso di Kurzweil possiamo dunque dire che la **LRT** è il risultato di questo processo di analisi:

a. L'osservazione che il ritmo evolutivo di diverse tecnologie ha natura esponenziale.
b. La formulazione di un semplice modello matematico che permette di spiegare i trend esponenziali di (a) e fare predizioni sul futuro (usando il modello e i dati storici).
c. L'elaborazione di un'analisi concettuale dell'evoluzione di una tecnologia e dei cambiamenti di paradigma che giustificano le variabili che compaiono in (b), spieghino l'adozione delle assunzioni fatte dal modello e forniscano un quadro complessivamente intuitivo dello sviluppo tecnologico nel suo complesso e della natura dei *feedback* positivi.

Tale ricostruzione ha sicuramente il pregio di separare in modo chiaro ed esplicito i vari punti dell'analisi di Kurzweil, permettendo non solo di valutare più facilmente le obiezioni, ma anche di apprezzarne l'indubbia originalità e lucidità.

[30] Vedi Kurzweil (2008), pp. 493-494 e la ricostruzione in Sandberg (2010), p. 5.

2.2 Critiche e repliche

Nonostante l'eccellente lavoro divulgativo e l'opera di convincimento che Kurzweil porta avanti da ormai più di un decennio, le critiche alla **LRT** non si sono ovviamente fatte attendere. È però importante (specialmente in un lavoro filosofico di questo tipo) distinguere, tra tutte le obiezioni fatte ai sostenitori della Singolarità, gli argomenti direttamente mirati verso **LRT** rispetto ad obiezioni più specifiche sulla visione globale (ad es., 'le macchine non potranno mai pensare come un essere umano'): in questa sezione ci occuperemo esclusivamente dei problemi sollevati da **LRT**, discutendo contro-argomenti specifici nelle apposite sezioni di filosofia della mente ed etica.

Ci sono fondamentalmente tre tipi di obiezioni che si possono sollevare verso **LRT**:

Os) L'obiezione *statistica*: non è vero che esistono i trend esponenziali.
Op) L'obiezione *popperiana*: il futuro non è necessariamente uguale al passato.
Om) L'obiezione di *Marty McFly*: predire la "forma" del futuro è estremamente complicato.

Affrontiamole una alla volta.

2.2.1 L'obiezione statistica

Per quanto riguarda (***Os***), l'obiezione principe è quella sollevata in più articoli da Theodore Modis[31]. Al di là di considerazioni squisitamente tecniche, la tesi di Modis è sicuramente interessante da un punto di vista puramente epistemologico. Nelle sue fasi iniziali, un trend esponenziale è molto simile (praticamente indistinguibile) da una curva logistica (la curva "a forma di *S*" incontrata in precedenza). La curva logistica è tipica di molti fenomeni biologici ed economici in cui, in qualche senso, esistono dei naturali limiti allo sviluppo imposti dalle risorse disponibili nell'ambiente: in questo senso, la critica di Modis può essere letta come una sfida a dimostrare che lo sviluppo tecnologico umano nel suo complesso (che, di fatto, ha accelerato fino ad ora) continuerà a farlo dato che è ugualmente possibile che invece ora incominci proprio a

[31] Vedi Modis (2006), Modis (2002).

rallentare[32] – tecnicamente parlando, Modis argomenta che già ora la curva logistica è una rappresentazione più accurata dei dati, ma la differenza è per sua stessa ammissione non molto alta. In altre parole, sia per Modis sia per Kurzweil viviamo in un'epoca speciale, ma mentre per il secondo siamo all'inizio di un ulteriore balzo in avanti, per il primo siamo al culmine dello sviluppo umano: da qui in poi, non possiamo che rallentare. Dal punto di vista strettamente epistemico, la sfida è quindi distinguere due scenari che appaiono indistinguibili all'uomo del nostro tempo, su base puramente numerica.

Difficile negare che lo scettico abbia sollevato qui un punto interessante. Tuttavia, il sostenitore della Singolarità può qui citare $\mathbf{SV}_2$ per rafforzare la sua posizione: infatti, dato che il prossimo grande paradigma dello sviluppo umano sarà segnato dall'esplosione di intelligenza, la risorsa più importante (almeno ad un primo livello) diventerà l'*informazione*; se è vero che la curva logistica è giustificata principalmente in casi di risorse limitate, un bene "immateriale" come l'informazione per sua natura non rischia di esaurirsi a causa di inerenti limiti ambientali – esiste probabilmente un limite teorico all'informazione che l'universo può contenere, ma è un limite così astronomico da rendere certamente possibile la Singolarità per come intesa da Kurzweil. Se quindi è vero che statisticamente una crescita logistica non può essere scartata sulla base dei dati storici, si può però sottolineare come le classiche assunzioni che sottostanno a quel modello di crescita falliscano per il bene più importante nelle previsioni sulla Singolarità[33].

Da un altro punto di vista, diversi commentatori[34] hanno preso di mira la ricostruzione complessiva della storia umana e il trend esponenziale evidenziato da Kurzweil nella serie di "eventi storici" che hanno portato dalla comparsa dell'uomo al sequenziamento del DNA[35]. In particolare, molte sono le accuse di "*cherry-picking*", ovvero di aver scelto accuratamente solo eventi nella storia umana che potessero comprovare **LRT** nel contesto più ampio possibile. Come evidenziato dal grafico dei cambi di paradigmi (Fig. 3), anche utilizzando quindici – indipendenti – fonti di "*eventi chiave*" lo sviluppo tecnologico appare esponenziale; inoltre altri recenti studi sui trend

[32] Vedi Modis (2002).
[33] Può essere utile anche un confronto con Sandberg (2010), p.3, dove si fa essenzialmente la stessa constatazione.
[34] Si veda ad es., *http://scienceblogs.com/pharyngula/2009/02/09/singularly-silly-singularity/.*
[35] Vedi Kurzweil (2008), p. 20.

di sviluppo nelle tecnologie collegate all'informazione[36] confermano l'osservazione generale di Kurzweil. La critica (per quanto sia tra le più popolari) appare lontana dal poter preoccupare davvero i sostenitori della Singolarità.

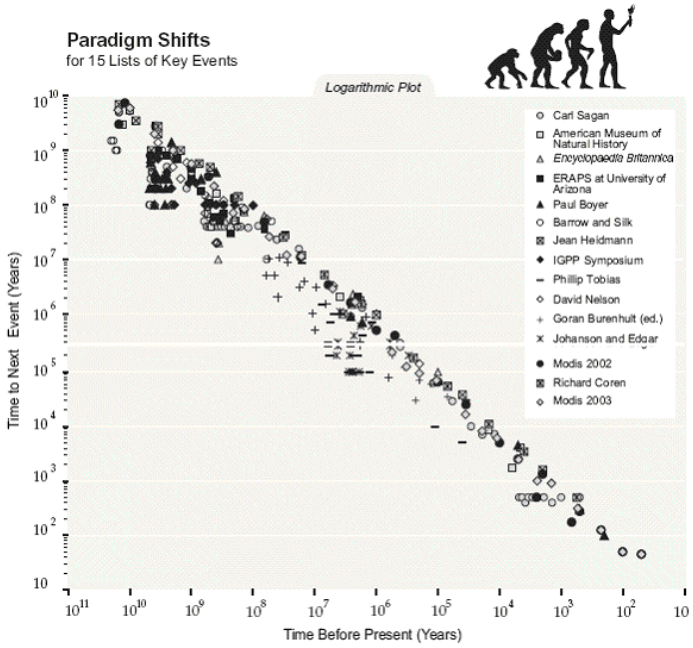

Fig. 3: Quindici concezioni dell'evoluzione con i cambiamenti di paradigma associati (Kurzweil (2008), p. 20).

2.2.2 L'obiezione popperiana

L'obiezione (***Op***) è una versione del problema dell'induzione[37] applicata al dibattito su **LRT**: il fatto che un certo fenomeno sia stato verificato *n* volte nel corso della storia non è motivo sufficiente per credere che si verificherà *n*+1 volte, soprattutto in un caso, come questo, dove non esistono chiare leggi fisiche a supporto della conclusione. Un'esilarante vignetta dell'*Economist*[38] dipinge una legge di Moore per i rasoi, sottolineando come il numero di lamette per rasoio stia crescendo ad un ritmo possibilmente esponenziale: se così fosse, fra qualche anno avremo rasoi da 15 lamette.

[36] Ad es., Nagy, Farmer, Tancik, Gonzales (2011) supportano l'idea di Kurzweil che l'accelerazione tecnologica sia super-esponenziale.
[37] Per una esposizione classica, si veda Popper (1963), Cap. 1.
[38] Si veda: *http://www.economist.com/node/5624861?story_id=5624861*.

L'idea dietro alla *reductio ad absurdum* è chiara: un semplice sguardo a un numero limitato di esempi non costituisce motivo per credere ad una continuazione futura del trend, soprattutto se altre considerazioni entrano in gioco (nel caso dei rasoi, la scarsa praticità di un rasoio a 15 lame). Nel caso della Singolarità – e quindi considerando lo sviluppo umano nel suo complesso – ci sono molte considerazioni extra numeriche per pensare che il trend non si mantenga: una guerra su scala mondiale con le armi di oggi (o, peggio, quelle del futuro) è il caso paradigmatico di evento che "resetta" l'orologio dello sviluppo umano.

Questo tipo di obiezione è ovviamente differente da (***Os***): mentre (***Os***) chiama direttamente in causa **LRT** e il lavoro di Kurzweil (sostenendo che esso sia basato su una errata interpretazione dei dati empirici), (***Op***) è compatibile con la verità di **LRT** per un periodo di tempo limitato della storia umana: infatti, quello che è in discussione non è tanto la validità di **LRT** fino *ad oggi*, quanto piuttosto la sua validità *domani*. Anche in questo caso, è indubbio che lo scettico disponga di una serie di buoni (e vividi) esempi che possano gettare ombra sulla portata di **LRT**. Se convincere del tutto lo scettico appare difficile, crediamo sia tuttavia possibile fare qualche interessante considerazione spostando leggermente l'attenzione da una domanda cui è quasi impossibile rispondere ('il trend continuerà?') a una domanda cui è semplice rispondere: scommettendo sulle previsioni fatte usando **LRT** a fine anni Novanta, avremmo vinto dei soldi?

Kurzweil (2010) è un recente e dettagliato report sulle predizioni contenute in Kurzweil (1999): ebbene, qual è il risultato di tali "scommesse"? Secondo Kurzweil, su 147 predizioni, 115 (78%) sono corrette alla fine del 2009, 12 (8%) sono "essenzialmente corrette": in totale, l'86% delle predizioni fatte utilizzando l'idea di uno sviluppo esponenziale della tecnologia sono risultate corrette – l'86% difficilmente convincerà lo scettico radicale, ma di certo anche per uno "scettico medio" è un risultato piuttosto stupefacente! Discorso chiuso? Purtroppo no: infatti, lo stesso giudizio di correttezza di una previsione è in qualche modo soggetto a trattativa. Ad esempio, Kurzweil profetizzava nel 1999 che in dieci anni "telefoni con tecnologia di traduzione" (ovvero: tu parli italiano, il tuo interlocutore sente in inglese) sarebbero stati 'comunemente usati', una funzionalità 'standard dei PC di qualsiasi persona'[39]; secondo

[39] Kurzweil (2010), p. 50.

Kurzweil, questa predizione nel 2009 era da giudicarsi "essenzialmente corretta". Al di là delle sottigliezze semantiche con cui Kurzweil tenta di difendere la tesi ('la predizione affermava che fossero di "uso comune", non presenti ovunque'[40]), è ragionevole affermare che i telefoni con tecnologia di traduzione non siano comunemente disponibili (altrimenti leggereste senza problemi questo testo in inglese). L'applicazione probabilmente migliore per la traduzione oggi disponibile, *Google Translate*, è evidentemente lontana – molto lontana! – da poter essere usata con affidabilità in una conversazione telefonica. Una valutazione di tutte le predizioni di Kurzweil (1999) esula dallo scopo di questa mappa filosofica, ma un punto generale (che tornerà più volte in seguito) vale la pena sottolinearlo immediatamente: tendenzialmente, le previsioni che concernono misure puramente *quantitative* (ad es. il costo in dollari di una determinata quantità di elaborazione) sono sorprendentemente accurate; viceversa, le previsioni che riguardano la *forma* che una particolare tecnologia ha preso (ad es., il *modo* di usare la tecnologia per comunicare) appaiono invece meno accurate (o comunque valutate con standard di "vaghezza" tali da renderle poco interessanti).

Infine, al critico di tipo (***Op***) rimane aperto un ulteriore spiraglio, quello della fondamentale incompletezza del modello matematico, il quale potrebbe adattarsi molto bene ai dati esistenti ma essere inadeguato al futuro: la semplicità del modello sarebbe dunque una ingannevole rappresentazione della realtà; quando altre variabili socio-economiche cominceranno ad interagire con V e W, il risultato non sarà più prevedibile dalla semplice formula proposta per **LRT**. Possiamo includere in questa macro-obiezione il lavoro pioneristico di Joseph Tainter, *The Collapse of Complex Society*, in cui la caduta di grandi imperi della storia viene analizzata attraverso la lente della complessità. Semplificando (ma non troppo), possiamo riassumere cosi l'idea portante di Tainter: le civiltà crescono aumentando di complessità; l'aumento porta indubbiamente benefici, ma ha anche un costo (in termini di energia, investimenti, etc.). L'osservazione chiave a questo punto è che la *complessità ha un ritorno marginale decrescente*, ovvero più una società è complessa, più è costoso renderla ancora più complessa: quando la complessità diventa insormontabile, la civiltà collassa e si disgrega in micro-civiltà più semplici. A sostegno della sua tesi, Tainter analizza

[40] Kurzweil (2010), p. 51.

diverse parti della società civile (agricoltura, gestione delle risorse, innovazione etc.), concludendo per ciascun caso che i dati esibiscono *un ritorno marginale decrescente*. Nel campo dell'innovazione, ad esempio, Tainter nota che 'se le spese di ricerca e sviluppo devono crescere del 4-5% l'anno per incrementare la produzione del 2%, questo trend non può continuare per sempre o arriverà il giorno in cui saremo tutti scienziati'[41]. Tainter non è comunque interamente pessimista riguardo alla società contemporanea:

> 'Quando un nuovo input viene introdotto in un sistema economico, sia esso una innovazione tecnica o un sussidio energetico, ha spesso potenzialmente la capacità di aumentare almeno temporaneamente la produttività marginale.'[42]

Ed è proprio attraverso questa ultima ammissione che possiamo ricondurre queste importanti riflessioni al nostro dibattito su **LRT**. Quello su cui Tainter e Kurzweil sono d'accordo è che prima o poi ogni tecnologia finirà per non essere più produttiva: tuttavia, mentre per Tainter (che guarda soprattutto al passato), lo sviluppo tecnologico diviene un rallentare l'inesorabile declino – il momento in cui l'energia e la creatività di una società non bastano più a sostenerne la complessità intrinseca –, Kurzweil pensa che la nuova era in arrivo, l'esplosione di intelligenza, segnerà un aumento cosi radicale in ogni singolo *input* da procrastinare indefinitamente il problema. Dato il diverso retroterra culturale e orizzonte temporale di analisi dei due ricercatori, non stupisce la parziale differenza di vedute, pur partendo da un approccio sostanzialmente simile – ovvero, il fatto di ricondurre l'intero sviluppo di una civiltà a un modello di produzione dell'*output* semplice, elegante e generale. Inoltre, anche grazie a questo confronto, risulta sempre più chiara l'importanza dell'esplosione di intelligenza nell'economia della visione generale prevista da Kurzweil: senza una vera Intelligenza Artificiale, Tainter potrebbe giustamente osservare che nessun altra innovazione, per quanto importante, possa cambiare radicalmente le carte in tavola per la nostra civiltà rispetto a quelle, fallite, del passato.

[41] Tainter (1990), p. 124.
[42] Tainter (1990), p. 124.

2.2.3 L'obiezione di Marty McFly

L'ultima obiezione, quella di "Marty McFly", è ben sintetizzata dalle parole del celebre linguista Steven Pinker:

> 'Non c'è la minima ragione per credere nella Singolarità in arrivo. Il fatto che si riesca a visualizzare un futuro con il potere dell'immaginazione non è una prova che sia probabile e nemmeno possibile. Considerate [...] il pendolarismo con il jet-pack, le città sotto l'acqua, gli edifici alti un miglio, le automobili a energia nucleare – tutte fantasie sul futuro che avevo da bambino e che non si sono mai avverate.'[43]

È innegabile che la nostra capacità di predire il futuro sia in alcuni campi drasticamente limitata (citando Yogi Berra, 'predire è molto difficile; soprattutto il futuro'); è altrettanto vero che anche se le meraviglie che il Pinker bambino si immaginava non sono qui, cose altrettanto incredibili (e cui nessuno pensava) sono diventate realtà grazie al progresso tecnologico: ad esempio, la più grande raccolta del sapere umano è gratuitamente disponibile online grazie a *Wikipedia*, il mio telefono cellulare riconosce facilmente qualsiasi motivo musicale recuperando a richiesta autore e titolo del brano (*Shazam* – e con *Musixmatch* me ne fornisce anche direttamente il testo) e, ancora più sorprendentemente, fMRI e *PET* permettono di osservare "*in vivo*" il funzionamento del cervello umano senza tagliare alcunché. In altre parole, rileggendo in modo caritatevole la provocazione di Pinker, l'obiezione a **LRT** non è tanto che il futuro sviluppo computazionale sia impossibile da prevedere di per sé, quanto piuttosto che è difficile prevedere l'impatto *preciso* che questo sviluppo avrà nelle nostre vite – un'osservazione che di fatto ben si sposa con la nostra precedente valutazione delle previsioni di Kurzweil. Se non era "difficile" prevedere che un'automobile di oggi potesse avere più potenza computazionale del *Saturn V* che portò l'uomo sulla Luna negli anni Sessanta, era praticamente impossibile immaginare che tale potenza computazionale avrebbe portato, ad esempio, al sequenziamento del DNA.

Se inquadriamo questo problema nel dibattito generale sulla Singolarità, la conclusione sembra essere non tanto che la Singolarità non ci sarà, ma che qualsiasi previsione su *cosa* accadrà di fatto all'accelerare della tecnologia appare futile. Questo

[43] Citato in *http://en.wikipedia.org/wiki/Technological_singularity*.

tipo di argomento è interessante perché ancora una volta è fondamentalmente indipendente dalle specifiche argomentazioni a sostegno di **LRT**: nulla, nel modello matematico proposto o nell'analisi dei cicli della tecnologia, permette di dire alcunché di sostanziale su come sarà il mondo vicino alla Singolarità; in particolare, le variabili *V* e *W* sono solamente *numeri*, la cui crescita esponenziale, se non sostanziata da considerazioni filosofiche forti, non ha alcuna rilevanza per il dibattito.

2.3 Possibilità epistemica, metafisica e realizzabilità pratica

Un equo bilancio degli argomenti, delle critiche e delle contro-risposte nel dibattito su **LRT**, ci porta ad ammettere che $\mathbf{SV}_1$ è quantomeno una possibilità *epistemica*: nulla di ciò che sappiamo sul mondo, allo stato attuale delle cose, ci fa pensare che sia impossibile che tale tesi sia vera. Se è vero che non abbiamo trovato conclusivi gli argomenti di Kurzweil, e che rimangono sicuramente da vagliare maggiormente alcune considerazioni statistiche e concettuali, è altrettanto vero che nessuna delle critiche sembra davvero scalfire l'immagine esponenziale del progresso tecnologico al cuore di $\mathbf{SV}_1$.

Il passaggio da possibilità epistemica a possibilità *metafisica* a realtà *empirica* è però ancora da compiere: da un lato, se la Singolarità passasse per controverse tesi filosofiche (come il funzionalismo in filosofia della mente) – e se queste si rivelassero *false* – avremmo motivo di pensare che, nonostante le apparenze, **LRT** non sia vera; dall'altro, anche se non ci fossero ostacoli *a priori* in tutta la visione proposta da Kurzweil, potrebbero comunque esserci, lungo il cammino, ostacoli di natura contingente che potrebbero rallentare, o fermare il processo. Distinguendo come sempre con attenzione i due tipi di passaggio (e lasciando la filosofia vera e propria al capitolo successivo), può essere utile dedicare la prossima sezione a tutta una serie di ostacoli empirici sul percorso verso la Singolarità.

2.3.1 Gli ostacoli sul cammino

Possiamo dividere in due grandi categorie gli eventi che potrebbero fermare la Singolarità: quelli *diretti* (ovvero tentativi esplicitamente volti a fermare lo sviluppo

tecnologico esponenziale) e quelli *indiretti* (ovvero eventi che non nascono con l'intento di danneggiare il progresso ma che finiscono per qualche motivo per rallentarlo o fermarlo indirettamente). Tra gli eventi diretti, c'è, ad esempio, la decisione dell'umanità di interrompere (o vietare con leggi/azioni militari/azioni politiche) lo sviluppo delle ricerche in IA: scoprire che la Singolarità porterebbe più costi che benefici (torneremo sulla bilancia dei pro e contro fra poco), può portarci a considerare razionale per l'umanità tentare di impedirla. Il problema con questo tipo di scenario è che lo sviluppo di IA (nel caso in cui ci trovassimo ad un passo dal primo vero computer intelligente) è analogo al dilemma del prigioniero: per qualsiasi Stato sulla terra che stia facendo ricerca nel campo, la strategia razionale è continuare a farla sia che gli altri Stati non smettano (altrimenti si rischierebbe di rimanere senza difese), sia che gli altri smettano (rimanendo cosi l'unico Stato ad averla). In altre parole, o l'umanità intera si accorda (ma come?) sulla cessazione delle attività legate alla IA, o anche solo l'esistenza di un gruppo di lavoro in uno Stato dissidente giustificherebbe il proseguo delle ricerche nel resto del mondo[44]. In altre parole ancora: se davvero ci trovassimo sull'orlo della Singolarità, non sembra ci sia alcuno schema di incentivi che, su larga scala, potrebbe mai razionalmente prevenire l'accadere del cambiamento di epoca.

Tra gli eventi indiretti c'è ogni sorta di catastrofe naturale (soluzioni "vintage" come un meteorite che colpisce la Terra) e sociale (soluzioni più moderne come rivoluzioni su larga scala, disastri biologici, etc.): se la probabilità del primo tipo di eventi è di fatto indipendente dal processo di sviluppo, non è così per i disastri ambientali e le rivoluzioni, la cui probabilità plausibilmente aumenta (o comunque non diminuisce) all'avvicinarsi della Singolarità. Se infatti lo sviluppo tecnologico i) continuerà ad aumentare il "*digital divide*" tra diverse parti del mondo e ii) produrrà tecnologie (come le nanotecnologie) il cui utilizzo sbagliato o imprudente potrebbe causare catastrofi, è facile apprezzare come la strada per la Singolarità sia lastricata di potenziali incidenti di percorso che rischiano di far precipitare di nuovo tutta l'umanità in un Medioevo tecnologico invece che in una nuova età dell'oro[45].

[44] L'escalation nucleare della Guerra Fredda seguiva né più ne meno lo stesso principio: il disarmo non conveniva a nessuno dato che non si poteva essere certi che l'avversario smantellasse il proprio arsenale. La differenza cruciale tra i due scenari è ovviamente il fatto che una bomba atomica, per quanto potente, non gode di intelligenza propria, a differenza di una mente artificiale.

[45] Alcuni di questi ostacoli ricompariranno nel capitolo sul futuro quando si discuterà della desiderabilità o meno degli scenari post-Singolarità.

2.3.2 Un primo bilancio

In conclusione, possiamo sottolineare che **LRT** non è l'unico modello astratto di Singolarità disponibile sul mercato: solo considerando il bel saggio di Sandberg (2008), infatti, ci sono una decina di modelli matematici che includono come possibilità $\mathbf{SV}_1$. Ovviamente, il fatto che si possano studiare modelli matematici in cui certe variabili mostrino un andamento esponenziale non è particolarmente originale né significativo di per sé: come abbiamo già avuto modo di evidenziare, un modello matematico è il risultato di una serie di assunti e definizioni; chiamare una variabile "conoscenza" rende il modello una *teoria della conoscenza* quanto ribattezzare "margherita" un cane lo renda un fiore. Quello che occorre sono dunque una serie di considerazioni concettuali che giustifichino le assunzioni fatte dal modello e colleghino le variabili che mostrano comportamenti di interesse alla realtà che il modello intende rappresentare. Da *questo* punto di vista, l'esistenza di un sempre più nutrito e variegato insieme di modelli è un ottimo segno per il sostenitore della Singolarità: invece che sulla sola **LRT**, si può dunque contare su una serie di idee e argomenti indipendenti – nel momento in cui un modello si rivelasse incompleto, oppure non fedele alla realtà, rimangono comunque molte altre considerazioni che supportano l'idea di una crescita esponenziale. Anche se, ancora una volta, l'evidenza è, nella migliore delle ipotesi, circostanziale e non conclusiva, abbiamo trovato un altro mattoncino da aggiungere sul piatto della bilancia a favore di $\mathbf{SV}_1$: complessivamente, crediamo che il caso a favore dell'accelerazione tecnologica sia ben supportato e che meriti grande attenzione.

Poiché dopo il primo *round* il sostenitore della Singolarità non solo è ancora in piedi, ma perfettamente in forza, dobbiamo dunque passare a vagliare la seconda parte del suo programma: anche concesso che la tecnologia abbia crescita esponenziale, cosa ci fa pensare che riusciremo ad accelerare l'intelligenza?

3. Intelligenza

"*Tutti sanno che una cosa è impossibile*
da realizzare, finché arriva uno sprovveduto
che non lo sa e la inventa."

(Albert Einstein)

3.1 Intelligenza e funzionalismo

Le conclusioni del capitolo precedente suggeriscono che la Singolarità sia uno scenario possibile, un potenziale punto di svolta nello sviluppo della civiltà umana, in un tempo relativamente vicino. Ovviamente, gran parte della forza di questa tesi deriva in modo strutturale dalla crescente "disponibilità" di intelligenza. In altre parole, per molti teorici della Singolarità, lo sviluppo di intelligenze non umane è condizione necessaria – anche se non ancora sufficiente – al raggiungimento dell'obiettivo: se avessimo la prova che un nuovo tipo di intelligenza artificiale (o ibrida) non possa essere realmente costruita, avremmo *ipso facto* un ottimo motivo (anche se non conclusivo) per dire che la Singolarità non potrebbe avvenire.

Come già notato in apertura, mentre l'idea stessa di accelerazione tecnologica ha in sé un po' di senso comune, l'idea di accelerazione di intelligenza è indubbiamente molto più "esotica". È difficile (per non dire impossibile) definire *intelligenza* in maniera esaustiva e non controversa; tuttavia, crediamo che una definizione popolare e intuitiva come quella di Howard Gardner possa inquadrare il problema in modo sufficientemente preciso per i nostri scopi:

> 'le capacità intellettive umane devono comprendere un insieme di abilità di *problem solving* che permettano a un individuo di risolvere genuini problemi (…) e devono anche comprendere la capacità di trovare o inventare nuovi problemi, gettando le basi per l'acquisizione di nuova conoscenza.'[46]

[46] Gardner (1993), p 64.

Se intelligenza è dunque approssimativamente uguale a 'risolvere problemi e trovarne di nuovi interessanti', è facile notare come, nel corso dei secoli, l'umanità nel suo complesso sia diventata più intelligente, spesso grazie all'interazione continua tra cose nel cervello e cose fuori di esso: l'invenzione della scrittura, i pallottolieri, i computer e tutti i vari *devices* spesso citati nel dibattito sulla Mente Estesa[47] hanno senza dubbio contribuito a migliorare le performance cognitive degli esseri umani e a permettere alle nuove generazioni di usufruire delle scoperte fatte da quelle vecchie. Tuttavia, è dubbio che i singoli individui oggi sul pianeta siano più intelligenti di qualche secolo fa; anche se fosse, tale miglioramento ha un tasso sicuramente troppo lento per meritare il nome di "accelerazione di intelligenza" e segnare dunque l'arrivo della Singolarità. Piuttosto la verità di $\mathbf{SV}_2$ (l'avvento di una nuova era per l'umanità caratterizzata dalla 'fusione di tecnologia e intelligenza umana'[48]) richiede un tipo completamente diverso di miglioramento delle capacità cognitive: carta e penna, monitor e *tablet* non sono sufficientemente veloci ed integrati con i nostri stati mentali per produrre un aumento *esponenziale* dell'intelligenza. Se questo è vero, appare quindi che qualsiasi tentativo di sostenere $\mathbf{SV}_2$ debba per forza implicare che l'intelligenza umana non sia qualcosa di esaurito, una volta per tutte, con la creazione dei cervelli biologici di *homo sapiens sapiens*: l'intelligenza è una proprietà che deve trascendere in qualche modo la dotazione che Madre Natura ci ha dato se vogliamo che possa migliorare esponenzialmente. In effetti, i sostenitori di $\mathbf{SV}_2$ sembrano confidare in una tesi *metafisica* nota come funzionalismo, ovvero l'idea che gli stati mentali siano realizzati nel cervello biologico ma possano essere replicati in altri sostrati a patto che alcune condizioni siano verificate.

È in questo punto del dibattito sulla Singolarità che tecnologi e futurologi incontrano la tradizionale filosofia della mente e delle scienze cognitive. Pertanto, lo scopo di questo capitolo è quello di tentare una contestualizzazione del fenomeno di accelerazione di intelligenza nel panorama scientifico e filosofico contemporaneo: attraverso la definizione dei concetti coinvolti e una breve ricostruzione storica sulla nascita e l'importanza del paradigma funzionalista faremo emergere le sottili connessioni tra metafisica, intelligenza artificiale e Singolarità.

[47] Il classico sull'argomento è Clark, Chalmers (1998). I temi legati alla Mente Estesa torneranno più volte nel resto del lavoro.
[48] Kurzweil (2008) p. 19.

3.1.1 Funzionalismo q.b.

Come prevedibile, le "ricette" proposte per l'accelerazione di intelligenza si basano tutte su un fatto fondamentale, ovvero che l'intelligenza sia in qualche modo una *tecnologia* e, come tale, possa essere esponenzialmente migliorata: se la Legge di Moore vale per i computer, perché non dovrebbe valere per gli essere umani?

Al di là del dibattito sulla Singolarità è in effetti semplice percepire come diversi i due seguenti "slogan": 'aumentiamo esponenzialmente l'altezza' vs 'aumentiamo esponenzialmente l'intelligenza'. Questo gap psicologico deriva fondamentalmente dalla diversa percezione che abbiamo delle due caratteristiche: mentre l'altezza è una proprietà prettamente biologica dell'essere umano, l'intelligenza ci sembra molto simile ad una proprietà "astratta" di un certo sistema. Se tale percezione è corretta, qualcosa di simile al funzionalismo deve essere vero – e se il funzionalismo è vero, allora diviene molto più ovvio equiparare l'intelligenza ad una tecnologia e da lì argomentare verso una possibile esplosione esponenziale.

In altre parole, se noi concepissimo la nostra mente come un prodotto esclusivamente biologico incorreremmo nei limiti tipici della nostra architettura (limite di consumo energetico, dimensioni fisse del cranio, limiti nella velocità degli impulsi nervosi, etc.) che renderebbero incredibilmente più complicato il raggiungimento dell'obiettivo. È quindi evidente fin da subito che sostenere $\mathbf{SV}_2$ di fronte a un cosiddetto "sciovinista del carbonio" sarebbe estremamente faticoso.

3.1.2 Cronache di un'ascesa

Cerchiamo a questo punto di entrare nel cuore concettuale del problema. Cominciamo con ciò che è più caro – e familiare – ad un filosofo, ovvero con alcune definizioni. In generale, definiremo *funzionalista* quella teoria ontologica degli *stati mentali* secondo cui qualcosa è uno stato mentale non in virtù di avere una determinata costituzione interna, ma piuttosto perché ricopre il ruolo causale appropriato in un dato sistema[49]. Il funzionalismo si contrappone tipicamente alle teorie dell'identità, secondo cui qualcosa

[49] Per una esposizione classica, si vedano gli argomenti in Turing (1950), Putnam (1967), Lewis (1980). Per un trattamento manualistico eccellente si veda Kim (1998), Cap. 3-4.

è uno stato mentale se è identico a un determinato stato fisico[50]: 'avere sete', 'credere che l'Inter vincerà il campionato', 'amare Gianni', sono tutti stati fisici[51]. Le teorie dell'identità sono considerate un primo passo per liberarsi del dualismo cartesiano (secondo cui gli stati mentali sono sostanze diverse da quelle presenti nel mondo fisico) e inquadrare il mentale in una immagine naturalistica del mondo: dato che il progresso della scienza sembra indicare una costante riduzione di fenomeni prima misteriosi a principi fisici, perché non potrebbe accadere lo stesso con gli stati mentali?

In letteratura l'argomento principale a sostegno del funzionalismo è noto con il nome di *realizzabilità multipla*[52]: l'idea di fondo è che intuitivamente ci aspetteremmo che cani, gatti e anche pipistrelli abbiano stati mentali, pur non condividendo quasi nessuna delle nostre strutture cerebrali. Tuttavia, se avere la credenza che 'c'è un predatore' è identico ad *avere le X-fibre* attivate, sembra proprio che un essere umano ed una pipistrello non possano essere in un identico stato mentale, dato che solo il primo possiede le fibre in questione. Ecco dunque che il funzionalismo arriva in soccorso e ci permette di superare i problemi della teoria dell'identità*:* poiché ciò che conta è il ruolo causale, le *U*-fibre umane e le *P*-fibre del pipistrello possono tranquillamente essere lo stesso stato mentale purché ricoprano nei due sistemi *ruoli funzionali analoghi.*

Definiremo dunque funzionalista qualsiasi teoria che implichi il bicondizionale:

FUN_{def}) $x \in S \leftrightarrow R_1(x,y)$ & $R_2(x,z)$& … $R_n(x,k)$

dove *S* è l'insieme degli stati mentali e *y*,*z*,*k* variano solo su membri di *S* (altri stati mentali) e su stati di input/output del sistema e $R_1 \ldots R_n$ sono relazioni funzionali.

L'idea che il pensiero sia una proprietà astratta di un sistema fisico non è in realtà un'invenzione degli ultimi decenni. Molti anni prima di Hilary Putnam ed Alan Turing, John Hobbes sosteneva la celebre 'pensare è calcolare':

> 'Quando una persona *ragiona*, non fa altro che concepire una somma totale risultante dall'*addizione* di parti o un resto derivante dalla *sottrazione* di una somma da un'altra […] in qualunque campo in cui c'è spazio per l'*addizione* e la *sottrazione*,

[50] In particolare ci riferiamo qui alle cosiddette teorie dell'identità dei tipi, dove è un tipo di stato mentale ad essere identico a un tipo di stato cerebrale.
[51] Vedi ad esempio Smart (1959).
[52] Vedi Putnam (1967).

c'è spazio anche per la *ragione* e dove non c'è spazio per le prime, la *ragione* non ha nulla da fare.'[53]

Tuttavia, tanti cambiamenti concettuali dovevano avvenire prima che il funzionalismo diventasse un paradigma di ricerca fruttuoso. Il punto centrale era, essenzialmente, capire cosa *davvero* volesse dire che 'pensare è calcolare'. Da una parte, la teoria della computazione e le nascenti scienze cognitive degli anni Cinquanta – linguistica ed Intelligenza Artificiale *in primis* – fornivano per la prima volta un linguaggio rigoroso in cui articolare l'idea; dall'altra, l'ontologia del funzionalismo era l'impostazione naturale per chiunque si proponesse come obiettivo ultimo di replicare il comportamento intelligente umano all'interno di un sistema artificiale.

Se dal punto di vista filosofico il funzionalismo nasce in un clima di superamento della teoria dell'identità, a livello scientifico il funzionalismo nasce negli anni della svolta cognitiva e della rivolta contro il comportamentismo, che vedeva la faccenda da un'angolatura decisamente differente:

'[…] naturalmente il comportamentista non nega l'esistenza degli stati mentali; preferisce semplicemente ignorarli. Il comportamentista "ignora" gli stati mentali nello stesso modo in cui il chimico ignora l'alchimia, l'astronomo l'astrologia, lo psicologo la telepatia e le manifestazioni parapsichiche. Il comportamentista non ha alcun interesse per gli stati mentali perché man mano che il fiume della scienza si allarga e diviene più profondo, questi vecchi concetti vengono risucchiati, per non ricomparire mai più.'[54]

Dove il comportamentismo vedeva una *black box* di cui si possono osservare solo input e output, il funzionalismo rende esplicita (per lo studio dell'intelligenza nel suo complesso) l'importanza dei singoli *passaggi* che da un dato input portano a quel particolare output. In estrema sintesi, per capire la mente bisogna *aprire* la scatola. Questo convincente cambiamento di prospettiva unito al grande successo delle prime

[53] Hobbes (2004), p. 34, corsivi miei.
[54] Watson (1920).

scienze cognitive[55] ha certamente contribuito a rendere l'ontologia funzionalista parte della pratica scientifica e, in parte, del nostro stesso senso comune.

3.1.3 Funzionalismo e Intelligenza Artificiale

Il funzionalismo, per motivi storici e naturali, ha un'amica di infanzia che è croce e delizia della sua esistenza: l'Intelligenza Artificiale. Abbiamo già incontrato più volte l'IA, confidando nella comprensione intuitiva del concetto senza mai indugiare in vere e proprie definizioni: parafrasando le parole di un suo pioniere, possiamo dire che l'IA è quella particolare disciplina che tenta di *riprodurre* in sistemi artificiali comportamenti che, se eseguiti da un essere umano, sarebbero giudicati intelligenti[56]. Il Sacro Graal dell'IA è dunque la costruzione di una macchina in grado di riprodurre tutte le facoltà cognitive degli esseri umani: percepire il mondo esterno, fare inferenze, provare emozioni, comunicare con altri sistemi intelligenti, manipolare l'ambiente esterno a proprio vantaggio, imparare dall'esperienza, etc.

Prima di proseguire nella riflessione, è bene ricordare che storicamente l'IA è un'affascinante storia di promesse mancate. Nata a metà degli anni Cinquanta tra grande clamore, ha fin da subito acceso la fantasia dei più illustri ricercatori del campo: più e più volte nei decenni scorsi abbiamo sentito dichiarare che la costruzione di un vero e proprio computer pensante fosse un traguardo "finalmente molto vicino". Tuttavia, come ben sappiamo, tali promesse non si sono *mai* avverate; all'interno dello stesso campo, si è passati da un tentativo titanico di riprodurre l'intelligenza umana ad una serie di sotto discipline, più o meno collegate tra loro, specializzate nella riproduzione di particolari "moduli" mentali: ad esempio, molti dei prodotti che oggi utilizziamo quotidianamente discendono in modo naturale dai risultati ottenuti dall'IA, sebbene l'etichetta di un tempo non venga più utilizzata nella sua accezione "omnicomprensiva" originale.

L'unione tra funzionalismo metafisico e IA produce la concezione computazional-rapprensentazionale della mente tipica delle scienze cognitive classiche: un sistema S_i è intelligente se, dati certi input, è in grado di elaborare le rappresentazioni mentali di tali

[55] Si pensi ad esempio all'impatto della linguistica generativa Chomskiana e ai primi risultati dell'IA, come il *Logic Theorist*. Vedi, ad es., Burattini, Cordeschi (2004), Cap. 1.
[56] Vedi Mc Carthy Minsky, Rochester, Shannon (1955).

input con computazioni che operano su tali rappresentazioni fino a produrre un certo output. Schematizzando,

(S_i) = INPUT → (rappresentazioni + computazioni) → OUTPUT

dove l'output corrisponde ad un certo comportamento osservabile. Tale impalcatura teorica (nel bene e nel male) ha caratterizzato la nascita e crescita delle scienze cognitive, spazzando via tutte le precedenti teorie della mente grazie alla capacità di rendere conto di comportamenti psicologici complessi e di fenomeni stratificati come l'apprendimento. Nelle parole di Fodor:

> 'vi sono fatti concernenti la mente che la teoria computazionale della mente spiega e che, altrimenti, non si potrebbero spiegare; e la sua tesi fondamentale – i processi intenzionali sono operazioni sintattiche definite su rappresentazioni mentali – può vantare una straordinaria eleganza.'[57]

Nei decenni successivi, l'amore tra ricercatori e filosofi e questo paradigma "straordinariamente elegante" era destinato a finire. Sempre Fodor commenta:

> 'Non avevo mai pensato che qualcuno potesse supporre che la teoria computazionale della mente contenesse quasi tutta la verità sulla cognizione; o addirittura che fosse prossima a rivelarci ogni cosa sul modo in cui funziona la mente.'[58]

Torneremo presto alle critiche rivolte al paradigma dall'*esterno*. Prima però occorre soffermarsi su una riflessione *interna:* anche ammesso che siano computazioni i rilevanti stati funzionali, in che senso sono rappresentazioni (ovvero: la manipolazione è simbolica o sub-simbolica)? Come prima approssimazione possiamo dire che il primo tipo di approccio prende il nome di *top-down*, e si lega al paradigma logicista dell'IA: nato sotto la stella della logica formale classica, questi approcci sostengono che il miglior modo per riprodurre la mente umana sia quello di *deingegnerizzare* la struttura di alto livello delle funzioni cognitive; il secondo è indicato con il termine *bottom-up* e

[57] Fodor (2001), p. 3.
[58] Fodor (2001), p. 3.

si concentra sulla modellizzazione non cognitiva della mente, "creando" intelligenza dall'emergere di certi pattern di elaborazione sub-simbolica,come quella che avviene tipicamente nelle reti neurali. Schematizzando:

$_{A}$TD) Approccio *top-down:*

$_{A}TD_1$) il livello di stimoli e risposte è quello macroscopico, cioè è direttamente accessibile alla nostra esperienza comune;

$_{A}TD_2$) la conoscenza è rappresentata in modo localizzato tramite simboli, ognuno con un significato autonomo e globale;

$_{A}TD_3$) gli effetti delle stimolazioni esterne consistono unicamente su processi computazionali che agiscono sui simboli.

$_{A}$BU) Approccio *bottom-up:*

$_{A}BU_1$) il livello di stimoli e risposte è quello microscopico, cioè in termini di caratteristiche elementari non direttamente accessibili alla nostra esperienza;

$_{A}BU_2$) la conoscenza è rappresentata in modo distribuito tramite le relazioni tra le microunità cognitive; un sistema molto usato è quello delle reti neurali;

$_{A}BU_3$) gli effetti delle stimolazioni esterne consistono unicamente in particolari processi computazionali che modificano le relazioni non simboliche tra gli elementi base del modello (ad. es., i nodi della rete neurale).

È chiaro che tale dibattito è interno alla visione *latu sensu* computazionale della mente[59]: il disaccordo tra le parti non si riferisce al valore computazionale degli stati funzionali, e nemmeno all'esistenza di rappresentazioni; tuttavia, dove l'approccio logicista vede manipolazione diretta di simboli, il connessionista vede significati emergere come regolarità statistiche da elementi privi di valenza simbolica. In particolare, entrambi gli approcci sono funzionalisti in quanto per entrambe le teorie di modellazione l'intelligenza non è altro che una proprietà astratta ed è il risultato del funzionamento di un certo sistema fisico. Tuttavia, le due tradizioni si sono rivelate quasi complementari, sia nei successi, sia nei fallimenti: se ($_{A}$TD) ha avuto il merito di riuscire fin da subito a modellare capacità astratte di alto livello (ad es., la deduzione

[59] Burattini, Cordeschi (2004), p. 98.

matematica), ha però fallito in modo eclatante nell'interazione con in mondo esterno[60]; viceversa, sebbene ($_{A}$BU) abbia prodotto interessanti sistemi fisici in grado di interagire con successo nell'ambiente, non ha mai avuto grande successo nel replicare le capacità di alto livello[61].

Al di fuori della comunità degli studiosi di IA, un numero crescente di filosofi e scienziati ha cominciato a dubitare della visione funzionalista computazionale della mente. In particolare, il motivo del contendere diviene squisitamente ontologico: mentre i sostenitori dell'IA *forte* rimangono ancorati ad una lettura metafisica del funzionalismo (i.e.: pensare è calcolare e gli stati mentali sono stati funzionali), i sostenitori dell'IA *debole* attribuiscono al calcolo una funzione puramente euristica (nella migliore delle ipotesi). In particolare, una riproduzione funzionalmente identica di un comportamento intelligente non è motivo sufficiente per ascrivere intelligenza reale al sistema, in analogia con quanto avviene nel tentativo di riprodurre e predire sui nostri computer la meteorologia: anche qualora si riuscisse a riprodurre un uragano in modo perfetto, i ricercatori presenti in sala non dovrebbero preoccuparsi di essere trascinati via poiché, di fatto, non si tratterebbe di un uragano *reale*, ma soltanto di una sua simulazione. Allo stesso modo, un sostenitore dell'IA debole pensa che quando un computer effettua una deduzione logica non stia *davvero* ragionando ma stia solo simulando un ragionamento.

Ovviamente, se il funzionalismo è vero, la falsità dell'IA debole discende in modo automatico: se fare una deduzione *significa* essere in una serie di stati funzionali, ogni qualvolta gli stati sono implementati, avremo una deduzione vera e propria – e *non* una simulazione della stessa.

3.1.4 Funzionalismo e Singolarità

A questo punto siamo pronti per affrontare nel dettaglio il legame tra funzionalismo e Singolarità: se, come abbiamo detto nelle pagine precedenti, la Singolarità presuppone un'accelerazione di intelligenza e quest'ultima a sua volta presuppone l'IA in qualche senso, allora potremmo dire, come prima approssimazione, che essere funzionalisti è condizione quantomeno necessaria per credere nell'avvento della Singolarità. In altre

[60] Ad es., vedi il classico Brooks (1990) per le critiche al paradigma logicista.
[61] Ad es., vedi le considerazioni in Bringsjord (in stampa).

parole, chiunque non accetti l'ontologia funzionalista avrebbe gioco facile nel dimostrare l'impossibilità della Singolarità, almeno seguendo il corso normale degli eventi (ad es., senza chiamare in causa l'avvento di Marziani in grado di aumentarci il Q.I.).

Prima di proseguire nella discussione, occorre però affrontare l'ultima batteria di argomenti esterni contro il paradigma funzionalista:

vsF$_1$) *Il problema del "mondo"*: il funzionalismo computazionale è un modello del mentale eccessivamente *astratto*.
vsF$_2$) *Il problema della "semantica"*: il funzionalismo non tiene conto del fatto che la mente è qualcosa di più della semplice manipolazione di simboli formali.

Innanzitutto appare ovvio che la critica espressa da (**vsF**$_1$) pesca a piene mani nel problema di modellazione della mente esposto nel paragrafo precedente e rimasto di fatto irrisolto. Tale critica dice fondamentalmente che, se guardiamo com'è fatto il mondo, l'analogia mente-software tanto cara al funzionalismo non sembra reggere: un sistema intelligente S_i è tale in virtù del fatto che "ha una mente", ma tale mente è, di fatto, *incorporata*, e mente e corpo sono immersi in un determinato ambiente. Poco male, dal momento che il modello computazionale può essere facilmente esteso fino ad incorporare altre parti di mondo: secondo la *embodied cognition*[62]: è un errore focalizzarsi solo sui processi "interni" all'organismo dato che mente, corpo, ambiente e azioni hanno pari dignità nell'elaborazione dei processi cognitivi. In questo modo, verrebbe a risolversi anche il problema dell'apprendimento: la cognizione incorporata potrebbe essere la soluzione teorica – e pratica – alla diatriba tra ($_A$TD) e ($_A$BU) in quanto parte dal presupposto che i pensieri siano il risultato della capacità dell'organismo di agire nell'ambiente; più precisamente, un determinato S_i, imparando a controllare i suoi movimenti e ad agire in un certo modo, sviluppa la consapevolezza delle sue proprie percezioni e abilità di interazione. Non solo: tali processi di basso livello sarebbero inoltre essenziali allo sviluppo di quelli di alto livello, come ad esempio, il linguaggio. Usiamo le parole di Di Francesco (2004) per chiarire ulteriormente il punto:

[62] Vedi ad es. le sei tesi della cognizione incorporata come presentate in Wilson (2002).

‘Avere esperienza di quello stato corporeo che chiamiamo emozione significa in questo quadro percepire certi mutamenti essenziali del proprio corpo (visceri, muscoli, respiro, eccetera), i quali avviano complesse interazioni tra il cervello e il corpo.’[63]

Per quanto decisamente assenti dal funzionalismo dei primi tempi, questo tipo di osservazioni si colloca in modo estremamente naturale in una visione computazionale della mente: occorre semplicemente che il funzionalista “duro e puro” ammetta che la cognizione umana abbia luogo in un sistema complesso in cui vari *pattern* interagiscono tra loro in modo differente dai *pattern* di un computer. In altre parole, occorre ricordare che il funzionalismo (come caratterizzato in (FUN_{def})) è una teoria ontologica sulla *natura* degli stati mentali, secondo la quale gli stati mentali *sono* stati funzionali; non c’è nulla nel funzionalismo che implichi un rifiuto di una elaborazione delle informazioni distribuita nel corpo: se l’evidenza empirica ci insegna che le relazioni funzionali corrette fanno parte di catene di stati funzionali che si estendono al di fuori del cervello (e, come vedremo, del corpo), il funzionalista ha tutto il diritto di includere questa evidenza nella teoria – a dire il vero, è il funzionalista, molto più del teorico dell’identità, ad avere una teoria così flessibile da accomodare senza problemi l’eterogeneità della cognizione umana. La teoria funzionalista, di per sé, non si impegna a fornire un elenco delle relazioni funzionali che “contano”: è compito delle scienze cognitive indicare i ruoli funzionali e cosa li realizza – l’importante è che la natura degli stati mentali rimanga *funzionale*, i.e. potenzialmente separabile dall’implementazione biologica. La diatriba tra scienza cognitiva classica e scienza cognitiva *nuova* (qui analizzata nella sua corrente *embodied*) è pertanto un falso dilemma filosofico: se di scontro si può parlare, è uno scontro puramente *empirico*. Da una parte, c’è una visione molto astratta della mente, retaggio della prima IA anni Sessanta, dall’altra c’è un modello più dinamico della cognizione, che ha fatto proprio il linguaggio dei sistemi dinamici e complessi e riscoperto l’importanza dei concetti cibernetici di *feedback* e retroazione[64]. Tuttavia, *anche* da un punto di vista empirico, non si vede (slogan sensazionalisti a parte) come sottolineare giustamente l’elaborazione che avviene nel

[63] Cfr. Di Francesco (2004), p. 119.
[64] Si veda Wiener (1948). Vedi anche la postfazione di Massimo Marraffa in Bechtel, Abrahamsen, Graham (2004).

corpo o l'apporto dell'ambiente alla cognizione squalifichi l'importanza di un'elaborazione centralizzata di alcune informazioni. In quest'ottica, la nuova scienza cognitiva può quindi essere pensata come un "allargamento" del funzionalismo computazionale, fino ad includere stati funzionali *non immediatamente* rappresentazionali: se la scienza cognitiva classica poneva l'accento *solo* su stati funzionali con un chiaro contenuto semantico (desideri, credenze, intenzioni, etc.), la nuova scienza cognitiva ha il merito di aver sollevato il problema dell'esistenza di stati che sono *funzionali* e *informazionali*, ma non rappresentazioni semantiche di alto livello. Riscoprire il ruolo delle emozioni, del corpo, dell'interazione ambientale è un modo per allargare la base delle computazioni necessarie per dotare un sistema fisico di intelligenza: le computazioni esplicite e coscienti su desideri e credenze sono solo una parte di ciò che ci rende intelligenti.

La conclusione provvisoria è dunque una sostanziale vittoria del funzionalismo sui suoi critici: nella sua accezione filosofica, il funzionalismo non è per nulla intaccato da (**vsF**$_1$); dal punto di vista empirico, il funzionalismo nella sua variante computazionale non è falsificato dalla nuova evidenza, la quale tuttavia impone un ripensamento adeguato del tipo di modellazione necessaria per catturare *tutti* gli stati informazionali che concorrono allo sviluppo dell'intelligenza. Per quanto riguarda il legame con la Singolarità e l'Intelligenza Artificiale, la riproduzione dell'intelligenza è legata alla possibilità di principio di separare gli stati mentali dalla base biologica e alla possibilità di computare le rilevanti relazioni funzionali: anche in questo ambito, il fatto che l'elaborazione delle informazioni avvenga sotto forma di rappresentazioni complesse o con modelli ibridi è concettualmente ininfluente.

Passiamo ora a (**vsF**$_2$), la più famosa tra le critiche mosse contro il funzionalismo:

vsF$_2$) *il problema della "semantica"*: il funzionalismo non tiene conto del fatto che la mente è qualcosa di più della semplice manipolazione di simboli formali.

La questione è molto semplice: un computer è per definizione un dispositivo di manipolazione di simboli formali, compie cioè una serie di operazioni puramente sintattiche, *ma* noi sappiamo dalla nostra esperienza personale che la mente è qualcosa

di più della manipolazione di simboli formali, *quindi* il funzionalismo computazionale è sbagliato. Nelle parole di John Searle, il più agguerrito sostenitore di questa obiezione:

> 'La mente non potrebbe essere soltanto un programma per computer perché i simboli formali di un programma non sono di per se stessi sufficienti a garantire la presenza del contenuto semantico che si trova nelle menti reali.'[65]

Più formalmente, l'argomento complessivo di Searle è dunque il seguente[66]:

Sea$_1$) I programmi sono sintattici.
Sea$_2$) La mente ha una semantica.
Sea$_3$) La sintassi non è da sola sufficiente per generare la semantica.
C) I programmi non sono sufficienti per fare una mente.

A supporto di (**Sea**$_3$), in un famoso articolo del 1982, intitolato *The Myth of the Computer,* Searle presenta per la prima volta l'ormai leggendario *argomento della stanza cinese*[67]. L'esperimento mentale è semplice e può essere così riassunto: immaginate di essere chiusi a chiave in una stanza e di avere a disposizione delle scatole con molti simboli in cinese (che, ovviamente, non conoscete) e un manuale in cui sono presenti le regole di combinazione dei simboli. All'esterno della stanza ci sono degli omini che vi passano nella stanza dei fogli scritti in cinese e aspettano che ne mandiate dei vostri indietro. Le scatole con i simboli sono il *database* del nostro sistema; il manuale con le regole corrisponde al programma; i fogli scritti in cinese che gli omini vi danno dall'esterno sono gli *input* del sistema, i risultati che fornite voi corrispondono agli *output.* La questione è chiara: sebbene io produca *output* e riesca a farlo anche velocemente – perché ho imparato molto bene ed in fretta il manuale che mi hanno dato – io di fatto non posso dire di *capire* il cinese. Nelle parole di Searle:

> 'se io non riesco a capire il cinese solamente sulla base dell'implementazione di un programma per computer per comprendere il cinese, allora non lo può fare nemmeno

[65] Cfr. Searle (1998), p. 8.
[66] Cfr. Searle (1998), p. 9.
[67] Originariamente l'argomento non aveva un nome specifico. Venne battezzato l'*argomento della stanza cinese* solo in seguito.

nessun altro computer digitale solamente su quella base, perché nessun computer digitale possiede qualcosa che io non ho.'[68]

Ritornando alle "etichette" introdotte in precedenza, è facile notare come Searle non discuta la possibilità di usare un calcolatore per effettuare simulazioni di qualsiasi tipo (IA debole), ma attacca l'idea dell'equivalenza tra simulazione e realtà (IA forte).

L'argomento, intuitivamente convincente, è stato però criticato da molti punti di vista. Innanzitutto, è importante notare che lo *stesso* ragionamento, applicato agli esseri umani produce risultati paradossali: consideriamo come stanza il nostro cervello, come omini i nostri neuroni. La critica di Searle in questa versione diventa dunque che i neuroni non capiscono niente, quindi *non* esiste comprensione – ma allora anche il cervello umano non capisce niente! Questa semplice *reductio* ricorda la risposta di Dennett e Hofstadter in Dennett, Hofstadter (1981): nello scenario immaginato da Searle, la comprensione non va attribuita ai meccanismi dentro la stanza, ma è una proprietà globale della stanza stessa; se è vero che *noi* non capiamo niente, la stanza è invece intelligente (solo che non ce ne possiamo accorgere dall'interno, un po' come i nostri neuroni non possono rappresentarsi l'intelligenza *globale* del nostro cervello). Rivisitando l'argomento di Searle su sintassi e semantica, Chalmers (1996) offre un altro tipo di *reductio*:

Sea*$_1$) Le ricette sono sintattiche.
Sea*$_2$) Le torte sono croccanti.
Sea*$_3$) La sintassi non è da sola sufficiente a generare la "croccantezza".
C*) Le ricette non sono sufficienti per fare una torta.

Infine, possiamo considerare un ultimo tipo di obiezione all'argomento di Searle, basata sulla contestazione della plausibilità della stanza cinese (Block (2002)); l'idea di fondo è che Searle stia "barando" con le nostre intuizioni, chiedendoci di immaginare qualcosa di familiare (essere umani che svolgono semplici compiti meccanici) e poi portando questa situazione all'estremo (esseri umani che usano database per tradurre in modo effettivo da una lingua ad un'altra): è in questo passaggio dal familiare

[68] Cfr. Searle (1998), p. 8.

all'estremo che si nasconderebbe una mossa ingannevole – ovvero l'impossibilità di tradurre da una lingua all'altra con una semplice tabella di conversione dei simboli: qualsiasi tabella siffatta sarebbe così grande da non poter essere memorizzata nell'intero universo. In altre parole, l'unico modo per simulare intelligenza è con un insieme di regole *intelligenti*: l'argomento di Searle non funziona perché ci invita erroneamente a pensare che produrre un comportamento che sembra intelligente sia possibile con un insieme di procedure totalmente stupide (che, nel caso della traduzione di lingue umane, sarebbero così inefficienti da essere praticamente impossibili da realizzare nell'universo).

Un'altra critica nello spirito di (**vsF**$_2$) è quella formulata dal fisico Roger Penrose, che chiama in suo aiuto i risultati del teorema di Gödel per mostrare che la mente non può essere un sistema formale. L'argomento è il seguente[69]:

Pen$_1$) L'IA "forte" ci dice che la mente è un sistema formale.

Pen$_2$) Un sistema formale "scopre" nuove verità manipolando tramite le regole di dimostrazione gli assiomi di base (l'unico modo che un sistema formale ha per produrre conoscenza è "dimostrare").

Pen$_3$) Kurt Gödel ha dimostrati in Gödel (1931) che all'interno di un sistema formale è possibile costruire un enunciato *G* tale che esso non è dimostrabile (né lo è la sua negazione) all'interno del sistema stesso.

Pen$_4$) La mente umana riconosce l'enunciato *G* come vero, anche se non può dimostrarlo.

Pen$_5$) *Quindi* la mente umana *non* è un sistema formale – e il progetto dell'IA forte è insensato.

(**Pen**$_1$) si limita a descrive ciò che, nell'essenza, significa la metafora mente-software; (**Pen**$_2$) racconta, in parole più semplici, il seguente concetto: un sistema formale apprende la verità di un enunciato deducendolo formalmente da un insieme di assiomi assunti come veri; infine, (**Pen**$_3$) ci ricorda lo scomodo risultato gödeliano per cui un sistema formale sufficientemente potente non può dedurre l'enunciato di Gödel *G* né la sua negazione, e, tuttavia, noi "riconosciamo come vero" *G*, in quanto afferma

[69] Penrose (2000).

di se stesso di non essere dimostrabile. In altre parole, un sistema formale non può stabilire la verità o falsità del suo enunciato di Gödel *G* poiché non può dedurla in nessun modo: dato che un programma di IA è un sistema formale sufficientemente potente, *allora* il programma non potrà riconoscere la verità/falsità del suo enunciato *G*. Ma, poiché noi "riconosciamo" la verità di *G*, da questo Penrose conclude che noi non possiamo essere programmi di IA – e la mente *non* è un software.

Sebbene la strategia di Penrose sia affascinante, sembra filosoficamente poco calzante. È sbagliato – o per lo meno incompleto – affermare che la nostra mente riconosce la verità di *G*. Ciò che la nostra mente riesce a fare è riconoscere la verità del *condizionale* 'se il sistema è consistente allora *G* è vero'. Ovviamente, non c'è nessuna garanzia che la mente umana sappia riconoscere la consistenza di qualsiasi sistema formale, ad esempio quello che rappresenta la nostra mente: dunque, *nessuno* può concludere che *G* sia vero in senso assoluto. Per chiarire il punto, supponiamo di doverci trovare di fronte al sogno dei programmatori di IA, ovvero ad un sistema formale effettivamente in grado di riprodurre la cognizione: chi ci garantisce che saremmo in grado di riconoscere come vero l'enunciato gödeliano *G* di un sistema così complesso? E se noi non siamo in grado di farlo, ciò vuol dire che non siamo poi tanto diversi dal quel sistema formale. Infine, si può obiettare che resta una questione *empirica* e non teorica, se la mente umana sia consistente oppure no: come noto, in sistemi *paraconsistenti G* è dimostrabile[70]. Astraendo dall'argomento, la visione di Penrose è una visione fondamentalmente non computazionale del cervello e del mondo fisico: per Penrose esistono sistemi fisici (gli esseri umani) che risolvono problemi cognitivi in modo "non computazionale". Tuttavia, è difficile inquadrare questa ontologia in una visione coerente: innanzitutto, i cervelli umani potrebbero risolvere problemi in modo diverso dai computer (così come le reti neurali risolvono problemi in modo diverso da sistemi deduttivi), senza che questo implichi che risolvano *tipi* di problemi diversi – in questo caso, la descrizione computazionale sarebbe differente (ad es., connessionismo vs. approccio deduttivo oppure computazione classica vs. computazione quantistica), ma sarebbe pur sempre una descrizione *computazionale*, sottoposta ai limiti di tutte le computazioni. Se invece Penrose vuole sostenere che i cervelli risolvono problemi non computabili (come l'argomento goedeliano sembra fare

[70] Per un approccio analitico a contraddizioni, gödelizzazioni e logiche paraconsistenti, cfr. Berto (2006).

intendere), possiamo concludere che gli argomenti presentati non sono sufficienti a supportare una tesi così forte. Non solo è infatti dibattuto che il cervello mostri comportamenti non computabili, ma è, più in generale, discusso che esista *in natura* qualcosa che non sia computabile: all'opposto dello spettro metafisico in cui si colloca lo studioso inglese, esistono infatti metafisiche *totalmente* computazionali, in cui non solo il cervello, ma tutto il mondo fisico è ridotto a computazione[71]. In questa ottica, è l'informazione la variabile fisica più fondamentale[72] e l'evoluzione dell'universo è analoga all'evoluzione di un mondo digitale che calcola, ad ogni istante t, il proprio stato per $t + 1$.

Finora siamo stati bravi a smarcarci dai principali attacchi sferrati da filosofi, fisici, neuroscienziati e nuovi scienziati cognitivi al funzionalismo, ma la guerra non è ancora finita. L'ultima questione, forse la più spinosa in assoluto, rischia di mettere a dura prova il nostro nucleo di tesi pro-Singolarità: tale critica è nota in letteratura come:

vsF$_3$) *il problema della "coscienza"*: il funzionalismo non riesce a rendere conto della dimensione soggettiva/qualitativa della coscienza.

Il problema è discusso come problema dei *qualia*, recentemente dibattuto con ingegno e notevoli argomenti da Chalmers (1996). L'argomentazione di Chalmers è in realtà un argomento K.O. ad ampio spettro, mirato non solo al funzionalismo, ma a tutte le ontologie fisicaliste nella filosofia della mente contemporanea. L'idea si basa sul rapporto tra il concetto di sopravvenienza e riducibilità; in particolare, Chalmers argomenta che *A*-proprietà sono riducibili a *B*-proprietà *se e solo se* le *A*-proprietà sopravvengono logicamente sulle *B*-proprietà. Cosa significa in concreto? Prendiamo il caso delle proprietà biologiche e delle proprietà fisiche; quando ci chiediamo se le prime si riducono alle seconde, dobbiamo chiederci: due mondi possibili che hanno identiche proprietà fisiche possono avere *diverse* proprietà biologiche? Se in un mondo un vombato si sta riproducendo, un mondo con identica conformazione atomica conterrà ovviamente un vombato che si sta riproducendo: in altre parole, una volta fissate le proprietà fisiche di un mondo, le proprietà biologiche sono automaticamente

[71] Vedi ad esempio Zuse (1982), Fredkin (1993), Wolfram (2002).
[72] Cfr. ad esempio la prospettiva *It from bit*, presentata nel pioneristico Wheeler (1990) e commentata, tra gli altri, da Chalmers (1996). Per una raccolta di saggi recente sull'argomento, si veda Zenil (2013).

stabilite – è *solo* per questo che possiamo parlare di riduzione possibile: se infatti le proprietà biologiche potessero variare in modo indipendente, la fisica non potrebbe da sola spiegarle, perché non potrebbe giustificare tale variazione. Quando parliamo degli aspetti qualitativi dell'esperienza, la situazione sembra però diversa: due mondi fisicamente identici sembrano poter avere *qualia* diversi (il famoso argomento degli zombie è volto proprio a dimostrare questo, o l'esempio di Kripke su dolore/solletico e *C-fibre*[73]), così come due sistemi fisici funzionalmente identici sembrano poter provare stati qualitativi *completamente* diversi – quindi, non c'è modo di ridurre i *qualia* (stati intrinseci, non relazionali) a stati funzionali (estrinseci, relazionali) o proprietà fisiche: il funzionalismo è dunque un'ontologia del mentale *incompleta*, poiché tralascia gli aspetti qualitativi della coscienza (la "dolorosità" del dolore o la "rossità" del rosso, etc.).

Dal punto di vista teorico, (**vsF**$_3$) è sicuramente un argomento molto importante: nonostante le repliche di materialisti di tutto il mondo, l'intuizione fondamentale di Chalmers sembra rimanere intatta. Dal punto di vista della Singolarità, ci possiamo però chiedere quale sia il reale effetto di (**vsF**$_3$) con un breve esperimento mentale. Immaginiamo, dunque, di voler creare un *cyborg* con lo scopo di farlo diventare il nostro nuovo migliore amico (d'ora in avanti JT$^+$). Salvando un po' di "estetica" – dobbiamo pur sempre portarlo in giro senza vergognarcene troppo – cominceremo con il costruirgli un corpo bionico molto simile al nostro: diversi giorni saranno dedicati alla riproduzione dei vari sistemi di senso; il mio migliore amico deve avere, almeno ad un primo livello, le *mie* stesse percezioni e non quelle del mio gatto o del mio iPhone. Superato il primo passo, il nostro JT$^+$ non sarebbe lui se non avesse una coscienza vera e propria; per raggiungere questo traguardo abbiamo davanti diverse strade e prenderemo quella che ci sembra più funzionale all'obiettivo: immaginiamo di fare un elenco di tutte le abilità che il nostro vecchio migliore amico JT possiede. Dato che noi conosciamo bene JT, non avremo molte difficoltà nello stilare un elenco molto preciso di *tutte* le sue funzioni cognitive a noi tanto care e supponendo di essere realmente capaci di scrivere il software che contenga tutto ciò – anche JT dovrebbe tentare di riprodurre una MR$^+$ –, cominceremo a pensare al sistema più intelligente per impiantarglielo in testa. Alla fine dei vari interventi JT$^+$, addormentato e pronto per il

[73] Vedi Kripke (1980).

mondo reale, si sveglia e si comporta *esattamente* come il suo originale: possiamo ipotizzare, ad esempio, che appena ci vede cominci a raccontarci dei suoi infiniti problemi amorosi. Non solo: ad un certo punto potrebbe addirittura esprimerci il suo desiderio per una tavoletta di cioccolato o una lattina di coca-cola e spiegarci esattamente quello che prova nello svuotarci il frigorifero, per la perdita della sua bella, etc. A questo punto, ci sembra proprio a tutti gli effetti il nostro migliore amico: dato che i suoi stati funzionali rispecchiano quelli del nostro cervello, quale motivo abbiamo per pensare che i corrispondenti stati qualitativi non siano istanziati? È cruciale qui interpretare in modo corretto le conclusioni dell'argomento di Chalmers: Chalmers *non* sostiene che sia impossibile costruire un robot senziente, sostiene semplicemente che essere senziente non significa essere in un certo stato funzionale. Ma se esiste una correlazione tra stati funzionali e stati qualitativi, JT^{+}, avendo i "giusti" stati funzionali, avrà i "giusti" correlati qualitativi delle sue esperienze. D'altra parte, se l'istanziazione dei *qualia* non fosse garantita in qualche modo dalle nostre leggi di natura, lo stesso rapporto tra neuroni e coscienza non sarebbe più possibile oggetto di indagine scientifica: i *qualia* potrebbero infatti variare in modo casuale da essere umano a essere umano[74].

Nessuno degli argomenti contro il funzionalismo sembra dunque sancirne la disfatta: in particolare, anche nell'ipotesi in cui i *qualia* non fossero riducibili funzionalmente rimarrebbe salvo il principio secondo cui due sistemi funzionalmente identici hanno una vita mentale identica. Ricordiamo che la nostra valutazione del funzionalismo è in questa sede legata alla possibilità o meno della Singolarità ed è proprio da questo punto di vista che è importantissimo sottolineare come una variante ancora più debole del "funzionalismo senza *qualia*" sia probabilmente sufficiente. Supponiamo che non solo i *qualia* ma anche altri stati mentali non siano riducibili funzionalmente: ciò implicherebbe il fallimento *totale* della riproduzione della mente umana ma non la possibilità di costruire sistemi intelligenti infinitamente utili allo sviluppo tecnologico.

Consideriamo le tre caratteristiche che la mente umana possiede in un'ottica di progresso: apprendimento, capacità di ragionare e capacità di comunicare. Anche assumendo che l'IA forte si sopravvaluti, non sembra possibile che queste tre capacità chiave non siano riproducibili: gli algoritmi di apprendimento sono alcuni tra i più

[74] Su questo "Principio di Invarianza", si veda Chalmers (1996), pp. 248-250.

grandi successi dell'IA degli ultimi anni, la capacità deduttiva è già oggi comparabile a quella degli esseri umani e l'intero paradigma della linguista generativa si basa sul fatto che la comunicazione discenda dalla computazione su numero semplice di regole e di simboli. Alla prima categoria appartengono oggetti utilissimi come i filtri *anti spam* e tutti i software specializzati nel riconoscimento vocale. Ad oggi, è possibile non solo insegnare ad un computer a riconoscere la nostra voce ma anche a riprodurla in un altro linguaggio simulando intonazione e prosodia[75]. Al secondo caso appartengono tutti i sistemi di ragionamento automatico: per fare un esempio caro ai filosofi, Edward Zalta e Paul Oppenheimer[76] hanno scoperto con un sistema di dimostrazione automatico una semplificazione della prova originale dell'argomento di Sant'Anselmo. Nel terzo e ultimo caso sono inclusi tutti i tentativi di comprensione del linguaggio naturale (*Watson*, *Venexia* e *Wolfram Alpha*, sono solo alcuni tra gli esempi più importanti) che si sforzano di dare risposte più o meno sofisticate alle domande poste dall'utente. Anche supponendo che il funzionalismo sia (quasi) falso per la maggioranza degli stati mentali e che l'IA debole sia vera, la possibilità di una Singolarità sembrerebbe – all'interno di questo scenario – solo *posticipata*: di fatto, un sistema che simula la predizione di un terremoto è ugualmente utile all'uomo rispetto ad un sistema che lo predice realmente. Come vedremo in seguito, esiste la possibilità concreta – anche se meno eccitante dell'alternativa – che la Singolarità tecnologica venga raggiunta con sistemi che un teorico dell'IA o un filosofo della mente difficilmente considererebbe *intelligenti*.

3.2 Migliorare, costruire e scaricare una mente

Come accennato nel Capitolo 1, ci sono fondamentalmente due approcci principali all'accelerazione dell'intelligenza: da una parte, migliorare il materiale di partenza biologico dell'uomo, dall'altra costruire *ex novo* una mente artificiale. Nel mondo della Singolarità si discute infine di un terzo, "ibrido" progetto: il *mind uploading*, ovvero l'idea di scaricare il contenuto della mente – come se fosse un file in una chiavetta USB – su supporto digitale e utilizzare una versione accelerata rispetto al cervello biologico per aumentare le prestazioni cognitive.

[75] Si veda ad es. la dimostrazione di Rick Rashid di *Microsoft Research* nel Novembre 2012 (il video è disponibile online all'indirizzo *http://www.youtube.com/watch?feature=player_detailpage&v=Nu-nlQqFCKg#t=450s*).
[76] Oppenheimer, Zalta (1991).

In questa sezione analizzeremo in dettaglio tutti e tre questi approcci.

3.2.1 Migliorare una mente

Come accennato più volte nei paragrafi precedenti, l'accelerazione di intelligenza potrebbe non avere – una volta sviluppatasi in modo evidente – la forma del nostro JT^{+} ma potrebbe banalmente rientrare sotto la categoria del miglioramento dell'intelligenza umana come attualmente è. Non solo: una coesistenza dei due tipi di accelerazione è ovviamente possibile, dato che, potremmo prima diventare più intelligenti *noi* e poi essere in grado di costruire tanti JT^{+}. Il dibattito sulle varie forme di potenziamento è già molto sviluppato in letteratura come sottoprodotto del potenziamento di qualità fisiche e genetiche (eugenetica, *doping*, etc.). Seguendo Sandberg (2011), in questa sezione discuteremo brevemente le principali forme di potenziamento già disponibili, lasciando all'ultimo capitolo le implicazioni per individui e società di tali strumenti.

Innanzitutto, quando parliamo di "potenziamento" nel contesto della Singolarità, parliamo di cose non convenzionali: escludiamo tutta una serie di attività potenzianti già presenti da secoli nel nostro patrimonio culturale come l'educazione, un ambiente stimolante, la lettura, etc. Ciò a cui facciamo riferimento, al contrario, possiede fondamentalmente tre caratteristiche:

Novità. Il potenziamento non deve avere forme di regolamentazione già in atto all'interno della società; parliamo quindi di tecnologie così nuove e innovative da non poter essere facilmente inquadrate nelle esistenti categorie istituzionali.

Miglioramento. Le tecnologie potenzianti devono avere un rapporto tra costi e benefici più vantaggioso rispetto a quelle convenzionali (una pillola in grado di aumentare le tue capacità cognitive è "meglio" che studiare sui libri per vent'anni).

Possibilità. Le tecnologie potenzianti devono aprire tutta una nuova serie di possibilità (e problemi correlati) all'interno della società in cui viviamo.

Sempre seguendo Sandberg (2011), cominciamo la nostra panoramica:

- *Droghe*. Sostanze stimolanti, come nicotina e caffeina, sono usate da molto tempo per migliorare la cognizione umana: nel primo caso, abbiamo un notevole miglioramento nell'attenzione e nella memoria a breve termine; la caffeina invece riduce drasticamente la stanchezza sia mentale sia fisica[77]. Un altro fattore decisivo sembra essere legato all'alimentazione: certi tipi di diete e l'assunzione di alcune forme di integratori influenzano la nostra cognizione; in particolare, oggi sappiamo che per un funzionamento ottimale il cervello necessita di una buona dose di glucosio: di conseguenza, ad una maggiore disponibilità di glucosio durante il compimento di compiti particolarmente difficili corrispondono migliori performance[78].

- *Stimolazione magnetica trans cranica.* La TMS può aumentare o diminuire in modo evidente l'eccitabilità della corteccia cambiandone i livelli di plasticità. Si è inoltre dimostrato che tale tecnica può avere effetti benefici notevoli nello svolgimento di alcuni compiti motori tra cui, l'apprendimento motorio[79], la coordinazione[80], la memoria di lavoro[81], la classificazione e il consolidamento dei ricordi durante il sonno[82], etc. Purtroppo, la TMS, ha due grandi svantaggi. Il primo è che non si conoscono ancora gli effetti collaterali a lungo termine, il secondo è che le differenze tra cervelli diversi rendono costoso – a livello di tempo – trovare il punto esatto in cui la stimolazione ha efficacia.

- *Modificazione genetica*. Per quanto riguarda la memoria in particolare, esistono modificazioni genetiche sperimentate sui topi che si sono dimostrate efficaci nell'acquisizione e nel mantenimento di nuovi ricordi. Poiché la struttura cellulare della memoria sembra essersi conservata piuttosto intatta lungo l'evoluzione è plausibile pensare che ci siano molti margini di miglioramento nelle tecniche genetiche sulla memoria[83]. Tuttavia, se parliamo di intelligenza nel suo complesso, gli studi attuali suggeriscono che ci sia un gran numero di variazioni genetiche responsabili per

[77] Rusted *et al*. (2005), Tieges *et al.* (2004).
[78] Fox *et al.*(1988), Sunram-Lea *et al*. (2002)
[79] Nitsche *et al.* (2003)
[80] Antal *et al,* (2004)
[81] Fregni *et al.* (2005),
[82] Marshall *et al*. (2004)
[83] Vedi Bailey *et al.* (1996).

l'individuo, ma ciascuna lo è per una piccola frazione della differenza di performance tra due individui[84].

- *Modificazione prenatale*. La somministrazione di supplementi[85] ai ratti ha migliorato di gran lunga le loro performance a livello di cambiamenti nello sviluppo neurale[86]. Inoltre, sembra dimostrato che una specifica integrazione della dieta materna alla fine della gravidanza sia efficace anche per gli esseri umani[87]. Tuttavia, un problema empirico è legato al "buon" senso prevalente nella nostra comunità: attualmente, nonostante questi strumenti siano facilmente a disposizione, al momento le raccomandazioni mediche alle madri sono più dirette ad evitare danni che a potenziare il figlio.

- *Impianti cyborg*. Negli ultimi decenni c'è stato un importante sviluppo nell'interfaccia cervello/computer. In particolare, è oramai possibile per un essere umano paralizzato controllare il mouse con un solo elettrodo impiantato nel cervello[88]. Tuttavia, la maggior parte di questi studi si sviluppano per sopperire a deficit funzionali in soggetti non sani: un essere umano perfettamente sano, non vorrebbe utilizzare tale tipo di strumento, ma plausibilmente preferirebbe comandi vocali, tattili, etc.

- *Mente estesa*. Diverse tipologie di hardware esterni fanno parte dell'uso comune da molto tempo: dai semplici "carta e penna", alla nostra calcolatrice, fino ai personal computer di ultima generazione. Clark, Chalmers (1998) commenta questa "estensione" nel mondo sottolineando il fatto che è la mente stessa ad estendere i propri confini al di fuori del cervello, fino ad incorporare naturalmente questi strumenti. Dal punto di vista del potenziamento cognitivo il trend interessante è riuscire a rendere i pezzi di mondo sempre "meno estranei" all'utente che ne fa uso.

– *Intelligenza collettiva*. La maggior parte della cognizione umana è *attualmente* distribuita in menti separate. Tuttavia, questo sistema di distribuzione potrebbe essere

[84] Craig, Plomin (2006).
[85] Il supplemento in questione è la *colina*, altrimenti nota come vitamina J.
[86] Mellott *et al*. (2004).
[87] Helland *et al.* (2003).
[88] Kennedy, Bakay (1998).

migliorato con sistemi di collaborazione collettiva. Alcuni famosi esempi, tra tutti: *Wikipedia*, *Linux* e i cosiddetti mercati della predizione (in cui gli individui possono scommettere sulla probabilità o meno che un certo evento si verifichi). Può essere interessante notare come tutti questi sistemi consentano procedure di correzione dell'errore automatiche ed eguaglino (quando non superano) le performance di singoli esperti nel campo[89].

- *Nuovi sensi*. Esistono veri e proprio tentativi di *incorporare* nuovi sensi. Recentemente, ad esempio, si sono fatti vari esperimenti con la cosiddetta *sensibilità magnetica*: una volta inserito un piccolo magnete sotto la pelle delle dita di un individuo, il soggetto è in grado di percepire campi magnetici esistenti grazie al loro effetto immediato sul magnete incorporato[90]. Anche se lontano dall'essere un reale esempio di potenziamento cognitivo, tale sviluppo dimostra in modo chiaro la possibilità di sviluppare sensi *completamente* diversi da quelli con cui siamo abituati a convivere.

Nonostante le possibili ricadute pratiche di tutte queste tecnologie e le domande filosofiche ed etiche sollevano, appare chiaro che sostenere $\mathbf{SV}_2$ sulla base *solamente* di questo miglioramento di intelligenza non è del tutto convincente: la maggior parte di queste tecnologie di potenziamento, infatti, sfrutta l'hardware biologico esistente in modo superiore a quello "naturale", ma lo trascende solo in parte. Concettualmente, il potenziamento attraverso l'ibridazione con hardware artificiali è già una porta aperta ad una sorta di funzionalismo: è naturale dunque, per un sostenitore dell'*enhancement*, affiancare a queste tecnologie i risultati dell'Intelligenza Artificiale vera e propria.

Migliorare la mente è utile, probabilmente necessario: ma è solo costruendone una che $\mathbf{SV}_2$ può diventare realtà.

[89] Vedi Giles (2005), Raymond (2001) per la robustezza all'errore e Hanson *et al.* (2003) per l'accuratezza predittiva.
[90] Laratt (2004).

3.2.2 Costruire una mente

SV$_2$ collega lo sviluppo esponenziale con il particolare momento storico in cui ci troviamo a vivere: secondo **SV**$_2$, fra poco avverrà una *rottura tecnologica* significativa, tale da proiettarci in una nuova era – la cosiddetta "esplosione di intelligenza". Il tema è noto fin dagli anni Sessanta grazie alle intuizioni di Irving John Good, che nel 1965 profetizzava che all'arrivo di una macchina più intelligente dell'uomo, 'poiché costruire macchine è un'attività intellettuale, una macchina ultraintelligente potrebbe disegnare macchine ancora migliori'[91]. Curiosamente, l'argomento è anche uno dei pochi passati al vaglio della comunità filosofica, grazie all'interesse per il tema di David Chalmers (che ha dedicato alla Singolarità il lungo e articolato Chalmers (2010)). Possiamo ricostruire l'argomento per **SV**$_2$ nel seguente modo:

P$_1$) È possibile costruire un sistema più intelligente di qualsiasi essere umano.
P$_2$) Un sistema di intelligenza *L* può (eventualmente insieme ad altri sistemi di intelligenza *L* o inferiore) costruire un sistema di intelligenza *L*+1 – e così via.
P$_3$) In qualche decina di anni sapremo costruire un sistema più intelligente di qualsiasi essere umano.
C) In qualche decina di anni raggiungeremo la Singolarità.

Come evidente, l'argomento tocca molteplici punti caldi del dibattito sulla natura della mente; a differenza di altri argomenti correlati però, l'argomento di **SV$_2$** ha una portata decisamente "pratica", con conseguenze potenzialmente incredibili sulla vita di ogni abitante del pianeta. Esaminiamo dunque, premessa per premessa, le possibili obiezioni.

Ad una prima, distratta analisi, **P**$_1$ potrebbe sembrare una "semplice" riproposizione delle classiche tesi dell'IA forte. A dire il vero, tale visione "distratta" è sicuramente aiutata dal fatto che, come abbiamo visto, la maggior parte (se non la totalità) dei sostenitori della Singolarità sia *anche* entusiasta sostenitrice dell'IA forte. Tuttavia la forza di **P**$_1$ è di essere in realtà compatibile con una qualsiasi ontologia del mentale che non precluda la possibilità di creare sistemi "estremamente utili" all'uomo. Per capire meglio questo punto, consideriamo nuovamente la posizione "reazionaria" di John

[91] Vedi Good (1965).

Searle, il quale sostiene che un computer rappresenterà *al massimo* una simulazione del pensiero umano ma non possiederà mai davvero stati mentali: in questa ottica, non importa quanto veloci ed efficienti diventeranno le macchine a risolvere problemi complessi, esse non saranno *mai*, letteralmente, esseri pensanti, in quanto il pensiero è un prodotto squisitamente biologico. Anche assumendo questa prospettiva, possiamo riscrivere l'argomento come un'esplosione di "utilità ed efficienza" invece che di "intelligenza", ma la portata pratica del ragionamento rimane pressoché invariata: supponendo di avere a disposizione un "computer medico" che fa diagnosi perfette in qualche millisecondo, il fatto che stia "davvero pensando" o solo "simulando intelligenza" non ha alcuna rilevanza pratica; sia in un caso, sia nell'altro una tale invenzione garantirebbe da sola il guadagno di molti anni nell'aspettativa di vita umana.

In quest'ottica, per negare $\mathbf{P}_1$ bisognerebbe sostenere una tesi alquanto strana, ovvero l'impossibilità teorica per l'uomo di costruire sistemi (artificiali, biologici o ibridi) che lo superino nello svolgimento di tutte le attività cognitive "interessanti" – tesi che, per attività limitate come fare di conto e giocare a scacchi, sappiamo già essere falsa. Non solo: anche posizioni ultra-conservative (ad es., 'un computer non saprà mai comporre una sinfonia') sono perfettamente compatibili con $\mathbf{P}_1$, fintanto che l'attività che si ritiene "unica" non copre le aree fondamentali per il progresso scientifico e tecnologico – guarda caso le aree per cui sembra intuitivamente più semplice che un computer faccia i primi progressi significativi.

$\mathbf{P}_2$ costituisce il passo ricorsivo dell'argomento, la garanzia che, in un certo senso, la prima macchina ultraintelligente sarà l'unica macchina ultraintelligente fatta dagli uomini, essendo le successive costruite da generazioni sempre più intelligenti di altre macchine – come dice un adagio famoso tra i sostenitori della Singolarità, la risposta più appropriata a 'Ci saranno mai macchine intelligenti quanto l'uomo?' sarebbe 'Per pochissimo tempo'. Se accettiamo $\mathbf{P}_1$, un modo per negare $\mathbf{P}_2$ è sostenere che c'è un *naturale limite massimo* allo sviluppo dell'intelligenza (e che questo limite è ragionevolmente vicino): macchine sempre più intelligenti satureranno presto questo spazio astratto di capacità cognitive, arrivando nel giro di qualche generazione al "massimo livello di intelligenza possibile" e quindi precludendo una crescita adatta alla Singolarità[92]. Tuttavia, sembra difficile trovare ragioni indipendenti per sostenere

[92] Vedi anche i commenti in Chalmers (2010), pp. 19-20.

questa ipotesi: qualsiasi sia la definizione di intelligenza che si desidera adottare, non sembrano esserci motivi per pensare di non poter estendere l'intelligenza dell'uomo di svariati ordini di grandezza (ad es. più memoria, più velocità elaborativa, maggiore capacità inferenziale, minor numero di *bias* cognitivi, sensi aumentati etc.). Nelle parole di Chalmers (2010), $\mathbf{P}_2$ dipende dalla costruzione di una IA con "metodi estendibili", ovvero con metodologie tali da poter essere applicate più volte aumentandone la portata[93].

Alcune metodologie discusse nel mondo della Singolarità (e che menzioneremo più avanti), come l'emulazione diretta del cervello, non sono buoni candidati: se emulassimo un cervello umano in un computer, potremmo guadagnare sicuramente nella velocità di esecuzione dei processi, ma non migliorare le altre qualità oltre un certo livello (oltre ad un certo punto, la velocità di pensiero non migliora più significativamente le performance cognitive). Altri metodi appaiono invece molto più estendibili: se la prima IA sarà costruita grazie alla programmazione diretta, è molto probabile che sviluppi successivi possano essere ottenuti dalla ottimizzazione e dal miglioramento del codice sorgente. Un potenziale controargomento[94] è che, anche se alcuni metodi sono estendibili, possono risultare marginalmente sempre meno efficaci nel costruire nuove intelligenze: ad es., la prima IA potrebbe essere il 10% più intelligente di noi ma solo il 5% più brava nel costruire macchine pensanti – se questo è vero, in poche generazioni un miglioramento di intelligenza non porterà alcun miglioramento nella capacità di costruire sistemi intelligenti. Tuttavia, anche assumendo questa prospettiva conservatrice, è probabile che un aumento ragionevole, anche se non illimitato, nelle capacità intellettive – accoppiato a migliori prestazioni computazionali – sia sufficiente per generare almeno in parte una esplosione di intelligenza le cui ricadute sulla nostra vita siano tangibili.

Infine, valutare $\mathbf{P}_3$ ci riporta alle considerazioni fatte in precedenza sulla capacità di prevedere la forma dello sviluppo tecnologico. In particolare, lo scettico potrebbe sostenere che non abbiamo alcuna idea del punto in cui siamo nello sviluppo di un software adatto a dare vita ad una vera intelligenza; non solo, lo sviluppo tecnologico e i bisogni umani sono così poco prevedibili che potrebbe addirittura essere che i prossimi "sistemi intelligenti" abbiano un'intelligenza ed una utilità completamente *diversa* da

[93] Vedi Chalmers (2010), pp. 12-14.
[94] Vedi Chalmers (2010), p. 27.

quello che ci aspettiamo. Guardando con oggettività alla storia dell'Intelligenza Artificiale sicuramente una lezione appresa è che molte cose che apparivano difficili si sono rivelate semplici – come battere il campione del mondo di scacchi – mentre, viceversa, cose che apparivano semplici si sono rivelate pressoché impossibili – come tradurre automaticamente da una lingua all'altra. Non solo, se la storia recente insegna qualcosa, in molti campi è stato un approccio totalmente "*non* intelligente" a dominare il mercato fino ad ora: i successi nella traduzione automatica di *Google Translate* non sono infatti dovuti ad un avanzamento nella IA in senso stretto, quanto piuttosto a raffinate tecniche statistiche di *machine learning* e ad enormi volumi di dati[95]. D'altra parte, esistono anche casi di spettacolare successo in cui è più difficile capire quanta intelligenza sia davvero coinvolta. Il *DARPA Challenge* è un concorso a premi organizzato per la prima volta nel 2004 dalla Difesa Americana: la sfida è costruire un veicolo autonomo che attraversi da solo i 240 km del deserto del Mojave (California). Nel primo anno, il risultato migliore è stato di 11,78 km percorsi, ma solo un anno dopo 22 su 23 auto hanno superato quel risultato e ben cinque hanno raggiunto il traguardo! La tecnologia usata da quei veicoli è stata ulteriormente migliorata negli ultimi anni: oggi (Febbraio 2013) tre stati americani (Nevada, Florida, California) permettono la circolazione nelle strade urbane delle macchine senza guidatore di *Google* (Sebastian Thrun, leader del progetto a *Google*, era il responsabile del team di Stanford che vinse il DARPA Challenge nel 2005): complessivamente, sono più di 300.000 le miglia percorse dalle *Google Car* senza incidenti. Altri due casi di successo che meritano una sicura menzione sono *Watson*, il super computer IBM che ha battuto gli esseri umani nel gioco a quiz statunitense *Jeopardy*, e *Siri*, l'assistente virtuale di Apple che interpreta e risponde alle domande degli utenti utilizzando il motore computazionale di *WolframAlpha*. Forse più che per le altre premesse, ci troviamo per $\mathbf{P}_3$ a non poter convincere lo scettico – che ha, dalla sua, ottimi esempi – ma non siamo d'altra parte evidentemente sconfessati dall'evidenza – avendo dalla nostra parte sia casi pratici sia argomenti teorici a supporto della fattibilità del progetto.

Mettendo insieme argomenti a favore e contro ciascuna premessa, un primo bilancio equilibrato appare affermare che, sebbene non ci sia motivo per dare l'esplosione di

[95] Semplificando, il sistema traduce il sintagma inglese *E* nel sintagma italiano *I* utilizzando le frequenze relative a *E* ed *I* presenti in documenti con traduzione certificata, come i testi dell'ONU.

intelligenza per scontata, tuttavia non esistono motivi *a priori* per ritenerla impossibile. Se questo è vero, anche se assegnassimo una probabilità soggettiva molto bassa al suo verificarsi (diciamo una su diecimila) la portata di tale evento è tale da giustificare fin d'ora la massima attenzione possibile. Nella parole di Chalmers:

> '[l'avvento di una IA] avrà un impatto enorme sul mondo. Per cui, anche se la probabilità di una Singolarità è bassa, dobbiamo pensare seriamente alla forma che potrà avere.'[96]

3.2.3 Scaricare una mente

Per concludere questa panoramica, occorre parlare della tecnica decisamente più esotica e meno nota a un'*audience* di filosofi: il *mind uploading* (d'ora in avanti M*up*). In poche parole, l'idea del M*up* è quella di scaricare il contenuto della mente su un supporto digitale ed utilizzare dunque una simulazione computazionale per "far girare" la mente scaricata all'interno del computer: tra la mia mente e una canzone mp3 non ci sarebbe dunque nessuna differenza di utilizzo e portabilità. Esistono principalmente due tecniche di M*up*: il M*up* forte, e il M*up* debole. Per il primo, la chiave è partire dal raccogliere digitalmente tutte le informazioni necessarie a riprodurre una mente: che sia attraverso *nanorobot* che si legano ai neuroni, o attraverso scansioni di tecniche simili alla fMRI, l'idea è che quello che dobbiamo scaricare da un determinato cervello sono i parametri biologici sottostanti al funzionamento cognitivo. Una volta ottenuta questa informazione, possiamo facilmente copiarla in un computer, il quale, attraverso un modello del funzionamento sinaptico, può simulare l'evoluzione del sistema (e dunque, plausibilmente, riprodurre l'intelligenza in un contesto in cui la velocità di trasmissione degli impulsi non è più limitata dalla biologia)[97].

La versione debole del M*up* parte da premesse diverse: il modello computazionale della mente non prende, come parametri, una data configurazione neuronale, ma piuttosto una serie di "vissuti" della mente in questione. L'algoritmo dovrebbe infatti riprodurre una mente e una personalità a partire da oggetti macroscopici, come le e-

[96] Si veda Chalmers (2010), p. 29.
[97] Si veda Bostrom, Sandberg (2010).

mail, le foto delle vacanze, i ricordi di amici etc. Non serve conoscere come il cervello di JT implementa una mente: possiamo riprodurre JT (in una versione molto fedele) mappando su un modello computazionali i suoi gusti, i suoi ricordi, le sue preferenze, i suoi pattern comportamentali.

È ovviamente difficile valutare questo paradigma di ricerca in assenza di fatti più concreti: tuttavia, la "Bibbia" dell'*Institute for the Future of Humanity* sull'argomento è un comprensivo e ragionato tentativo di valutare tutte le tecnologie coinvolte nel M*up* e discuterne criticamente assunzioni scientifiche e ragionevoli prospettive sul futuro. Va detto che i tentativi esistenti di usare conoscenza "biologica" per riprodurre direttamente intelligenza sono per ora largamente infruttuosi: le mappature esistenti di sistemi nervosi di organismi semplici (come la drosofila) hanno reso evidente negli ultimi anni l'inadeguatezza di una visione ingenua sulla simulazione. Conoscere il network alla perfezione non è sufficiente a produrre la complessità di comportamento che si osserva nell'organismo – la semplice mappatura non garantisce alcuna comprensione dei meccanismi di base coinvolti. Più che una prova dell'impossibilità di realizzare il M*up,* questi primi esperimenti sono una prova di quante incognite ancora rimangano da risolvere. È comunque importante citare qui il *Blue Brain Project* del Politecnico di Losanna e tutto il progetto del *Connectone*, che unisce la neurobiologia alla teoria dei network: anche se in fase ancora embrionale, l'esistenza di questi progetti e l'interesse della comunità scientifica dimostrano quantomeno come sia sempre più diffusa la credenza che la comprensione e riproduzione a basso livello del cervello umano sia, con finanziamenti e tecniche adeguate, finalmente alla portata dell'umanità.

Per quanto riguarda il M*up* nella sua versione più soft, la valutazione critica è ancora meno facile, non esistendo un generale consenso (né una discussione dettagliata e condivisa) sugli assunti modellistici di base. Preliminarmente, può essere utile ricordare che l'idea che la mente e la personalità di un individuo siano il risultato del settaggio di un numero di parametri piuttosto limitato è da diverso tempo parte della psicologia cognitiva: la famosa *Teoria dei Cinque Tratti della Personalità*[98] è infatti un approccio descrittivo piuttosto rispettato nel campo. Secondo la teoria, sarebbero cinque i parametri da settare per ottenere tutte le possibili personalità umane: nell'ottica M*up* questo vorrebbe dire mappare e-mail, ricordi etc. su un modello di soli cinque parametri

[98] Vedi Backstrom, Larsson, Maddux (2009) e McRae, Costa (2008).

per ottenere una copia digitale della personalità di un soggetto. Tuttavia, è molto prematura qualsiasi forma di entusiasmo: la teoria dei cinque tratti è infatti puramente descrittiva (cioè una generalizzazione statistica sulla base di estesi test di personalità), ma non ha alcun valore predittivo né, soprattutto, alcun modello computazionale sottostante che ne giustifichi la struttura e ne permetta l'implementazione.

3.3 Natura e cultura all'avvento della Singolarità

In chiusura del capitolo sull'intelligenza, può essere utile riprendere in modo organico i riferimenti fatti nel testo alla teoria della Mente Estesa, più volte evocata nel corso di questo lavoro. Andy Clark e David Chalmers hanno infatti argomentato in modo convincente che una conseguenza diretta del funzionalismo è l'estensione della mente ad oggetti al di fuori del cranio[99]. Più in dettaglio:

i) La mente è costituita da stati mentali;
ii) se uno stato mentale è in parte ambientale, allora non è più possibile considerarlo dentro la testa;
iii) se uno stato mentale non è nella testa, allora la mente si estende nel mondo;
iv) esistono alcuni stati mentali non dentro la testa (alcune credenze e desideri);
dunque per (ii), (iii) e *modus ponens*, possiamo concludere che
v) la mente si estende nel mondo.

Per sostenere (iii) Clark e Chalmers ci chiedono di immaginare la seguente situazione:

Il caso di Otto e Inga. Inga, una giovane studentessa di filosofia, viene a sapere da un amico che c'è una mostra a Palazzo Reale e che tale museo si trova a Milano in Piazza Duomo. Inga richiama alla memoria dov'è Piazza Duomo e si reca alla mostra. A questo punto sembra chiaro che:

[99] Vedi Clark, Chalmers (1998).

I_a) Inga *crede* che il museo sia in Piazza Duomo.
I_b) La credenza in (I_a) era presente anche prima che Inga consultasse la sua memoria.

Otto, giovane ospite di Ville Turro, soffre di Alzheimer precoce e così porta sempre con sé un taccuino in cui scrive tutte le informazioni che via via acquisisce durante il giorno; tutte le volte che ha bisogno di ripescarle lo sfoglia. Anche lui vuole andare alla mostra perciò apre il suo taccuino e cerca l'indirizzo del Museo, legge che è in Piazza Duomo e ci va.

O_a) Otto *crede* che il museo sia in Piazza Duomo.
O_b) La credenza in (O_a) era presente anche prima che Otto consultasse il taccuino.

Se (I_a) e (I_b) sono vere e indiscutibili, e così pare, allora sono altrettanto vere e indiscutibili (O_a) e ($O_{b)}$ e ciò implica che si possono avere stati mentali *spread into the world*. Il taccuino di Otto è quindi da considerarsi funzionalmente identico a una memoria biologica; e poiché, come sappiamo, per il funzionalista "funzionalmente identico" equivale a identico *tout court*, abbiamo trovato un genuino caso di mente sparsa nel mondo.

Parleremo di nuovo di Mente Estesa (in relazione all'Etica) nell'ultimo capitolo. Quello che vorremmo sottolineare qui è che, *se* la tesi di Clark e Chalmers vale per Otto, per i nostri *smart phones* e PC, *a fortiori* l'estensione della mente nel mondo che avverrà eventualmente con la Singolarità sarà (manco a dirlo) *esponenzialmente* più evidente: i confini tra biologia/dotazione di partenza/produzione autonoma di conoscenza e strumenti/apprendimento/integrazione di informazioni si fanno così labili da diventare impercettibili. Immaginiamo, solo per un istante, una mente perfettamente digitale connessa con Internet e altre menti digitali: dove inizia e finisce la sua conoscenza?

4. Futuro

"*Studia il passato se vuoi prevedere il futuro*."

(Confucio)

4.1 Potenziamento, semi-immortalità e *mind uploading*

Nel secondo e terzo capitolo di questo lavoro abbiamo argomentato che la Singolarità sia una possibilità reale e concreta per lo sviluppo umano nel suo complesso. Nonostante alcune incognite restino ancora da chiarire e alcune difficoltà rimangano da risolvere, sembra tutt'altro che impossibile l'impatto rivoluzionario che tale sviluppo tecnologico esponenziale potrebbe avere sulle nostre vite. Ovviamente, che qualcosa sia possibile non la rende di fatto *desiderabile*. Ad esempio, potremmo accorgerci che la Singolarità Tecnologica comporti conseguenze tanto sgradevoli da volerne impedire *attivamente* l'avvenimento. Viceversa, potremmo trovarci a concludere – dopo una approfondita analisi – che i benefici superino di gran lunga i danni collaterali e quindi, decidere, di investire un numero ancora maggiore di risorse per promuoverne l'arrivo.

In questo capitolo ci occuperemo dunque di valutare la desiderabilità dello scenario proposto e, in particolare, ci cimenteremo nel difficile compito di *prevedere* gli effetti della Singolarità, valutando l'impatto che questo processo avrà per gli individui e la società nel suo insieme.

4.1.1 Potenziamento

Il dibattito sulle conseguenze del potenziamento umano è senza dubbio quello più sviluppato e approfondito in letteratura. Il motivo fondamentale è che, tra tutti gli "scenari", l'*enhancement* è il più facile da comprendere e da prevedere, dato che le tecnologie potenzianti sono *già* parte del nostro immaginario collettivo e, in parte, della moderna pratica scientifica. La desiderabilità su larga scala dell'avvento di tecnologie

potenzianti, sia a livello fisico, sia a livello cognitivo, sembra ovvia e viene bene espressa da un semplice concetto: chi non vorrebbe, in fondo, essere (più) intelligente? D'altra parte, ci sono delle altrettanto ovvie preoccupazioni. Una prima questione può essere ben riassunta facendo un'analogia con i miglioratori di performance sportive, il cosiddetto *doping*: quello che possiamo dire è che l'utilizzo di sostanze artificiali in contesti competitivi crea una sorta di "corsa alle armi" tra tutti quelli che le utilizzano. In altre parole, dato che qualcuno all'interno di uno specifico contesto utilizza un potenziatore, gli altri attorno a lui hanno un interesse ad utilizzarne uno di un tipo ancora più forte. Il secondo problema, difficilmente risolvibile, è quello dell'equità: se è vero che ognuno di noi può essere *potenziato* – essere a t_1 più forte, più veloce, più intelligente, etc. di come è al tempo t – l'alterazione della distribuzione naturale delle caratteristiche umane rischia di produrre un "divario artificiale" tra chi usufruisce degli strumenti di *enhancement* e chi no.

Per quanto riguarda l'analogia con il *doping*, una risposta semplice è a disposizione in quanto il potenziamento cognitivo è infatti solo superficialmente simile al doping: mentre il secondo è un bene *posizionale* – il suo valore dipende solo dal fatto che io posso usufruire dei potenziatori artificiali e gli altri no – il primo non lo è affatto – il motivo ovvio è che conviene a tutti essere più intelligenti, qualsiasi sia il caso. Essere "più intelligenti" ha un valore *intrinseco* che è indipendente dal fatto che JT o MR siano più o meno intelligenti degli altri. Inoltre, è bene sottolineare che, mentre la differenza tra una gara di 100 ciclisti normali e 100 ciclisti dopati avrebbe conseguenze sociali trascurabili (certo, 100 dopati faranno tempi assoluti migliori rispetto ai 100 "*nature*"); se prendiamo 100 scienziati "normali" e 100 potenziati, la società nel suo complesso non potrà che risultare migliorata dall'attività dei secondi, anche se rimarranno differenze di intelligenza nel secondo gruppo! Il fatto indiscusso che ci sarà sempre un Einstein non toglie utilità per gli altri di essere più intelligenti di prima.

Passiamo al secondo argomento. Una prima risposta potrebbe essere che il rischio evidenziato è sicuramente un rischio reale, tuttavia, da questo punto di vista è importante non dimenticare che già esistono tecnologie di potenziamento a disposizione dell'umanità da secoli e che la società (quantomeno quella occidentale) è riuscita nel tempo (con discreto successo) ad evitare effetti di disuguaglianza massiccia. Solo per fare un esempio sotto gli occhi di tutti, la lettura è una delle forme di potenziamento

cognitivo più antiche: leggere e acquisire conoscenza (tramite libri) è stata – e sarà sempre, sebbene passando per media diversi – la prima forma di miglioramento cognitivo. Non a caso, le società moderne combattono l'analfabetismo e introducono l'istruzione obbligatoria per tutti cittadini che ne fanno parte. Anche se (purtroppo) non siamo abituati a considerare la "cultura" come un potenziamento vero e proprio (vista l'abitudine a considerarla parte indispensabile del nostro sviluppo mentale e sociale), è evidente il successo che la società nel suo complesso sia riuscita ad ottenere: nella maggior parte dei paesi occidentali, infatti, l'istruzione è garantita *indipendentemente* da fattori culturali e socio/economici[100].

Portando l'esempio dell'istruzione nel nostro scenario della Singolarità, un primo motivo di speranza è quindi che il legislatore possa riconoscere l'utilità massiccia di tali tecnologie e, di conseguenza, si impegni a renderle disponibili su larga scala. Con un primo bilancio, possiamo dunque affermare l'esistenza di un *diritto negativo* all'*enhancement*: ogni individuo, consapevole dei rischi e delle conseguenze delle proprie azioni, si fa carico (privatamente) dei costi, guadagnandosi in tal modo pieno diritto di accedere alle tecniche di potenziamento. Come abbiamo visto solo poche righe fa, la questione del *diritto positivo* – ovvero del fatto che la società nel suo complesso promuova, di fatto, il potenziamento – è più complicata e dipende innanzitutto dalla capacità delle istituzioni di *comprendere* il valore potenziale che tale promozione può avere nei riguardi di tutte le persone coinvolte nel processo.

4.1.2 Semi-immortalità.

Una delle conseguenze più "vistose" del miglioramento tecnologico è senza dubbio il prolungamento della vita media: anche solo limitandoci al secolo passato, l'aspettativa di vita per un abitante del nostro paese è aumentata di quasi 20 anni in poco meno di 100. Ovviamente, è facile notare che l'invecchiamento di una grossa parte di popolazione che in precedenza era destinata a morire e lasciare spazio alle nuove generazioni porta con sé una serie di conseguenze tutt'altro che irrilevanti: sul piano individuale, banalmente, il soggetto si trova a dover pianificare una vita molto più lunga e molto più sana che in precedenza; specularmente, la società dovrà convivere con una

[100] All'Unità di Italia, il 78% della popolazione era analfabeta (Chistolini (2001), p. 46).

crescente parte di popolazione poco produttiva – dubitiamo di avere la stessa capacità produttiva a 90 anni, sebbene in uno stato di ottima forma psico/fisica – e in costante bisogno di assistenza.

Affrontiamo gli argomenti uno alla volta. Anche in questo caso, l'avvento della Singolarità sembra – almeno a prima vista – l'avverarsi di un sogno: chi non vorrebbe vivere mille anni? Come in precedenza, una facile domanda porta con sé tutta una serie di considerazioni pro e contro la possibilità. Sul piano "esistenziale" l'argomento a sfavore maggiormente citato in letteratura è quello della *noia*[101]: semplificando, c'è una paura concreta per la mancanza di stimoli adeguati a rendere interessante e degna di essere vissuta una vita così lunga. Sul piano filosofico, invece, gran parte del dibattito si è concentrata sui temi dell'identità personale e sulla relazione tra continuità *metafisica*, Io e prolungamento della vita. Per quanto riguarda il primo aspetto – la paura della *noia* – sembra impossibile compiere una generalizzazione: in particolare, è facile argomentare che anche su un lasso temporale di 100 anni, a persone diverse corrispondono diversi modi di intendere il concetto "vita stimolante" e, ovviamente, se da 100 passiamo a 1000, questo tipo di divergenze non può che diventare esponenzialmente più evidente. Due considerazioni: la prima, analoga a quella utilizzata per il potenziamento cognitivo, è che, per chi lo desidera, sembra esserci un *diritto negativo* innegabile al prolungamento della vita – sarà l'individuo stesso a decidere se è in grado di sopportare la noia oppure no; in seconda battuta, la tecnologia che ha creato il problema, potrebbe anche aiutarci a risolverlo: ad esempio, aumentando il ventaglio di percezioni a nostra disposizione, rendendo più interessante l'assimilazione delle informazioni, garantendo nuove esperienze fisiche e mentali che a stento riusciamo ad immaginare, etc.

Al diritto negativo si contrappone la questione del diritto positivo: ovvero, la società nel suo complesso ha interesse a promuovere la semi-immortalità? La questione è ancora più complicata che nel caso precedente, perché finisce inevitabilmente col sovrapporsi a tutta una serie di questioni legate alla tutela della salute, dell'individuo, etc.; in particolare, in molte nazioni con il *welfare state* la tutela della salute dei propri cittadini è già un diritto: già *ora*, se la tecnologia lo permette, è la società stessa a farsi carico del prolungamento della vita dei suoi componenti, basti pensare a quanto le

[101] Si veda ad es. Jonas (2009).

aspettative di vita media siano aumentate a seguito del *boom* economico. Ciò fa pensare a un utilizzo massiccio da parte della società del futuro (prossimo) di tecnologie mediche avanzate atte a prolungare realmente la nostra vita. Da una parte, esiste una possibilità ovvia: dato che l'*invecchiamento* può essere considerato di per sé una malattia, qualsiasi tecnologia che eviti tale degenerazione diventa un trattamento *desiderabile* al pari di qualsiasi altro trattamento già in essere (cure per il cancro, cellule staminali, etc.); dall'altra, si potrebbe argomentare che le "tecnologie di prolungamento della vita" siano un trattamento *accessorio* rispetto alla salute primaria dell'individuo e che quindi non sia nell'interesse della società promuoverne l'utilizzo su vasta scala (né più né meno di come vengono trattati gli interventi di chirurgia plastica al giorno d'oggi). Questa seconda prospettiva appare però particolarmente complicata da perseguire: se infatti l'accesso a trattamenti estetici non introduce ragionevolmente grandi distorsioni, quello alle tecnologie radicali di prolungamento della vita chiaramente sì. Il risultato quindi è che, rispetto all'attuale situazione, il progresso medico e biologico si potrebbe inserire senza alcuno sforzo nell'attuale impianto sociale, rendendolo così, oltre che negativo, anche un diritto positivo. Per amore di discussione, possiamo porci un ulteriore interessante questione che va molto al di là del puro dato di fatto: è desiderabile che la società *tuteli* la vita dei propri individui? La risposta esula dal contesto tecnologico in senso stretto e interessa in modo molto più ampio il dibattito sul *paternalismo* delle società contemporanee[102]. Semplificando, due sono le prospettive che vanno ad aprirsi: nel momento in cui la tutela della salute è considerata un diritto positivo – come abbiamo visto in precedenza – l'avvento della Singolarità non dovrebbe alterare questo accadimento; se, al contrario, non viene considerata come tale, è importante sottolineare come questo non abbia la forza di cancellare l'esistenza del corrispondente diritto negativo.

Dal punto vista filosofico, una grande attenzione è stata dedicata all'approfondimento dei temi dell'identità personale nel contesto di un'aspettativa di vita esponenzialmente più lunga rispetto a quella entro cui siamo abituati a pensarci. Un argomento che viene spesso menzionato in letteratura è quello dei "piani di vita", ovvero, l'idea che l'identità di una persona nel tempo sia costituita, in parte, da una

[102] Diversi studiosi, ad esempio, partendo dalla costatazione che le decisioni umane, lungi dall'essere l'esito di un calcolo logico sono influenzate da specifici limiti cognitivi, hanno sostenuto l'opportunità di interventi governativi paternalistici che proteggano i cittadini da eventuali conseguenze indesiderate delle loro scelte. Per una prospettiva cognitiva, vedi Thaler, Sunstein (2009).

certa *continuità* di gusti, credenze, desideri e, ovviamente, di pianificazione sequenziale per il raggiungimento degli obiettivi[103]. Su questa base, l'obiezione ovvia ad una vita di mille anni appare immediata: nel corso di un periodo di tempo così *ampio* i cambiamenti che possono avvenire tra le diverse fasi temporali della vita di una persona sono cosi *grandi* da essere globalmente incoerenti; sarebbe perciò impossibile ascrivere ad un solo Io tutto quello che avviene nel corso di un esistenza così lunga[104]. È evidente che l'argomento così espresso ha un certo fondamento, anche se è bene fare una prima precisazione – chiediamoci: quando si dice che 'la coerenza interna è costitutiva dell'identità personale', che cosa si intende esattamente? A onor del vero, il dibattito sull'identità personale vede moltitudini di teorie differenti che da anni si fronteggiano su una serie di argomenti classici. Fondamentalmente tre sono le teorie principali:

1. *La teoria psicologica* secondo cui l'identità personale diacronica dipende da una certa continuità degli stati mentali ('stessa mente = stessa persona')[105].
2. *La teoria fisica* secondo cui l'identità personale diacronica dipende da una certa continuità della composizione fisica ('stesso corpo = stessa persona')[106].
3. *La teoria animalista* secondo cui l'identità personale diacronica dipende puramente da questioni biologiche ('stesso animale = stessa persona')[107].

Dal punto di vista dell'argomento sulla "coerenza della vita", appare chiaro che solo la teoria psicologica ne è maggiormente colpita: infatti, per tutte le altre alternative appare concepibile il fatto che anche in una vita di soli cinquanta anni si dia continuità fisica/biologica senza coerenza nella pianificazione della stessa (ad es., perdita di memoria); tuttavia, sfortunatamente per il sostenitore della Singolarità, risulta evidente che il criterio da privilegiare in tale prospettiva è senza ombra di dubbio il primo, che ben si sposa con una visione funzionalista del mentale, poiché permette l'idea che diversi sostrati fisici supportino uno stesso Io nel tempo. In altre parole, il sostenitore della Singolarità, oltre a parteggiare – come abbiamo nel terzo capitolo – per il funzionalismo, si troverà a privilegiare la teoria psicologica dell'identità, essendo

[103] Vedi Williams (1973).
[104] Vedi Barazzetti, Reichlin (2011).
[105] Vedi Parfit (1984).
[106] Vedi Williams (1970).
[107] Vedi Olson (1999).

l'unica a garantire il risultato sperato – persistenza dell'Io in diversi/o sostrati/o fisici/o – in un'ottica di IA e M*up*[108].

In questa prospettiva, la pianificazione temporale è dunque collegata alla continuità dei miei stati mentali; posso pianificare qualcosa per settimana prossima *solo* nel momento in cui ha senso considerarmi come *esistente* la settimana prossima. Come posso, dunque, pianificare qualcosa tra mille anni? Una prima risposta potrebbe essere che il problema è *già* presente nella nostra vita attuale, sebbene in versione *soft*: conversioni religiose, shock improvvisi o decisioni di vita radicali, sono esempi di eventi – seppur su piccola scala – che in qualche modo *interrompono* la "normale" continuità psicologica del soggetto in questione. Tuttavia, l'esistenza di tali episodi non pare condizionare la nostra possibilità di pianificare *tout court* – male che vada, si cambia "rotta". D'altra parte, l'argomento assume che il valore della pianificazione rimanga inalterato all'avvento della Singolarità, ovvero, come noi pianifichiamo al tempo *t* una vita di 100 anni faremo lo stesso in una vita di 1000. A pensarci bene, questo tipo di cambiamento è già successo nella storia dell'umanità: se un antico romano pianificava una vita in un orizzonte di una quarantina d'anni, l'uomo del secolo scorso poteva permettersi il lusso di pianificare su quasi il doppio del tempo. Eppure nulla – *nulla* – in questo cambiamento ha introdotto difficoltà di pianificazione; semplicemente, la percezione individuale del tempo si è *dilatata* in base alle aspettative di vita. L'avvento della Singolarità potrebbe innescare un fenomeno analogo: quando una vita di 1000 anni sarà "disponibile" è molto probabile che le categorie psicologiche dell'uomo si dilateranno di conseguenza.

Esiste, inoltre, un'altra assunzione nell'argomento che può essere dibattuta e cioè l'assunzione dell'importanza stessa dell'identità personale: è difficile notare questa "premessa nascosta" perché, intuitivamente, appare ovvio che quello che ci interessa nella pianificazione sia il permanere dell'identità. Di fatto, nella vita di tutti i giorni – escludendo scenari di fantascienza – l'identità personale di chi pianifica non è mai davvero messa in dubbio. Come vedremo in modo approfondito nella sezione successiva sul *mind uploading*, è proprio questa tesi chiave a rischiare di essere falsificata. Se il gemellaggio tra funzionalismo e Singolarità ci induce a considerare reale la possibilità che le persone siano *type* e non *token*, è naturale pensare a situazioni

[108] Cfr. Di Francesco (1998), Cap. 3.

in cui – grazie alle nuove tecnologie – abbiamo individui diversi con la mia stessa continuità psicologica. Se questo è vero, il concetto di pianificazione diventa assai meno scontato: non è incoerente pensare che io possa pianificare gli studi del mio discendente MR^2 (economia alla Bocconi) e del mio discendente MR^3 (filosofia al San Raffaele). Perché, in fondo, questa pianificazione non dovrebbe essere legittima? Riaffronteremo questo problema nel dettaglio nella prossima sezione, ma possiamo evidenziare che il bilancio appare piuttosto chiaro: nessuno degli argomenti qui esposti sembra davvero mettere in dubbio la *desiderabilità* sotto il profilo del prolungamento della vita.

4.1.3 *Mind uploading*

È ora il turno dell'argomento più controverso e meno conosciuto all'interno del dibattito sulla Singolarità: mentre il potenziamento umano e la semi-immortalità sono questioni ormai ampiamente discusse in letteratura, il M*up* rimane, forse per la sua apparente connotazione "fantascientifica", piuttosto trascurato. Riepiloghiamo brevissimamente la tecnologia in campo: il M*up* è quella particolare *tecnica* per cui il contenuto di informazione di una mente umana viene ricreato in un supporto digitale. Ne esistono fondamentalmente due tipi, il *soft* M*up*, che riproduce i contenuti grazie ad una sorta di modellazione computazionale dei ricordi scritti di un individuo (mail, video digitali, lettere cartacee, diari di infanzia, etc.) e l'*hard* M*up*, in cui i contenuti vengono riprodotti scaricandoli *direttamente* dal cervello (con, ad esempio, l'utilizzo delle nanotecnologie.). Concentriamoci per ora su ciò che è comune ai due tipi, ovvero il *risultato*: una replica digitale *perfetta* della mia mente (MR^2). Fino ad oggi, come il termine stesso "individuo" tende a suggerire, le persone sono sempre state considerate esemplari *unici* e *irripetibili*: anche due gemelli omozigoti sono infatti, per molti aspetti, visibilmente diversi. Nel momento in cui diviene possibile riprodurre una mente *una volta*, diviene ovviamente possibile farlo *n* volte: in quest'ottica, MR è unica ed irripetibile nel senso in cui lo è *Wish you Were here* o *L'attimo fuggente*, potenzialmente presente in infiniti luoghi e supporti diversi! A questo punto, affrontiamo il problema distinguendo tra filosofia e società. La questione filosofica, anticipata col il problema dell'identità personale della sezione precedente, fa emergere una versione del reale quantomeno bizzarra: in un mondo "pre-Singolarità", la

persistenza dell'Io non è da nessun punto di vista problematica, ovvero, quando mi trovo a pianificare la mia esistenza, so *benissimo* che non ci saranno dubbi sul fatto che sarò io a condurre quella stessa esistenza su cui progetto e nessun altro; nel mondo della Singolarità, invece, questa immagine si frantuma letteralmente: nel momento stesso in cui pianifico cosa farò domani, devo sempre avere in mente l'eventualità che ci sia più di un possibile candidato per fare quello che ho pianificato (MR^2, MR^3 o, perché no?, entrambe). Ancora peggio, *tutti* i candidati potrebbero essere indistinguibili tra loro – come la versione di *Wish You Were Here* sul mio iPod, quella conservata in una cartella sul desktop del mio PC e quella su vinile del 1975. Come già previsto in Parfit (1984), questo tipo di situazione confligge prepotentemente con l'idea intuitiva che la persistenza dell'io sia ciò che davvero conti quando si parla di persone: ovvero, quando prendo una decisione, è importante che l'agente in questione sia sempre lo stesso, al momento di agire e al momento di subirne le conseguenze. L'introduzione del M*up* apre la porta a tutta una serie di scenari in cui potrei non essere *io* a soffrire per le azioni sbagliate o gioire per le scelte corrette, ma i miei cloni. Questo ci impone di affrontare due questioni:

i. Di fronte alla imminente morte del mio corpo, è *desiderabile* sopravvivere sotto forma di cloni *non* identici?
ii. Ammesso di volerlo fare, qual è lo stato psicologico di una persona che si trova a vivere attraverso i suoi cloni?

Affrontando (i), ci accorgiamo che le opzioni sembrano essere fondamentalmente tre: la prima è negare *in toto* che la sopravvenienza attraverso coloni sia desiderabile, la seconda è invece *abbracciare* un'idea "parfittiana" e sostenere che ciò che conta davvero è la semplice continuità degli stati mentali – pertanto, la sopravvivenza attraverso cloni diventa desiderabile tanto quanto la persistenza normale dell'identità. La terza ipotesi è un compromesso tra le due visioni: se infatti la desiderabilità non è concepita come tutto-o-nulla, ma come una relazione che "ammette gradi", si potrebbe sostenere che la sopravvivenza attraverso cloni sia, ad esempio, meno desiderabile di quella *standard*, ma *più* desiderabile della morte.

A questo punto, possiamo dire che la prima posizione è quella del senso comune: non mi importa che qualcuno di molto simile me vinca al superenalotto, sono *io* a dover vincere. Allo stesso modo, non importa se il mio clone si ammala di una brutta malattia, ma gioisco del fatto che non sono *io* ad averla. Di contro, la seconda opzione è chiaramente controintuitiva. Tuttavia, possiamo provare a rendere il ragionamento più robusto in questo modo: supponiamo di avere un individuo (MR) che, ad un certo punto della sua esistenza, viene *uploadato* su due diversi supporti. L'argomento assume, ovviamente, che MR non possa essere identico ad entrambe le sue copie e, d'altra parte, che non ci sia motivo di pensare che sia identico a una e non all'altra – essendo identici per definizione i contenuti. Possiamo, però, concepire la situazione in modo diverso, aiutandoci con un'analogia. Consideriamo la nostra MR che si duplica allo stesso modo di una strada che si biforca: in quest'ottica, MR non si sdoppierebbe in due successori – così come la strada non si "sdoppia" in senso stretto – ma "propagherebbe" semplicemente la propria continuità psicologica in due rami differenti. Con un rovescio completo della prospettiva standard, possiamo affermare che prendere sul serio la teoria parfittiana, significa identificare con una e una sola persona tutti i vari flussi che possono essere ricondotti ad una comune origine. Come le strade e i fiumi, le persone si *ramificano* e si "fondono" in altre entità senza per questo perdere l'identità del flusso degli stati mentali – allo stesso modo in cui gli affluenti sono parte del Po. I due cloni sono, da ultimo, parte di un individuo più grande (MR) che comprende la carriera di MR pre e post M*up*: quello che davvero conta è l'identità del processo *causale* e non l'idea di persona tanto cara al senso comune. A questo punto è bene precisare che, di fatto, i due cloni sono parti distinte dello *stesso* processo mentale, ma *non* sono persone identiche: quello che succede quando *noi* sopravviviamo è solo *accidentalmente* la persistenza della stessa mente. In fondo, secondo Parfit, anche all'interno di MR avviene lo stesso processo, seppur su minima scala. Noi siamo radicalmente diversi rispetto a quando avevamo tre anni e saremo molto diversi rispetto a quando ne avremo ottanta, ma non mettiamo certo in discussione il fatto di essere, in un certo senso, sempre lo stesso individuo.

La terza soluzione, invece, riconosce in qualche modo dei meriti alla teoria parfittiana senza spingersi così in là da accettare *in toto* le conclusioni: l'idea qui è che noi persistiamo e sopravviviamo in modo standard, ma in mancanza di una vera

sopravvivenza (come identità) è comunque preferibile la sopravvivenza di un nostro clone alla cessazione totale del nostro flusso di stati mentali – ad es., se il sogno della mia vita fosse quello di scrivere un romanzo e non potessi finirlo perché prossima alla morte, vorrei che a portarlo a termine fosse una persona "il più possibile simile a me". È tuttavia difficile ipotizzare come (di fatto) le persone reagirebbero a un'invenzione del genere: qualcuno sarebbe entusiasta, qualcuno palesemente refrattario, qualcun altro (forse la maggioranza) sarebbe possibilista. L'unica previsione certa che possiamo fare sull'argomento è che, di certo, tale rivoluzionario cambiamento richiederà un ripensamento critico di molte delle nostre abitudini. Così come fino ai primi del Novecento, le opere d'arte erano considerate dei *token* e l'avvento della tecnica le ha rese molto più simili a dei *type* – cambiando radicalmente il rapporto con l'arte –, lo stesso potrebbe accadere per l'Io nel mondo della Singolarità: chiedersi dove va a finire l'arte nell'epoca della sua riproducibilità, è simile al chiedersi che fine fa l'Io nell'epoca della duplicabilità.

Oltre alle prospettive esistenziali esistono conseguenze tutt'altro che banali sul piano sociale dello sviluppo delle tecnologie post-Singolarità che sono state largamente ignorate all'interno del dibattito[109]. In generale, come noto, il valore di un bene deriva dall'incontro tra la domanda e l'offerta[110]. Per fare un esempio banale, il petrolio ha molto valore perché tanta gente lo richiede e la sua disponibilità in natura è scarsa: se domani inventassimo una macchina ad idrogeno, il petrolio smetterebbe di avere valore perché la domanda crollerebbe. Non solo, l'idrogeno sarebbe a costo zero, poiché, nonostante la domanda in aumento esponenziale, la sua disponibilità è tanto grande – l'idrogeno è la risorsa più abbondante nell'universo – da creare una dinamica in cui l'offerta rimarrebbe infinitamente più grande della domanda. La prospettiva è analoga nel caso del mercato del lavoro: alcuni attori del mercato propongono un compenso per svolgere determinati compiti e altri attori offrono, banalmente, il proprio tempo e le proprie capacità per svolgere tali compiti; è facile constatare che meno persone riescono a fare un determinato lavoro, più quel lavoro sarà ben pagato – questo è uno dei motivi principali per cui Tiger Woods è un "attore" più pagato rispetto ad un impiegato delle poste. Come nel caso del petrolio, dunque, il mercato del lavoro attuale si basa

[109] Ma si vedano Hanson (1994) e Hanson (2008) per una analisi squisitamente economica.
[110] Questa, ovviamente, è una semplificazione: tuttavia è sufficiente ad introdurre chiaramente il problema. Per un approfondimento, si veda ad esempio Frank (2007), Cap. 2 e Cap. 4.

sull'assunzione che le risorse principali di scambio – in sintesi, le capacità cognitive – siano limitate. È facile notare come il M*up* potrebbe essere l'equivalente del motore all'idrogeno per il mercato del lavoro. Supponiamo infatti di poter produrre sistemi artificiali in modo rapido ed economico, clonando sistemi naturali esistenti: in poco tempo, l'offerta di capacità cognitive aumenterebbe in modo esponenziale e, se la domanda non dovesse crescere parallelamente alla prima, il prezzo con cui la risorsa "intelligenza" viene scambiata non potrebbe far altro che precipitare. Come nessun attore sul mercato pagherebbe più un euro per un barile di petrolio, allo stesso modo non si spenderebbe più un centesimo per il lavoro di una mente umana.

La prospettiva globale è dunque molto più seria di quello che si tende ad assumere. Se la maggior parte delle previsioni apocalittiche post-Singolarità – ad. es., macchine che conquistano la terra, caccia ai cyborg, etc. – sono spesso molto citate, ma risentono inevitabilmente di echi fantascientifici, la prospettiva di un mondo in cui la competizione economica finisce per emarginare le menti umane a favore di quelle artificiali appare uno scenario assolutamente realistico. In altre parole, la razza umana non rischierebbe l'estinzione per il dominio della terra, ma per la mancanza di condizioni economiche adeguate al proprio sostentamento! Ovviamente, anche se abbiamo affrontato questo discorso all'interno della sezione sul M*up*, l'IA forte pone interrogativi identici: la differenza però – cruciale – è che da molti punti di vista il M*up* è una questione più semplice da trattare. Anche se gli esseri umani fossero impossibilitati a produrre da zero un sistema artificiale pensante, una de-ingegnerizzazione a basso livello dei meccanismi cerebrali potrebbe essere sufficiente per il M*up*: potremmo riuscire a duplicare menti umane senza mai dover capirne il funzionamento e quindi, anche se l'IA fallisse, la sola possibilità del M*up* è sufficiente a generare lo scenario socio/economico che abbiamo appena descritto.

4.2 Individuo, etica e società

L'ultimo argomento che desideriamo trattare riguarda le possibili conseguenze della Singolarità sull'individuo in quanto *agente morale* che fa e subisce azioni e vive in un determinato contesto.

In particolare, discuteremo del rapporto tra "possedere una mente" (naturale o artificiale) e "avere diritti/doveri" e analizzeremo il problema di agire moralmente in un mondo in cui i confini dell'Io si fanno sempre più labili

4.2.1 Individuo ed Etica

Come abbiamo visto nelle precedenti sezioni, la stabilità dell'Io è messa a dura prova dalle tecnologie della Singolarità. È curioso che già David Hume, uno dei primi filosofi a dubitare dell'esistenza di un Io sostanziale, inizi la sua carriera Etica dichiarando che 'non c'è nulla di più certo ed indubitabile dell'Io'[111]. In particolare, il filosofo sembra suggerire che, sebbene in fase "metafisica" si possa legittimamente dubitare dell'esistenza di un Io sostanziale, quando si passa alla riflessione Etica è in un certo senso la materia stessa a presupporlo. Oltre alla "realtà dell'Io", esiste un altro problema, altrettanto grave da risolvere all'interno del dibattito sulla Singolarità, ovvero l'attribuzione di diritti e doveri morali, che normalmente fanno riferimento a persone umane. Il problema è duplice: da una parte è difficile stabilire quali sistemi naturali o artificiali meritino un certo tipo di attenzioni speciali, dall'altra la disgregazione dell'Io pone la questione di identificare nel mondo quali sistemi facciano in effetti parte o meno del dato individuo che vogliamo tutelare. Affrontiamo i problemi uno alla volta.

Innanzitutto, proviamo a individuare alcune caratteristiche "strutturali" – una sorta di primitivi – dell'Etica *generalmente intesa*, indipendentemente dai singoli modi usati per declinare una teoria morale.

> **PDE**$_0$) l'Etica dipende dall'Io.
>
> **PDE**$_1$) Il giudizio morale dipende dalla possibilità di una corretta individuazione del soggetto nel mondo. Se non possiamo indicare chi agisce, non possiamo giudicare il suo operato come buono o cattivo e neppure se egli stesso lo è.
>
> **PDE**$_2$) C'è una differenza assiologica tra ciò che è parte di un soggetto e ciò che non lo è. In pratica, questa tesi codifica l'idea intuitiva secondo cui io posso vantare diritti privilegiati sulle parti di me che su altri oggetti non posso vantare:

[111] Una trattazione chiara dell'argomento humeano sui rapporti tra Io ed Etica è esposta in Di Francesco (1998).

> un danno alla mia persona – ad esempio, la ferita ad un braccio – non è equiparabile al furto della mia calcolatrice.

Chiediamoci ora quali siano i problemi che emergono dall' inserimento dell'Etica nel *framework* della Mente Estesa, così come introdotti nel Capitolo 3: (**PDE**$_1$) ci dice che il giudizio morale dipende dalla corretta individuazione del soggetto nel mondo. In altre parole, per formulare un giudizio morale sulle azioni di *x* del tipo: 'il soggetto *x* ha agito bene/male', prerequisito essenziale è l'individuazione di *x* nel mondo. Nello specifico, questo significa che nel momento in cui non è possibile distinguere il soggetto dalle azioni che compie (o dagli altri soggetti presenti in scena), non è parimenti possibile individuare l'origine delle azioni morali e valutarne gli effetti. Il problema è stato sollevato in modo più che brillante da Diego Marconi con il famoso argomento delle *ragazze pigre*[112]:

> *Il caso delle ragazze pigre*. Emma è una ragazza pigra che odia il latino. Suo padre è un uomo ricco e permissivo e le compra un traduttore elettronico che Emma porta sempre con sé durante le versioni e di cui si fida ciecamente. Anna è pigra e suo padre non è decisamente ricco come quello della sua amica Emma. Tuttavia, suo padre è un latinista geniale e, siccome fa lo scrittore, passa sempre in casa le sue giornate. Tutte le volte che ad Anna serve una traduzione, il padre esaudisce i suoi desideri ed esegue il compito simultaneamente supportato dalla piena fiducia della figlia.

Il punto retorico è dunque il seguente: se Emma e il dispositivo sono un *unico* sistema cognitivo e quindi un'*unica* mente, cosa dire di Anna e del padre? Dovremmo considerare anch'esso un unico sistema cognitivo (un'unica mente), ma, chiediamoci, è davvero plausibile? Chi fa le traduzioni? *Quante* menti ci sono in casa di Anna?

Nel nostro discorso l'argomento di Marconi è particolarmente rilevante perché rende vivido il problema di assegnare responsabilità: se nel caso di Emma, non sembrano esserci grossi problemi psicologici, per Anna la questione non pare altrettanto semplice: di chi sono, infatti, i *meriti* (e le *colpe*) della traduzione? Il problema è che riusciamo a

[112] Vedi Marconi (2005).

capire chi ha agito solamente in relazione al punto di vista: secondo Anna è lei a tradurre, secondo il padre è lui a tradurre, ma cosa dovrebbe dire l'arrabbiato professore di latino (che crede che Anna sia solo una grande ignorante[113])? Questo mal di testa è praticamente inevitabile ogni qual volta si prenda sul serio la prospettiva della Mente Estesa – e, *a fortiori*, scenari post-Singolarità come il M*up*: tracciare il soggetto nel mondo, le sue azioni e i nessi causali rilevanti (tutte procedure *essenziali* per l'attribuzione di responsabilità) diventa un'impresa ai limiti del paradosso. Si badi bene che questo problema *non* dipende dalla nozione di una specifica teoria Etica: o l'Etica è colpita in uno dei suoi punti fondamentali *oppure* la Mente Estesa, viste le conseguenze, è una teoria insostenibile[114].

(**PDE**$_2$) ci dice invece che c'è una differenza assiologica tra ciò che è parte di un soggetto e ciò che non lo è. Questa tesi codifica l'idea intuitiva secondo cui posso vantare diritti privilegiati sulle parti di me che non posso vantare su altri oggetti (un danno al mio cervello, non è paragonabile al furto della mia agenda). Neil Levy, in un articolo sui rapporti tra Mente Estesa e neuroetica[115] fa quasi la stessa osservazione quando fa notare l'assunzione tradizionale che vuole che un intervento al cervello (quindi *interno* al mio corpo) abbia una valenza unica e qualitativamente distinta da qualsiasi altro tipo di intervento, specie se "ambientale":

[113] Ovviamente "ignorante" è anch'esso un termine che risente della prospettiva da cui si tracciano i confini: per il professore, infatti, Anna non sa tradurre, ma Anna, aiutata da Chalmers e Clark, può ribattere che, incorporando qualche funzione cognitiva esterna, anche il latino più difficile diventa per lei comprensibile.

[114] Inutile dire che siamo più propensi per la prima ipotesi che per la seconda.

[115] Doverosa mi sembra a questo punto una nota consistente su quello che è forse l'unico ponte tra Etica e mente estesa che per ora è stato gettato nel panorama filosofico. Il punto centrale è che la teoria della Mente Estesa nega che gli stati mentali siano limitati ai soli stati cerebrali. Questo si scontra con quelli che Levy considera come presupposti della neuroetica contemporanea, ovvero a) c'è qualcosa di speciale e distintivo dell'intervenire sul cervello per alterare gli stati mentali; b) la neuroetica si occupa di stabilire se è moralmente accettabile intervenire direttamente sugli stati mentali. Se la teoria della mente estesa è vera, (a) e (b) non possono più essere accettati acriticamente. Prendiamo, come esempio, il fatto di dover calmare una classe di studenti particolarmente agitati e rumorosi. Per fare ciò si può sedare la classe con dei farmaci, oppure dare loro un'educazione extrascolastica più severa. Se entrambe le strategie producessero lo stesso effetto, un filosofo può ritenerli moralmente differenti *solo* se accetta (a), e quindi valuta l'intervento "cerebrale" in modo diverso da quello "ambientale". Ma, se la tesi della mente estesa è vera, non c'è più modo di giudicare in modo moralmente diverso le due strategie. Per quanto riguarda (b), Levy considera il dibattito sul miglioramento delle abilità mentali: è eticamente corretto utilizzare strumenti esogeni per migliorare le performance cognitive? Ovviamente, *se* la teoria della mente estesa è vera, allora gli stati mentali sono composti sia dal cervello, sia dall'ambiente, per cui intervenendo su di esso (sky, ipod, tv, PC, internet, dvd…) stiamo *già* da decenni migliorando le nostre performance ed intervenendo sui nostri stati mentali. Quindi la domanda da porsi non è più "è giusto eticamente intervenire sul cervello per migliorare le performance?" ma "*quali* interventi è giusto fare?". La morale conclusiva che Levy trae dai suoi argomenti è che la mente estesa sembra far perdere terreno alla neuroetica, ma in realtà la rafforza: se infatti la mente si estende nel mondo, la neuroetica estende con essa i propri confini di studio.

> ‘Se gli stati mentali non sono confinati nel cervello, allora la tesi per cui gli stati neurologici sono specialmente problematici, o unici nella loro natura, necessita di una difesa più dettagliata di quelle che solitamente sono assunte. Accettare la tesi della Mente Estesa […] ci richiede di pensare e rigettare l’idea […] che gli interventi nel cervello siano l’unico modo in cui possiamo direttamente intervenire sugli stati mentali che costituiscono la nostra identità.’[116]

Riprendiamo il caso di Otto e Inga e immaginiamo che a Inga rubino l’agenda: rubare l’agenda ad Inga, che ha una memoria funzionante, non sembra avere lo stesso peso morale che se le si prendesse la testa a bastonate fino a causarle danni permanenti alla memoria. Per ora, tutto sembra andare come previsto da (**PDE**$_2$). Ma consideriamo Otto: in questo caso prendere la testa di Otto a bastonate sembra sbagliato tanto quanto rubare il taccuino. Questo esempio non solo ribalta la prospettiva di (**PDE**$_2$), ma getta nuova luce sul caso di Inga: il nostro giudizio intuitivo sulla maggiore importanza del cervello per Inga rispetto all’agenda si trova a essere spiegata su basi del tutto differenti dalla relazione esterno/interno. Come il caso di Otto rende evidente, è *vero* che è più grave un danno al cervello di Inga che alla sua agenda, ma solo perché in Inga è il cervello a svolgere certe funzioni cognitive. In poche parole, quello che conta nel giudicare la gravità di un’azione non è tanto il luogo in cui si è subito un danno, bensì ciò che esso comporta! È il supporto di elaborazione quello che conta e non *dove* esso è situato, e questo spiega perché un danno al cervello di Inga è uguale, funzionalmente e “moralmente”, a un danno al taccuino di Otto. Inoltre, poiché, come abbiamo visto, è spesso difficile distinguere i confini dei sistemi elaborativi, le inconsistenze di (**PDE**$_2$) sono palesi.

Che cosa dire, infine, di (**PDE**$_0$)? La caduta di (**PDE**$_1$) e (**PDE**$_2$) non fa altro che mostrare come sia difficile ridare un ruolo morale efficace a un Io esteso e frammentato; il legame concettuale tra Etica e Io esplicitato da (**PDE**$_0$) sembra diventare l’etichetta di una categoria vuota, un po’ come “il cavallo blu del re dei pinguini”: non c’è nulla di contraddittorio o di intrinsecamente sbagliato in tali etichette, ma “semplicemente” il mondo non contiene oggetti che soddisfano i requisiti richiesti.

[116] Cfr. Levy (2007), p. 7.

Mostrare che i mattoncini dell'Etica hanno problemi a reggere nell'ottica della Mente Estesa e, di conseguenza, negli scenari tecnologici proposti, non significa *rigettare* l'Etica *tout court* ma semplicemente far notare l'importanza che tali argomenti hanno in un'ottica post-Singolarità. Decidere di chiudere gli occhi davanti alle conseguenze che la frammentazione dell'Io, la clonazione della mente e l'IA nel suo complesso portano con sé in un futuro possibile non lontano è un atteggiamento superficiale. Se ammettiamo la necessità di un'Etica – seppur riformulata – non possiamo fare altro che prendere in considerazioni gli scenari proposti e sviluppare delle adeguate riflessioni filosofiche. Alla domanda 'Otto ha ancora *bisogno* di un'Etica nonostante sia sparso nel mondo?' (o, parallelamente, 'I cloni di MR hanno bisogno di un'Etica nonostante siano semplicemente cloni?') ci sentiremmo di rispondere con un *sì*. Di contro, l'ulteriore questione delle menti artificiali – dal dispositivo di traduzione di Anna, ai cyborg perfetti di *Blade Runner* – va ancora abbondantemente indagata. Se, da una parte, ci troviamo nella scomoda situazione di dover rincorrere l'Io nel mondo, dall'altra questi nuovi scenari tecnologici sembrano vedere la nascita di nuove forme di vita sulle quali ragionare in termini etici.

Per quanto riguarda il secondo punto, ovvero l'attribuzione di diritti e doveri a sistemi artificiali, se in un mondo pre-Singolarità i dilemmi etici sono (a torto o ragione) piuttosto rari – ad esempio, tirare un calcio a un sasso o uccidere una formica vengono difficilmente considerati un problema insormontabile –, un mondo post-Singolarità non permette più di rimandare una riflessione adeguata su tutti gli enti del reale. Se siamo in cerca di condizioni per l'attribuzione di uno *status* morale è conveniente iniziare dalla nozione di persona: anche intuitivamente, sono le persone i soggetti morali più importanti sulla scena e sono sempre le persone gli enti verso i quali ci è richiesta una maggiore attenzione. A questo proposito, consideriamo le *sei condizioni* proposte da Dennett per identificare una persona[117]:

1. Essere razionale.
2. Possedere stati di coscienza/stati intenzionali.
3. Essere oggetto del cosiddetto *atteggiamento intenzionale.*

[117] Vedi Dennett (1976).

4. Essere in grado di *interpretare* gli altri soggetti attraverso l'atteggiamento intenzionale.
5. Essere capaci di *comunicazione* verbale.
6. Possedere una qualche forma di autocoscienza.

Commentiamo le condizioni una per una. Per quanto riguarda (1), la forma di razionalità minimale richiesta è un semplice adeguamento al ragionamento mezzi/fini: in altre parole, il comportamento di una persona è *interpretabile* come il tentativo di soddisfare certe preferenze a partire da certi mezzi e certe informazioni di base (io posso interpretare *x* come *razionale*). (2), (3), e (4) sono collegati al primo e, di fatto, sono tutti attinenti ad un termine tecnico specifico della filosofia dennettiana che si chiama, appunto, *atteggiamento intenzionale*. Cosa intende esattamente Dennett? Prendiamo un sasso e lanciamolo dalla finestra: per predire la fine che farà il sasso ci basta una conoscenza *fisica* di base. Prendiamo ora un sistema leggermente più complesso di un sasso, ad esempio una stanza contenente un termostato e supponiamo di utilizzare i controlli del termostato per aumentare la temperatura della stanza: una volta schiacciato il tasto "+", la nostra predizione è che, in poco tempo, la stanza diverrà più calda. Qual e la giustificazione di questa predizione? Di certo, non può essere la nostra conoscenza della fisica: nessun essere umano dispone di una descrizione fisica esaustiva di un termostato! Piuttosto, dice Dennett, la nostra predizione si basa in questo caso su una conoscenza *funzionale* dei meccanismi coinvolti (derivati dalla semplice connessione causale tasto-caldaia-termostato-temperatura). È importante apprezzare l'enorme vantaggio predittivo della visione funzionale: mentre il miglior fisico del mondo non potrebbe predire alcunché in questo caso, un bambino di dieci anni riuscirebbe a scaldare la stanza in pochi secondi. Saliamo ancora di complessità e consideriamo un nostro amico (JT) che seduto sul divano ci dice 'Ho sete'. La nostra previsione, in questo caso, sarà la seguente: JT si alzerà, andrà verso il frigorifero, ci ruberà una birra e la berrà alla nostra salute. La prospettiva fisica e quella funzionale non sono applicabili: non solo non sappiamo la "fisica degli umani", ma siamo molto lontani dal poter osservare *direttamente* l'elaborazione delle informazioni nel nostro cervello – e in quello di JT. Secondo Dennett, quello che accade nella scena appena descritta è un *ulteriore* salto di prospettiva, nient'altro che l'applicazione del cosiddetto atteggiamento

intenzionale: interpretando il comportamento di un sistema, assumendo che abbia certi desideri e certe credenze, siamo in grado di predire con grande accuratezza il comportamento di sistemi infinitamente complessi. In poche parole, le condizioni dennettiane ci dicono che ogni persona è tale anche perché è soggetta all'interpretazione intenzionale da parte degli altri ed è in grado di interpretare gli altri applicando lo stesso metodo. Infine, le ultime due condizioni sono in qualche modo mutuate dal senso comune: la capacità di esprimersi e quella di "percepire se stessi", sono infatti da sempre considerate caratteristiche proprie dell'uomo, indipendentemente dalle specifiche tesi di Dennett – ovvero sono indipendenti dallo status ontologico di desideri e credenze. Tali condizioni sono senza dubbio un punto di partenza più che plausibile per cominciare un dibattito sull'argomento: ad esempio, recenti studi di neuroimmagine hanno reso possibile l'applicazione di tali categorie anche a soggetti "in coma" o, comunque, in condizioni palesemente anormale[118]. Ovviamente, dal punto di vista della Singolarità la cosa interessante è capire se tali criteri possano essere estesi anche al di là degli esseri umani. Chiaramente, qualsiasi sistema artificiale come quello che l'IA vuole riprodurre soddisferebbe tutti e sei i requisiti. Senza ricorrere a bizzarri scenari fantascientifici, possiamo cominciare da un caso reale per apprezzare meglio il ragionamento che ci sta dietro: supponiamo di voler sfidare il nostro PC a scacchi con uno dei programmi attualmente in circolazione; quando giochiamo contro il computer, ciò che facciamo è nientemeno che trattare il programma con l'atteggiamento intenzionale. Infatti, per vincere, *non* utilizziamo di certo una teoria fisica sui processori, né una teoria funzionale sugli algoritmi che sottostanno al software: per vincere, quello che dobbiamo fare è tentare di *prevedere* il comportamento del PC, ascrivendogli un desiderio – vincere la partita – e tutta una serie di credenze corrispondenti alle diverse strategie possibili. Solo in questo modo abbiamo una speranza di vincere ed è solo in questo modo che il suo comportamento ha un senso, altrimenti percepiremmo ogni mossa come scollegata dalla successiva. Se questo è vero, esistono già sistemi per cui (1-3) sono verificate. Quello che manca oggi, sono gli ultimi due passaggi: già esistono sistemi artificiali in grado di comunicare in limitati domini in modo molto efficace – lo stesso termostato ci dice, tutto sommato, qualcosa di importante sulla stanza –, ma non è ovviamente questo il tipo di comunicazione

[118] Vedi Monti, Laureys, Owen (2010) e Monti *et al*. (2010).

richiesto. Esiste, dunque, un modo per rendere questa intuizione più robusta? Diciamo, innanzitutto, che la capacità di comunicazione ammette *gradi*: fatta 100 quella umana di padroneggiare perfettamente sintassi, semantica e pragmatica[119], possiamo dire che la capacità attuale dell'IA migliore in circolazione, si aggira probabilmente intorno al 20 – ha ormai una sintassi molto buona, una semantica sufficiente e specializzata in aree molto limitate, ma un'area pragmatica spesso totalmente inesistente. In particolare, quello che manca è una teoria perfettamente generale e ricorsiva del linguaggio naturale: come ha notato e spesso ripetuto da Chomsky, il fatto più stupefacente della nostra competenza linguistica è la capacità di capire e produrre un numero infinto di enunciati, pur possedendo una mente finita. Come argomentato da Davidson[120], è dunque la *ricorsività* la proprietà chiave del linguaggi umani: se è vero che anche le api posso comunicare tra loro, la loro comunicazione è limitata ad un set ben definito di messaggi – questo è totalmente diverso per gli umani che sono in grado di associare condizioni di verità ad enunciati mai sentiti prima. Per quanto riguarda l'aspetto pragmatico, ovvero la capacità dell'individuo di comportarsi appropriatamente a fronte della domanda 'Puoi passarmi il sale?', possiamo congetturare che sia in parte legato anche al punto (6): se infatti, sintassi e semantica possono progredire in modo autonomo – grazie ai corrispettivi sviluppi delle relative discipline formali – appare *impossibile* acquisire una competenza pragmatica senza una rappresentazione degli stati mentali piuttosto complessa. È chiaro comunque, che in questi ultimi due punti si sta parlando di differenze di grado e *non* di natura tra esseri umani e IA: l'avvento della Singolarità dovrebbe accorciare le distanze fino a chiudere per sempre questo *gap*. In quel momento, allora, molti sistemi artificiali si troveranno a soddisfare le condizioni di Dennett una per una e se oggi staccare staccare la spina ad un PC è un comportamento moralmente irrilevante, domani potrebbe essere tanto grave quanto commettere un omicidio. Se questo è vero, ci si aspetta che gli esseri artificiali ricevano gli *stessi* diritti e doveri di un essere umano; questo significa, ad esempio, che l'utilizzo dei PC come schiavi dovrebbe essere considerato moralmente identico al caso umano.

È piuttosto interessante notare che il cambiamento di prospettiva apre senari inediti anche nell'altra direzione: supponiamo infatti di disporre di menti artificiali e di

[119] L'esatta relazione tra sintassi, semantica e pragmatica è ovviamente oggetto di dibattito (si veda a proposito Bianchi (2003), Cap. 4). Tuttavia, tutti i partecipanti alla discussione concordano che una teoria completa della comunicazione umana debba in qualche modo tenere conto di tutti questi tre ingredienti.
[120] Vedi Davidson (1967).

programmarne il contenuto secondo le nostre esigenze. Quando il sistema artificiale "nascerà", sarà dunque quanto di più possibile vicino alle nostre aspettative: ma se questo tipo di intervento è concesso con le menti artificiali perché dovrebbe essere vietato per sistemi naturali? In altre parole: dato che una delle caratteristiche più interessanti delle menti artificiali è poterle produrre "su misura" – e data l'equivalenza stabilita sopra grazie alle condizioni di Dennett – come possiamo allora condannare lo stesso tipo di selezione su organismi umani? In altre parole ancora: come possiamo immaginare di produrre menti artificiali "a piacere" e condannare al contempo forme di eugenetica quando si applicano agli esseri umani?

4.2.2 Soggetti morali e politica

Per concludere questa sezione, ci occuperemo brevemente di due aspetti all'intersezione tra sviluppo tecnologico e vita pratica. Il primo è implicitamente già presente in molte considerazioni fatte nella sezione precedente: quali diritti assegnare ad un sistema artificiale?

Il problema ha radici molto più profonde della Singolarità. Per molti aspetti, il pioneristico Turing (1950) aveva già chiaramente individuato un problema di "doppio standard": qualsiasi motivo abbiamo per dire che un computer che sembra intelligente è "solo" una simulazione, possiamo applicare lo stesso discorso alle persone[121]. Anche accettando questa "equivalenza funzionale morale", per cui due sistemi, S_1 e S_2 hanno identici diritti e doveri morali se sono funzionalmente analoghi, rimane il problema di stabilire in quale momento dello sviluppo tecnologico le macchine potranno ambire ai primi diritti e doveri. Intuitivamente, la nostra morale pre-Singolarità sembra assegnare una qualche sorta di "primato morale" a sistemi biologici complessi[122]: è moralmente ininfluente uccidere (senza motivo) una formica, è abbastanza ripugnante uccidere un cane, è decisamente sbagliato uccidere un uomo. Lo stesso discorso si potrebbe *prima facie* applicare anche alle Intelligenze Artificiali, aggiungendo addirittura una

[121] Vedi anche Aaronson (2013), pp. 33-34.
[122] Non ci importa qui stabilire se questo sia in sé giustificato o meno, ma capire come l'attuale atteggiamento dominante possa/debba cambiare in una ottica di Singolarità.

connotazione maggiormente oggettiva al giudizio: dato un sistema artificiale è infatti più semplice, rispetto a un sistema biologico, poter quantificare in qualche modo le sue funzioni e quindi stilare un "ranking" di complessità, una sorta di "coscienziometro"[123]. A quel punto, a maggior coscienza si può far corrispondere un maggiore *status* morale: qualsiasi siano le intuizioni morali e gli argomenti che razionalizzano il nostro comportamento attuale, sembra che in questo caso specifico si possano benissimo trasportare ad un mondo post-Singolarità. Ovviamente, un ordinamento di complessità non risolve il problema del livello minimo di coscienza necessario ad essere considerato un soggetto morale: come per la morale umana (si pensi al concetto di "maggiore età"), ad un certo punto la decisione potrebbe essere una semplice *stipulazione*[124].

Il secondo aspetto cui vogliamo accennare è quello dell'equità, anticipato nella sezione sul potenziamento cognitivo: dato l'avvento della Singolarità e dei suoi immensi benefici, come avverrà la distribuzione degli stessi? Indubbiamente il problema dell'equità è una delle preoccupazioni maggiori sollevate dai detrattori della Singolarità: mentre crediamo sia impossibile valutare la critica nel contesto allargato di tutte le problematiche legate alla Singolarità, possiamo concentrare la nostra discussione limitandoci all'*enhancement.* L'argomento dell'equità potrebbe dunque essere ricostruito come segue: se il potenziamento fosse disponibile solo ad una limitata fascia di popolazione, ci sarebbe il rischio di un circolo vizioso (chi può permettersi il potenziamento diviene più intelligente ed è così più probabile possa accedere ad ulteriori potenziamenti e così via). Come ben argomentato in Sandberg, Savulescu (2011), l'argomento assume che le tecnologie di *enhancement* siano disponibili semplicemente attraverso un apposito mercato, ma, per ragioni già discusse in precedenza, non è affatto ovvio che non sia la società nel suo complesso a farsi carico della distribuzione di tali strumenti. Per quanto interessante, questo dibattito non ci appare fondamentalmente *nuovo* rispetto ad altre sfide che la società attuale, in un modo o nell'altro, si trova ad affrontare su base quotidiana: anche senza essere funzionalisti estremi, è evidente che il “problema della pillola della intelligenza” sia molto simile al

[123] Quantificare il "livello di coscienza" è ovviamente, allo stadio attuale, poco più che una metafora. Tuttavia è importante sottolineare come anche in contesti biologici comincino a comparire metriche interessanti per questo scopo, come ad esempio la teoria dell'informazione integrata proposta in Tononi (2008).

[124] A marzo 2013, il Tribunale Civile di Milano ha dichiarato Google non responsabile dei risultati del suo algoritmo di ranking delle parole chiave, di fatto sancendo una sorta di indipendenza tra "creatore" e "creatura". È molto presto per parlare di Singolarità, ma è chiaro che il problema dello status morale dei software è già oggi un problema (in particolare, chi paga se il software sbaglia?).

“problema dell'Università di Harvard”. Infatti, una università prestigiosa è, tipicamente, molto costosa ma aumenta le probabilità di avere una vita di successo, con la quale aumentano le probabilità di poter mandare i propri figli alla stessa università, e così via. Non stiamo suggerendo che questo non sia un *grande* problema nella società contemporanea (non da ultimo, perché l'ammissione ad università prestigiose di persone meritevoli ma non abbienti porta benefici alla società intera); piuttosto, le considerazioni che si fanno in questo tipo di dibattito da decenni valgono probabilmente intatte per l'*enhancement* (compresa la giusta osservazione “libertaria” che anche in un mercato privato il costo delle tecnologie decresce esponenzialmente al loro miglioramento).

Ci sono ovviamente altri aspetti della teoria Etica e politica che potrebbero essere influenzati dallo sviluppo tecnologico – per dirne due, il potenziamento morale (i.e. possiamo usare la tecnologia per renderci moralmente più giusti?) e l'organizzazione sociale (i.e. in una società prevalentemente virtuale, le comunità "geografiche" che ruolo e senso mantengono?). Ci riserviamo la possibilità di analizzare criticamente questi e altri aspetti in un lavoro successivo, quando la stessa letteratura scientifica e filosofica sull'argomento sarà maggiore e più chiari saranno i possibili cambiamenti da tenere in considerazione.

5. Conclusione

"*Guardo al futuro con l'inguaribile ottimismo del dodo*."

(Anonimo)

5.1 Il futuro non è più quello di una volta

Abbiamo iniziato l'introduzione chiedendoci come sarà il mondo nel 2060. In parte abbiamo dato risposte, in parte abbiamo lasciato la porta aperta a nuove riflessioni sui vari argomenti (pratici e teorici) trattati. Prevedere il futuro con precisione non è ovviamente un compito semplice, neppure per le migliori menti di ciascuna era: scenari "impossibili" diventano realtà e "certezze" non si sono mai verificate. Senza dubbio, ci sono indizi, suggerimenti diretti e indiretti, "visioni" che possono contribuire alla formulazione delle nostre ipotesi, ma la forma precisa del futuro rimane, come più volte sottolineato, di difficile comprensione.

Anche con questi *caveat*, è giusto sottolineare come non abbiamo trovato davvero un argomento KO contro la Singolarità, la quale, al termine di questo lavoro, rimane a tutti gli effetti una possibilità per lo sviluppo umano. Se la Singolarità dovesse arrivare – anche in una forma più debole di quella ipotizzata dai suoi più entusiastici sostenitori –, porterà inevitabilmente sconvolgimenti in ciascuna delle aree della conoscenza: a quel punto la domanda cruciale smetterà di essere 'come sarà il mondo nel 2006?' per diventare 'siamo pronti ai cambiamenti epocali che affronteremo nel 2060?'. L'Io come oggi ce lo rappresentiamo sta già ora cominciando a frammentarsi: dal virtuale al personale, l'Io "umano" si disgrega nell'ambiente e nelle tecnologie a disposizione. Non solo, come argomentato, non è da escludere che presto l'Intelligenza Artificiale porterà alla creazione di veri e propri Io artificiali: *mind uploading* e/o IA classica cambieranno per sempre la faccia del nostro mondo. Tra le discipline filosofiche tradizionali, l'Etica è senza dubbio quella che sarà chiamata in modo più drammatico a reinventare se stessa: lo status morale degli agenti artificiali e il problema di "rincorrere" ragioni e

responsabilità in un mondo sempre più interconnesso richiedono plausibilmente un ripensamento critico di molte assunzioni tradizionali.

Lasciando per un momento il futuro per ritornare al presente, possiamo affermare che l'essere umano di oggi, l'uomo del 2013, non è pronto per l'avvento della Singolarità. Le discussioni sul tema occupano ancora uno spazio troppo ristretto, perfino in ambienti che dovrebbero essere "specializzati". Il motivo di tanta leggerezza riteniamo si debba andare a ricercare nella difficoltà concettuale degli argomenti presentati, da una parte, e nelle "responsabilità" che derivano dalla trattazione di tali questioni: prevedere il futuro e occuparsi delle conseguenze di ciò che accade è un compito *coraggioso*, oltre che di difficile svolgimento. La società attuale va realmente preparata alla possibilità della Singolarità: sebbene il lavoro divulgativo abbia la sua importanza, è l'azione a fare la differenza. Il progetto "implicito" in I^+ ha proprio a che fare con tale obiettivo, ma a questo punto la domanda è 'come fare?'. La risposta, breve e ovvia, è semplice: l'educazione. Non stiamo parlando, è bene sottolinearlo, dei cambiamenti del sistema educativo favoriti dallo sviluppo tecnologico (si pensi, ad esempio, alla rivoluzione recente dei corsi universitari proposti online). Piuttosto, l'idea è che l'intera educazione vada ripensata in funzione di un mondo che cambia più rapidamente dei programmi scolastici attuali - un mondo in cui la tecnologia riporta verso il soggetto tutta una serie di responsabilità che, per decenni, abbiamo delegato ad altri per ragioni di tempo, costo cognitivo, etc.

A livello personale, è questo il punto che maggiormente mi preme sottolineare in conclusione di *questo* lavoro: se (*se*) gli argomenti esposti sono effettivamente efficaci nello stabilire che la Singolarità sia una possibilità concreta (ancorché magari remota), la preoccupazione principale diviene *fare* qualcosa nella società attuale affinché l'avvicinarsi al processo dia i migliori risultati possibili con il minor numero di "effetti collaterali". Una Scuola 2.0 sarà l'arma ufficiale per affrontare i cambiamenti portati dalla Singolarità: se non diventeremo "più intelligenti" e "più etici" non riusciremo in alcun modo ad avere una società migliore di quella attuale. Di contro, anche se una Singolarità alla fine non dovesse arrivare, "allenarsi" alla vita in un modo più adatto ad un mondo in continuo cambiamento non potrà che portare benefici a tutti gli individui coinvolti.

Bibliografia

Aaronson, S., 2013, *Quantum Computing since Democritus*, Cambridge: Cambridge University Press.

Antal, A., Nitsche M.A., *et al.,* 2004, 'Direct Current Stimulation over V5 Enhances Visuomotor Coordination by Improving Motion Perception in Humans', *Journal of Cognitive Neuroscience*, 16(4): 521-7.

Backstrom, M., Larsson M. R., Maddux R. E., 2009, 'A Structural Validation of an Inventory Based on the Abridged Five Factor Circumplex Model', *Journal of Personality Assessment*, 91(5): 462-472.

Bailey, C.H., Bartsch D., *et al*., 1996, 'Toward a Molecular Definition of Long-Term Memory Storage'. *PNAS*, 93(24): 13445-52.

Barazzetti, G., Reichlin, M., 2011, 'Life Extension and Personal Identity', in Kahane, Savulescu , ter Meulen (2011), pp. 398-409.

Bechtel, W., Abrahamsen A., Graham G., 2004, *Mente, Cervelli e Calcolatori*, Roma-Bari: Laterza.

Béla, N., Farmer J. D., Trancik J. E., Gonzales J. P., 2011, 'Superexponential Long-Term Trends in Information Technology'**,** *Technological Forecasting and Social Change,* 78(8): 1356-1364.

Berto, F., 2006, *Teorie dell'Assurdo: I rivali del Principio di Non-Contraddizione*, Roma: Carocci.

Bianchi, C., 2003, *Pragmatica del linguaggio*, Roma-Bari: Laterza.

Bringsjord, S., in stampa, 'The Logicist Manifesto: At Long Last Let Logic-Based Artificial Intelligence Become a Field Unto Itself', versione 9.18.08 disponibile all'indirizzo http://kryten.mm.rpi.edu/SB_LAI_Manifesto_091808.pdf.

Brooks, R. A., 1990, 'Elephants Don't Play Chess', *Robotics and Autonomous Systems*, 6: 3-15.

Burattini, E., Cordeschi R., 2004, *Intelligenza Artificiale*, Roma: Carocci.

Chalmers, D., 1996, *The Conscious Mind*, Oxford: Oxford University Press.

Chalmers, D., 2010, 'The Singularity: A Philosophical Analysis', *Journal of Consciousness Studies*, 17:7-65.

Clark, A., Chalmers D., 1998, 'The Extended Mind', *Analysis*, 58:10-23.

Craig, I., Plomin R., 2006, 'Quantitative Ttrait Loci for IQ and Other Complex Traits: Single-nucleotide Polymorphism Genotyping Using Pooled DNA and Microarrays', *Genes Brain and Behavior*, 5: 32-7.

Davidson, D., 1967, 'Truth and Meaning', *Synthese*, 17: 304-23.

Dennett, D., 1976, 'Conditions of Personhood', in A. O. Rorty (ed.), *The Identities of Persons*, University of California Press.

Dennett, D., Hofstadter D., 1976, *TheMind's I*, New York: Bantam Books.

Di Francesco, M., 1998, *L'io e i suoi Sé. Identità Personale e Scienza della Mente*, Milano: Cortina.

Di Francesco, M., 2004, 'Mi ritorni in mente. Mente Distribuita e Unità del Soggetto'. *Networks*, 3-4: 115-139.

Fodor, A. J., 2001, *La Mente Non Funziona Così*, Roma-Bari: Laterza.

Fox, P.T., Raichle M.E., *et al*., 1988, ' Nonoxidative Glucose Consumption during Focal Physiologic Neural Activity', *Science*, 241(4864): 462-4.

Frank, R., 2007, *Microeconomics and Behavior (7th Edition)*, New York: McGraw-Hill.

Fredkin, E., 1993, 'A New Cosmogony', in *PhysComp '92: Proceedings of the Workshop on Physics and Computation*, IEEE Computer Society Press, pp. 116-121.

Fregni, F., Boggio P.S., *et al.*, 2005, 'Anodal Transcranial Direct Current Stimulation of Prefrontal Cortex Enhances Working Memory', *Experimental Brain Research*, 166(1): 23-30.

Gardner, H., 1993, *Frames of the Mind. The Theory of Multiple Intelligence*, New York: Basic Books:

Giles, J., 2005, 'Internet Encyclopaedias Go Head to Head', *Nature*, 438(7070): 900-1.

Good, I. J., 1965, 'Speculations Concerning the First Ultraintelligent Machine', *Advances in Computers*, Vol. 6.

Hanson, R., 1994, 'If Uploads Come First: The Crack of a Future Dawn', *Extropy,* 6(2).

Hanson, R., Polk C., *et al*., 2003, 'The Policy Analysis Market: an Electronic Commerce Application of a Combinatorial Information Market', *ACM Conference on Electronic Commerce 2003.*

Hanson, R., 2008, 'Economics of brain emulations', in P. Healey (ed.), *Tomorrow's People*, EarthScan.

Helland, I.B., Smith L., *et al*., 2003, 'Maternal Supplementation with Very-Long-Chain n-3 Fatty Acids During Pregnancy and Lactation Augments Children's IQ at 4 Years of Age', *Pediatrics*, 111(1): 39-44.

Hobbes, T., 2004, *Leviatano*, A. Pacchi (ed.), Roma: Laterza.

Jonas, H., 2009, *Il Principio Responsabilità. Un'Etica per la Civiltà Tecnologica*, Torino: Einaudi.

Kahane, G., Savulescu J., ter Meulen R (eds.)., 2011, *Enhancing Human Capacities*, Oxford: Wiley-Blakwell.

Kennedy, P.R., Bakay R.A.E., 1998, 'Restoration of Neural Output from a Paralyzed Patient by a Direct Brain Connection', *Neuroreport*, 9(8): 1707-11.

Kripke, Saul. 1980, *Naming and Necessity*, Cambridge: Harvard University Press.

Kurzweil, R., 1999, *The Age of Spiritual Machines*, New York: Viking Press.

Kurzweil, R., 2004, *Fantastic Voyage: Live Long Enough to Live Forever*, Emmaus: Rodale Books.

Kurzweil, R., 2005, *TheSingularity is Near*, New York: Viking Penguin.

Kurzweil, R., 2008, *La Singolarità è Vicina*, Milano: Apogeo.

Kurzweil, R., 2010, 'How My Predictions Are Faring', disponibile all'indirizzo web http://www.kurzweilai.net/images/How-My-Predictions-Are-Faring.pdf.

Laratt, S., 2004, 'The Gift of Magnetic Vision', *Body Modification Ezine*.

Levin, J., 2010, 'Functionalism', in E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*.

Levy, N., 2007, 'Rethinking Neuroethics in the Light of the Extended Mind Thesis', *American Journal of Bioethics*, 7(9): 3-11.

Lewis, D. K., 1980, 'Mad Pain and Martian Pain', in Ned Block (ed.), *Readings in Philosophy of Psychology*, Cambridge: Harvard University Press, pp. 216–32.

Marconi, D., 2005, 'Contro la mente estesa', *Sistemi Intelligenti*, XVII (3): 389-398.

Marshall, L., Molle M., *et al*., 2004, 'Transcranial Direct Current Stimulation During Sleep Improves Declarative Memory', *Journal of Neuroscience*, 24(44): 9985-92.

McCarthy, J., Minsky M., Rochester N., Shannon C., 1955, 'A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence', disponibile all'indirizzo http://www-formal.stanford.edu/jmc/history/dartmouth.pdf.

Mellott, T.J., Williams C.L., *et al*., 2004, 'Prenatal Choline Supplementation Advances Hippocampal Development and Enhances MAPK and CREB Activation', *FASEB Journal*, 18(1): 545-7.

McCrae, R. R., Costa P. T., 2008, 'Empirical and Theoretical Status of the Five-Factor Model of Personality Traits', in Boyle, J., Matthews, G., Sakloske, D. H. (eds.), *The SAGE Handbook of Personality Theory and Assessment*, 273-294, SAGE.

Modis, T., 2002, 'Forecasting the Growth of Complexity and Change', *Technological Forecasting & Social Change*, 69(4).

Modis, T., 2006, 'The Singularity Myth', *Technological Forecasting & Social Change,* 73(2).

Moore, G. E., 1965, 'Cramming More Components onto Integrated Circuits', *Electronics Magazine*.

Monti, M. M., Laureys S., Owen A.M., 2010, 'Diagnosing the Vegetative State'. *BMJ*, 341(c3765): 292-296**.**

Monti, M. M., Vanhaudenhuyse A., Coleman M.R., Boly M., Pickard J., Tshibanda J-F., Owen A. M., Laureys S., 2010, 'Willful Modulation of Brain Activity in Disorders of Consciousness', *New England Journal of Medicine*, 362: 579-589.

Nitsche, M.A., Schauenburg A., *et al*., 2003, 'Facilitation of Implicit Motor Learning by Weak Transcranial Direct Current Stimulation of the Primary Motor Cortex in the Human', *Journal of Cognitive Neuroscience*, 15(4): 619-26.

Olson, E. T., 1999, *The Human Animal: Personal Identity Without Psychology*, Oxford: Oxford University Press.

Oppenheimer, P. E., Zalta E. N., 1991,'On the Logic of the Ontological Argument', in James Tomberlin (ed.), *Philosophical Perspectives 5: The Philosophy of Religion*, pp. 509-529.

Parfit, D., 1984, *Reasons and Persons*, Oxford: Oxford University Press.

Penrose, R., 2000, *La mente nuova dell'imperatore*, Milano: Rizzoli.

Popper, K., 1963, *Conjectures and Refutations: the Growth of Scientific Knowledge*, London: Routledge.

Putnam, H., 1967, 'The Nature of Mental States', in W. H. Capitan, D. D. Merrill (eds.), *Art, Mind, and Religion*, University of Pittsburgh Press, pp.37-4.

Raymond, E.S., 2001, *The Cathedral and the Bazaar,* New York: O'Reilly.

Rusted, J.M., Trawley S., *et al*., 2005, 'Nicotine Improves Memory for Delayed Intentions', *Psychopharmacology*, 182(3): 355-65.

Sandberg, A., 2010, 'An Overview of Models of Technological Singularity', disponibile all'indirizzo web agi-conf.org/2010/wp-content/uploads/2009/06/agi10singmodels2.pdf.

Sandberg, A., Bostrom N., 2008, 'Whole Brain Emulation: A Roadmap', *Technical Report #2008‐3*. Future of Humanity Institute, Oxford University.

Sandberg, A., Savulescu J., 2011, 'Social and Economic Impacts of Cognitive Enhancement', in Kahane, Savulescu , ter Meulen (2011), pp. 92-112.

Savulescu, J., Bostrom, N., 2011, *Human Enhancement*, Oxford: Oxford University Press.

Searle, J., 1998, *Mind, Language And Society: Philosophy In The Real World*, New York: Basic Books.

Smart, J. J. C., 1959, 'Sensations and Brain Processes', *Philosophical Review*, 68: 141-156.

Sunram-Lea, S.I., Foster J.K., *et al*., 2002, 'Investigation into the Significance of Task Difficulty and Divided Allocation of Resources on the Glucose Memory Facilitation Effect', *Psychopharmacology*, 160(4): 387-97.

Tainter, J., 1988, *The Collapse of Complex Societies*, Cambridge: Cambridge University Press.

Thaler, R. H., Sunstein C. R., 2009, *Nudge: Improving Decisions about Health, Wealth, and Happiness*, New York: Penguin Books.

Tieges, Z., Richard Ridderinkhof K., *et al*., 2004, 'Caffeine Strengthens Action Monitoring: Evidence From the Error-Related Negativity', *Brain Research and Cognitive Brain Research*, 21(1): 87-93.

Tononi, G., 2008, 'Consciousness as integrated information: a provisional manifesto', *The Biological Bulletin*, 215(3): 216-242.

Turing, A. M., 1950, 'Computing Machinery and Intelligence', *Minds and Machines*, 59: 433-460.

Watson, J. B., Rayner, R., 1920, 'Conditioned emotional reactions', *Journal of Experimental Psychology*, 3(1), pp. 1-14.

Wheeler, J. A., 1990, 'Information, Physics, Quantum: The Search for Links', in W. Zurek (ed.), *Complexity, Entropy, and the Physics of Information*, Addison-Wesley, pp. 309-336.

Wiener, N., 1948, *Cybernetics: Or Control and Communication in the Animal and the Machine,* Cambridge: MIT Press.

Williams, B., 1970, 'The Self and the Future', *Philosophical Review*, 79(2): 161-180.

Williams, B., 1973, 'The Makropulos Case: Reflections on the Tedium of Immortality', in B. Williams (ed.), *Problems of the Self,* Cambridge University Press, pp. 82-100.

Wilson, M., 2002, 'Six View of Embodied Cognition', *Psychonomic Bulletin & Rev*, 9(4): 625-36.

Wolfram, S., 2002, *A New Kind of Science*, Champaign: Wolfram Media.

Zenil, H., 2013, ed., *A Computable Universe*, Singapore: World Scientific Publishing.

Zuse, K., 1982, 'The Computing Universe', *International Journal of Theoretical Physics*, 21: 589-600.