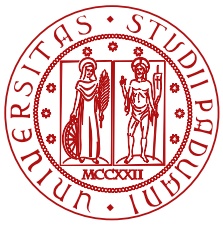

UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA MAGISTRALE IN INGEGNERIA INFORMATICA

Aspetti computazionali della Integrated Information Theory (IIT) della coscienza

Relatore
Prof. Bilardi Gianfranco

Laureando
Lacini Ralton

ANNO ACCADEMICO 2023-2024

Data di laurea 11/04/2024

Grazie Elettra per avermi ~~sopportato~~ supportato.

Sommario

Lo studio della coscienza è un campo di ricerca multidisciplinare che coinvolge diversi rami della scienza e della filosofia. Esplora i meccanismi alla base dell'esperienza cosciente, e indaga come l'esperienza soggettiva possa emergere da processi fisici. Considerata per molto tempo esclusivamente una questione filosofica, negli ultimi anni è diventata oggetto di un'intensa indagine scientifica. La Integrated Information Theory (IIT) è una delle principali teorie scientifiche sulla coscienza. Sviluppata dal neuroscienziato Giulio Tononi e collaboratori, cerca di spiegare le proprietà che il cervello umano e altri sistemi fisici devono soddisfare per essere considerati coscienti. La teoria illustra come un sistema fisico, modellato come una rete di componenti interconnessi, mostri una capacità di integrare informazione. Viene presentato un quadro matematico e definito una misura chiamata Phi (Φ), per quantificare il grado di coscienza e l'esperienza particolare che il sistema sta avendo. Questa tesi esplora gli aspetti matematici e computazionali della teoria ed elabora i suoi limiti computazionali.

Indice

Capitolo 1

Introduzione

1.1 Lo studio della coscienza

Lo studio della coscienza è un campo di ricerca multidisciplinare che coinvolge aree come la filosofia, la psicologia, le neuroscienze, la fisica e l'informatica. Una definizione esatta è difficile da formulare, tuttavia ognuno di noi possiede un'idea intuitiva su cosa sia. È l'abilità che possediamo noi esseri umani di riflettere su noi stessi, di percepire il mondo in prima persona, la consapevolezza che abbiamo di esistere. "Cogito ergo sum" di Cartesio significa letteralmente "penso dunque sono", e "sono" significa essere cosciente. In altri contesti la coscienza è l'abilità di percepire il mondo intorno a noi, di provare esperienze soggettive e di attribuirne significato. È quella *cosa* che perdiamo quando dormiamo, oppure siamo sotto anestesia totale, e riacquistiamo una volta svegli.

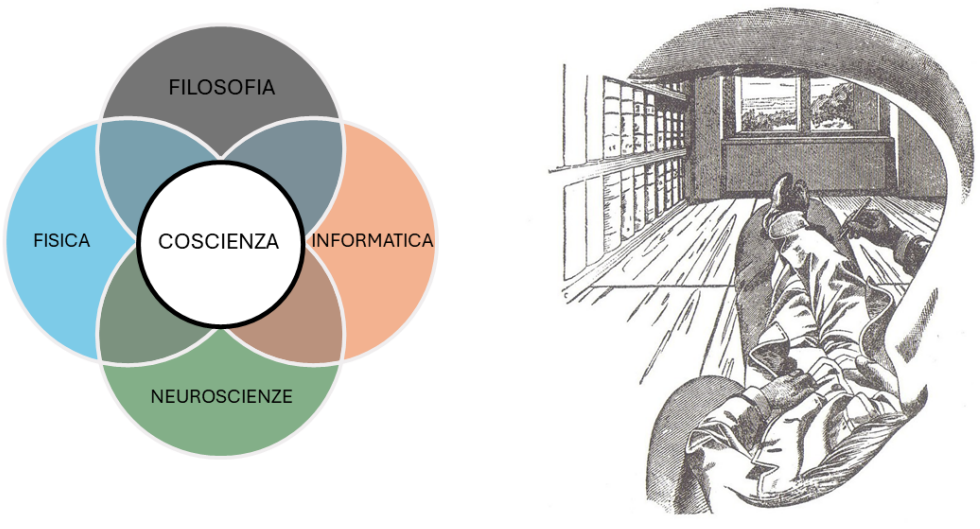

Figura 1.1
A sinistra, la coscienza come un campo di ricerca multidisciplinare. A destra, l'opera "visione dall'occhio sinistro" di Mach, spesso usata per discutere questioni legate alla percezione, alla prospettiva soggettiva e alla coscienza nel contesto della filosofia della scienza.

Sembra che la coscienza sia una proprietà emergente dell'attività celebrale. Il filosofo australiano David Chalmers definisce il problema "facile" e il problema "difficile" della coscienza [1]. Il problema "facile" della coscienza riguarda il modo in cui il cervello e i suoi meccanismi neurali ci abilitano a eseguire funzioni cognitive, acquisire ed elaborare informazioni, e generare risposte comportamentali. Ad esempio, come avviene l'elaborazione sensoriale, ossia come il cervello trasforma i segnali luminosi captati dagli occhi in immagini visive? Il secondo problema viene definito "difficile", perché anche se riuscissimo a spiegare completamente tutti i meccanismi neurali, non avremmo comunque una risposta dell'origine delle sensazioni qualitative, o "qualia". Ad esempio, anche se potessimo descrivere in dettaglio come il cervello elabora le immagini visive, resterebbe il quesito su come e perché questa elaborazione si traduce nell'esperienza vissuta del colore rosso, che è diversa per ciascun individuo e intrinsecamente legata alla propria coscienza.

Lo studio della coscienza non è fine solo ad una migliore comprensione dell'esperienza umana, ma risulta essere importante per le ricerche delle scienze cognitive e del comportamento, in ambito medico terapeutico, etico-sociale ed infine nell'ambito dell'intelligenza artificiale. Inoltre, non è ancora del tutto chiaro il ruolo della coscienza nell'interpretazione della meccanica quantistica e del problema di misurazione e del collasso della funzione d'onda. Gli studi sulla coscienza hanno radici antiche filosofiche e per molti anni non sono stati disponibili mezzi scientifici, come l'elettroencefalografia (EEG) e la risonanza magnetica funzionale (fMRI), per comprendere e studiare il problema. Con i recenti sviluppi tecnologici stiamo assistendo ad una transizione da una questione puramente filosofica ad una vera e propria disciplina scientifica. Sotto questo aspetto esistono molteplici programmi di ricerca che sono sfociati in diverse teorie per cercare di fare luce sull'argomento.

Capitolo 2

La Integrated Information Theory

2.1 Panoramica della teoria

Tra le diverse teorie [2] si analizzata la Teoria dell'Informazione Integrata della coscienza (IIT - Integrated Information Theory) [3] del neuroscienziato Giulio Tononi e collaboratori. L'unicità di questa teoria è che propone un modello matematico dettagliato e preciso che mira a spiegare come la coscienza possa emergere da un sistema fisico, quantificare il grado di coscienza e descrivere la struttura (in tremini di "qualia") dell'esperienza che sta avendo il sistema fisico in quello stato.

Nel caso del cervello umano spiega ulteriormente come alcune aree del cervello possano contribuire all'insorgenza della coscienza ed altre no, e come questa svanisca nei casi di sonno profondo, anestesia o convulsioni. Il cervelletto, per esempio, costituisce circa il 10% della massa cerebrale e contiene circa il 50% dei neuroni cerebrali, ma non ha la struttura adeguata per ospitare la conoscenza, mentre la corteccia celebrale posteriore sì.

Proposta nel 2004 ha subito diverse iterazioni e l'ultima pubblicazione, con la versione più recente della teoria, è del 17 ottobre 2023. Si noti che ogni riferimento alla "teoria", all' "IIT" o alla "pubblicazione" si rivolge esclusivamente a quest'ultima [3]. La teoria propone che un sistema interconnesso di unità può essere cosciente solo se questo possiede l'abilità di "integrare informazione", e la qualità dell'esperienza che sta avendo è identica ad una opportuna struttura chiamata Φ-struttura, ottenuta dal sistema sotto analisi.

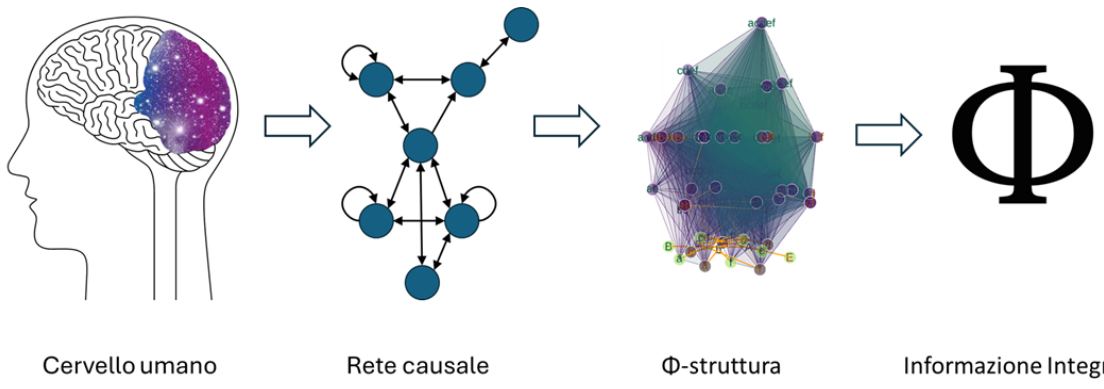

Figura 2.1
Passaggi dell'IIT, dal substrato fisico al calcolo di Φ

2.1.1 Premessa

Si suppone una familiarità con l'IIT e l'analisi segue fedelmente gli sviluppi della teoria come presentati nella pubblicazione più recente.
Alcune immagini, usate in questa tesi, provengono da pubblicazioni parallele e conferenze degli stessi autori dell'IIT, dunque la proprietà intellettuale appartiene agli stessi e il loro utilizzo in questa tesi è inteso esclusivamente a fini accademici e di ricerca, sotto il principio di "fair use".

Durante il suo percorso la teoria ha subito diverse critiche, tra cui Aaronson [2], Jake R Hanson [4] , Michael A. Cerullo [5], mentre più recentemente è stata etichettata come pseudoscienza in una lettera aperta firmata da più di 100 ricercatori sulla coscienza [6] .

Questa tesi non affronta gli aspetti filosofici e le loro implicazioni, ma si concentra sull'analisi del modello matematico proposto, sul calcolo della complessità computazionale, oltre che di spiegarla, nella speranza di renderla più accessibile, illustrandola con immagini ed esempi.

2.2 Metodologia

L'IIT, a differenza di altre teorie sulla coscienza, non comincia lo studio a partire da fattori comportamentali, funzionali o correlati all'attività neurali per cercare di spiegare come questi diano vita alla coscienza. Tradizionalmente, in campi come la fisica, l'indagine inizia con l'osservazione di fenomeni naturali da cui si sviluppano modelli matematici. Questo risulta difficile nello studio della coscienza poiché essa stessa rappresenta il fenomeno osservativo e l'esperienza da spiegare.

Per l'IIT essere cosciente vuol dire avere un'esperienza, quindi si useranno queste due nozioni interscambiabilmente. Quello che si fa in questo caso è di partire dalla fenomenologia dell'esperienza stessa e di cercare di delineare quali siano le proprietà essenziali ed irrefutabili che si possono assumere, gli assiomi della coscienza. Alcuni di questi principi che la teoria suggerisce possono apparire discutibili, ma questo va oltre lo scopo di questa tesi. Può tuttavia essere utile una riflessione su di esse notando che la loro negazione porterebbe a situazioni logicamente inconsistenti o paradossali. Vengono dunque accettate come punto di partenza, elencate e brevemente spiegate. Viene inoltre spiegato come esse vengono tradotte in postulati fisici.

2.3 Assiomi fenomenologici e postulati di esistenza fisica

0 - **Esistenza.** Qualcosa esiste. L'esperienza corrente esiste. Qualcosa esiste al di fuori della nostra esperienza. Esistere significa avere potere di causa ed effetto. Un'unità fisica esiste se ha il potere di produrre un effetto su altre unità e subire l'effetto di altre unità. Ne consegue

Figura 2.2
Assiomi dell'IIT con immagini ispirate all'opera di Mach.

l'esistenza di unità che si influenzano a vicenda, chiamati substrato di coscienza. Il concetto di "unità" si riferisce a qualcosa che possiamo assumere come atomico, senza struttura interna, che possiede uno stato, che possiamo osservare e manipolare. In questa astrazione possiamo pensare alle unità, facendo un parallelo con il cervello umano, come ad aree del cervello, strutture neurali, singoli neuroni, singole proteine, atomi o particelle elementari. In linea di principio ogni granularità del substrato deve essere considerata.

1 - **Intrinsecità.** L'esperienza è intrinseca a se stessa, all'essere cosciente. Il potere di causa-effetto deve essere intrinseco. Il substrato deve esercitare il potere di causa-effetto all'interno del substrato stesso.

2 - **Informazione.** L'esperienza corrente è specifica e informativa. È quella che si sperimenta in questo momento e differisce da ogni altra esperienza possibile. Il potere di causa-effetto che il substrato esercita deve essere specifico allo stato attuale e deve scegliere uno stato futuro del substrato e deve essere stato scelto da uno stato precedente.

3 - **Integrazione.** L'esperienza è unica e non riducibile alle sue parti. Deve scegliere gli stati di causa ed effetto come un tutt'uno, irriducibile a unità o gruppi di unità indipendenti. La quantità φ (informazione integrata) misura quanto il substrato è irriducibile rispetto alle sue parti.

4 - **Esclusione.** L'esperienza è definita, ha dei "bordi". L'esperienza visiva per esempio non si estende oltre il campo visivo delimitato. Il potere di causa-effetto deve anch'esso essere

definito esattamente dalle unità verso le quali viene esercitato.

5 - **Composizione.** L'esperienza, pur essendo unica, ha una sua struttura. È composta da sottoinsiemi di elementi che esercitano causa-effetto su altri sottoinsiemi dello stesso substrato (distinzioni), che "sovrapponendosi" con altri, formano delle relazioni. Per esempio: *vedo un libro blu. Il libro è sul tavolo*. Le distinzioni e le relazioni devono essere a loro volta irriducibili, e ad esse viene associato un valore di informazione integrata. La totalità delle distinzioni e relazioni formano una struttura di causa-effetto detta Φ-struttura.

La IIT propone la seguente identità tra la Φ-struttura e la qualità dell'esperienza. Ogni proprietà dell'esperienza è contenuta dall'interno della struttura. C'è una corrispondenza uno ad uno tra il modo in cui l'esperienza viene percepita e il modo in cui la Φ-struttura è composta. Inoltre, la somma dei valori dell'informazione integrata di distinzioni e relazioni è chiamata informazione integrata strutturale, viene indicata con la lettera Φ e corrisponde alla quantità dell'esperienza, della coscienza.

Non è facile avere una soddisfacente intuizione di questa identità e sul perché la coscienza possa essere spiegata interamente in questi termini. Se accettiamo che l'esistenza è l'abilità di esercitare causa-effetto, considerando ogni fonte di causa-effetto all'interno del sistema possiamo concludere di non aver tralasciato nulla. In questo senso non c'è più bisogno di altri ingredienti nella teoria.

È curioso osservare che l'ordine in cui vengono applicati gli assiomi e i postulati è guidato da una scelta ragionevole e naturale, ma non unica. Nella versione 3.0 dell'IIT, infatti, l'ordine in cui essi intervenivano era in ordine : esistenza, intrinsecità, composizione, informazione, integrazione, esclusione.

2.4 Principi ontologici

Oltre agli assiomi, l'IIT offre tre principi ontologici che intervengono in diversi punti durante l'analisi della teoria.

1 - **Principio dell'essere.** Questo si ricollega all'assioma zero e al principio realista della teoria. "Essere" o "esistere" significa avere potere di causa ed effetto. Inoltre "c'è un mondo la fuori", qualcosa esiste e persiste indipendentemente dalla nostra coscienza.

2 - **Principio dell'esistenza massima.** Questo principio viene offerto dall'IIT semplicemente come una buona spiegazione per determinare che, quando si tratta di esistenza, o potere di causa-effetto, esiste quel sottoinsieme di unità che esercita maggiormente questo potere.

3 - **Principio dell'esistenza minima.** Questo principio viene offerto dall'IIT per spiegare che, quando si tratta di requisiti per l'esistenza, nulla può esistere più di quello che esiste di meno. Si pensi al postulato dell'integrazione. Se un sistema è poco integrato, oppure non integrato, detto anche riducibile, allora il suo potere di causa-effetto è basso o nullo, e dunque anche la sua

esistenza è poca o nulla. Il principio stabilisce dunque che questa esistenza minima è l'unica esistenza.

In riferimento ai principi ontologici due e tre, nonostante a questo punto possano risultare non chiarissimi, la loro applicazione ha un'importanza fondamentale sulla direzione di sviluppo della teoria. Si vedrà successivamente che esse risulteranno essere delle scelte ragionevoli e naturali, pur rimanendo aperte a discussioni ed approfondimenti.

Capitolo 3

Struttura matematica

Per stabilire un punto d'inizio della teoria occorre implementare i postulati di esistenza fisica in termini operativi. Partendo dall'assioma zero, il substrato (sistema) più elementare che possiamo pensare è quello di due o più unità che si influenzano reciprocamente ed hanno anche la capacità di auto-influenzarsi.

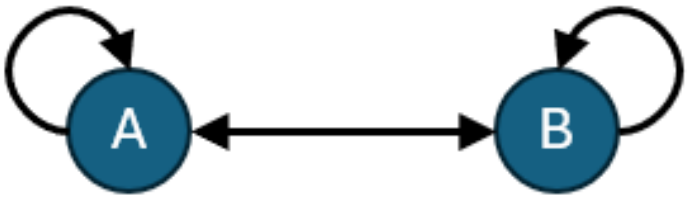

Figura 3.1
Semplice sistema dinamico di due unità.

Da questo punto in avanti i termini "potere di causa-effetto", "fare una differenza", "influenzare", "causare" vengono usati interscambiabilmente come sinonimi. Per quanto riguarda lo studio della causalità, il nesso tra causa ed effetto in termini matematici, non c'è tutt'oggi una convergenza unica [7].

Generalmente vengono usati gli strumenti della teoria delle probabilità per caratterizzare la relazione tra causa ed effetto, con l'idea che le cause aumentino le probabilità dell'effetto, a parità di tutte le altre condizioni. Con riferimento alla relazione

$$P(\, effetto \mid evento\,) > P(\, effetto \mid \neg evento\,)$$

ossia la probabilità di un effetto è maggiore quando accade un evento rispetto a quando non

accade, questo può suggerire che l'evento accaduto sia la causa dell'effetto. Tuttavia, se si pensa ad un semplice esempio dove si vuole prevedere la probabilità di pioggia (effetto), monitorando i dati dell'igrometro, è vero che, quando quest'ultimo segna un valore alto, la probabilità di pioggia aumenta, ma non si può dedurre che l'evento della misurazione alta dell'igrometro causi la pioggia. Questo perché la relazione $P(A|B)$ è di tipo osservazionale, mentre Pearl argomenta che la relazione giusta da considerare sia invece

$$P(\ effetto \mid do(evento)\ ) > P(\ effetto \mid do(\neg evento)\ )$$

a parità di tutte le altre condizioni[8] [9] [10, p. 53]. L'operatore "do" rappresenta un intervento nel sistema, l'imposizione dell'evento ad un particolare valore. Effettivamente, manipolando manualmente l'igrometro, questo non ha nessun effetto sulla pioggia, dunque la relazione viene violata.

Gli interventi o imposizioni sul sistema, uniti ai modelli grafici probabilistici, come le reti causali di Bayes (causal Bayes networks), si sono dimostrati efficaci per sviluppare e comprendere la causalità. Le reti causali di Bayes sono un sottoinsieme delle reti di Bayes, e si distinguono da quest'ultime per il senso degli archi orientati nel DAG (directed acyclic graph - grafo diretto aciclico). Nel caso delle reti causali, l'arco diretto indica una relazione di causa-effetto rispetto alla sola correlazione nelle reti di Bayes non causali.

3.1 Panorama generale ed assunzioni

Tornando all'esempio di un semplice substrato di coscienza o sistema in un determinato stato, siamo interessati alla sua evoluzione, e più precisamente a transizioni di stato a tempi discreti, da $t-1$, t e $t+1$. Come nel caso della granularità di quello che si considera "unità", una cosa simile si applica anche alla transizione temporale di stato. Potenzialmente ogni intervallo temporale ragionevole deve essere considerato. Per esempio, nel caso si considerino i neuroni come unità elementari, può risultare ragionevole studiare la transizione, in linea con i tempi delle trasmissioni nervose e sinaptiche, nell'ordine dei millisecondi.

Considereremo inoltre le seguenti assunzioni :

- Le unità sono in numero finito e lo spazio dei loro stati è finito. Uno spazio finito di unità può essere usato per approssimare localmente uno spazio molto più vasto, potenzialmente infinito. La stessa approssimazione viene fatta anche per lo spazio degli stati, ossia i possibili valori che ogni unità può assumere. Nell'esempio dello studio del cervello, non c'è ragione plausibile di considerare nel sistema, l'insieme di tutte le cellule del corpo umano, o addirittura

gli elementi che compongono l'ambiente circostante. Per semplicità, in alcuni passaggi di questa tesi, viene assunto che lo spazio degli stati per ogni unità sia binario.

- **Il sistema transita di stato in tempi discreti.** Per gli stessi motivi elencati sopra, i sistemi a stati discreti valutati in diverse granularità temporali possono essere delle buone approssimazioni di quelli a tempo continuo. Inoltre si assume che tutte le transizioni di stato delle unità siano già avvenute nell'intervallo temporale che si considera.

- **Il sistema è tempo invariante.** Questo significa che le probabilità di transizione da uno stato all'altro non dipendono dal tempo in cui avviene la transizione. Questa assunzione permette di semplificare la trattazione teorica del processo stocastico e rispecchia molti sistemi fisici e biologici.

La terminologia e la notazione usata per descrivere i passaggi successivi rispecchia esattamente quella usata nella pubblicazione. Il numero accanto alle varie formule segue anch'esso la pubblicazione per una corrispondenza uno ad uno.

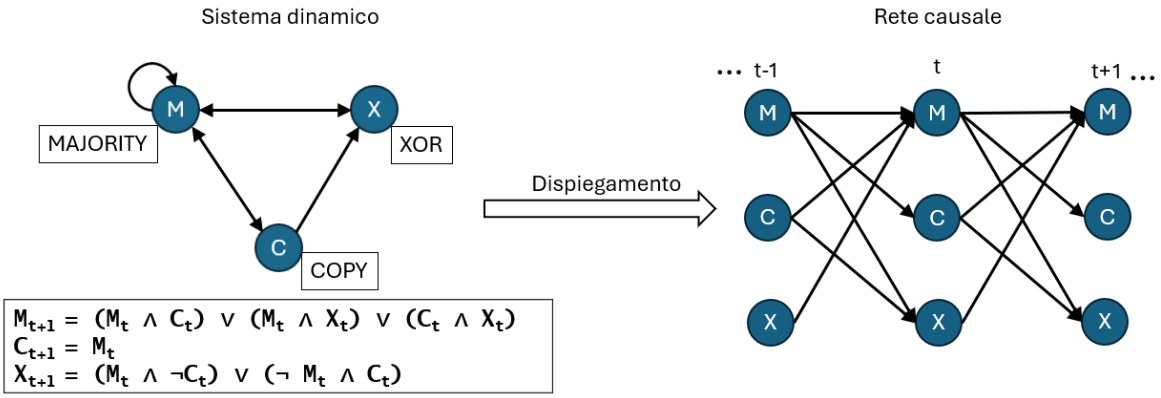

Figura 3.2

Un substrato (detto anche sistema o universo) $U = (U_1, U_2, ..., U_n)$ può dunque essere pensato come un insieme finito di $|U| = n$ unità che interagiscono, che possono essere osservate e manipolate. Lo stato u_i di ogni unità U_i è influenzato esclusivamente dalle interazioni con altre unità, e in questo modello viene rappresentata da una variabile aleatoria. Mentre nell'immagine sopra, lo stato di ogni ogni unità a tempo $t + 1$ è completamente determinato dallo stato delle unità a tempo t, ci si potrebbe chiedere la sorgente dell'alea in un sistema simile. Mentre la questione rimane fondamentalmente aperta, una possibile spiegazione ragionevole potrebbe essere la seguente. Il modello matematico non rappresenta esattamente il sistema fisico, ma una sua approssimazione. Queste approssimazioni risultano dalle assunzioni sul numero finito delle unità, dei loro stati, dalla finestra temporale nella quale le iterazioni si considerano "concluse",

oppure dalla natura intrinsecamente probabilistica della realtà più profonda descritta dalla fisica quantistica.

Con un abuso di notazione si intende per U sia l'insieme delle unità che compongono l'universo, inteso come l'insieme di tutte le unità che interagiscono, sia il vettore di variabili aleatorie $(U_1, U_2, ..., U_n)$. Indicheremo con Ω_{U_i} lo spazio degli stati dell'unità U_i, e con $\Omega_U = \Omega_{U_1} \times \Omega_{U_2} \times \cdots \times \Omega_{U_n} = \prod_i \Omega_{U_i}$ lo spazio degli stati dell'intero sistema. Indicheremo con $u \in \Omega_U$ lo stato attuale del sistema.

Il modello grafico probabilistico per rappresentare l'evoluzione temporale è quello delle reti causali di Bayes, e si ottiene dal dispiegamento temporale (unfolding in inglese) del sistema sotto esame. Successivamente, quando parleremo di rete causale, o DAG, faremo riferimento a questo modello grafico. I nodi del DAG rappresentano le unità, o equivalentemente, le variabili aleatorie a tempi discreti, mentre gli archi rappresentano relazioni di causalità.

3.2 Calcolo delle matrici di transizione probabilistiche

Indagheremo il sistema solo in due momenti diversi, e considereremo solo le transizioni da $t-1$ a t e da t a $t+1$, e useremo lettera minuscola per rappresentare lo stato delle unità al tempo corrente; mentre, specificato a seconda del contesto, si userà la stessa lettera minuscola con una barra sopra per indicare lo stato delle stesse unità nel momento successivo o quello precedente. Ad ogni intervallo temporale, dunque, le variabili aleatorie associate alle unità, vengono considerate diverse, ossia U_i a tempo t è diversa da $\bar{U}_i$ a tempo $t+1$. Siamo dunque interessati ad avere una completa descrizione del sistema mediante la sua matrice di transizione chiamata TPM (transition probability matrix in inglese), ovvero la probabilità di transitare dallo stato corrente u a quello futuro $\bar{u}$ per ogni u,$\bar{u}$. La TPM di U viene definita:

$$T_U = p(\bar{u} \mid u) \qquad u,\ \bar{u} \in\ \Omega_U \tag{1}$$

Viene inoltre assunta soddisfatta la proprietà locale di Markov, dove lo stato di una qualsiasi unità è indipendente da quello delle altre (esclusi i sui figli), dato lo stato dei suoi diretti genitori

(parents in inglese). Questo ci permette di fattorizzare le probabilità condizionali in (1) come

$$p(\bar{u}|u) = \prod_{i=1}^{|U|} p(\bar{u_i} \mid pa(\bar{u_i}))$$

Per genitori $pa(u_i)$ di u_i, si intende lo stato delle unità che sono la causa dell'unità u_i, oppure l'effetto delle quali u_i prende, oppure gli archi entranti del nodo u_i usando i termini della teoria dei grafi. Più in generale, i genitori di ogni unità al tempo $t+1$, sono un sottoinsieme di tutte le unità al tempo t, quindi le $\bar{u_i}$ sono indipendenti tra di loro dato lo stato attuale u. Possiamo dunque scrivere la formula (1) come :

$$p(\bar{u}|u) = \prod_{i=1}^{|U|} p(\bar{u_i}|u) \tag{2}$$

Questa assunzione è alla base delle reti di Bayes, e consente di rendere trattabili problemi dove il calcolo delle probabilità congiunte aumenta in modo esponenziale rispetto al numero delle variabili aleatorie. L'intuizione dietro questa assunzione è la seguente: una volta determinate tutte le cause di fenomeno, la conoscenza di altri fattori, che non siano effetti di esso, non influiscono sulla probabilità di quest'ultimo di verificarsi. Un esempio potrebbe essere il seguente modello causale.

In questo modello l'esito della pioggia è correlata con lo stato delle nuvole, la percentuale di umidità dell'aria, e dall'osservazione o meno di fulmini. In particolare, nonostante non siano i fulmini a causare la pioggia, una loro osservazione aumenta la probabilità che piova e viceversa. Una volta determinato però lo stato delle nuvole e quello dell'umidità sapere se ci sono fulmini oppure no, non influisce più sulla probabilità di pioggia. Lo stesso vale anche per i fulmini. Dato lo stato di nuvolosità e umidità, fulmini e pioggia diventano eventi indipendenti.

Al contrario invece, la pioggia continua a dipendere dai suoi effetti. Sapere che c'è il terreno bagnato cambia la probabilità di pioggia, anche nel caso si conoscano le sue dirette cause.

Si assume dunque che possiamo calcolare le probabilità condizionate delle singole unità in (2) e combinarle per avere una completa descrizione del sistema mediante la sua TPM. Dal punto di vista operativo, le singole probabilità $p(\bar{u_i}|u)$ si calcolano intervenendo sul sistema, forzando lo stato delle unità in ognuno dei possibili stati, e annotando la probabilità di $\bar{u_i}$. Matematicamente questa operazione viene indicata come $p(\bar{u_i}|do(u))$, dove $do(u)$ indica che

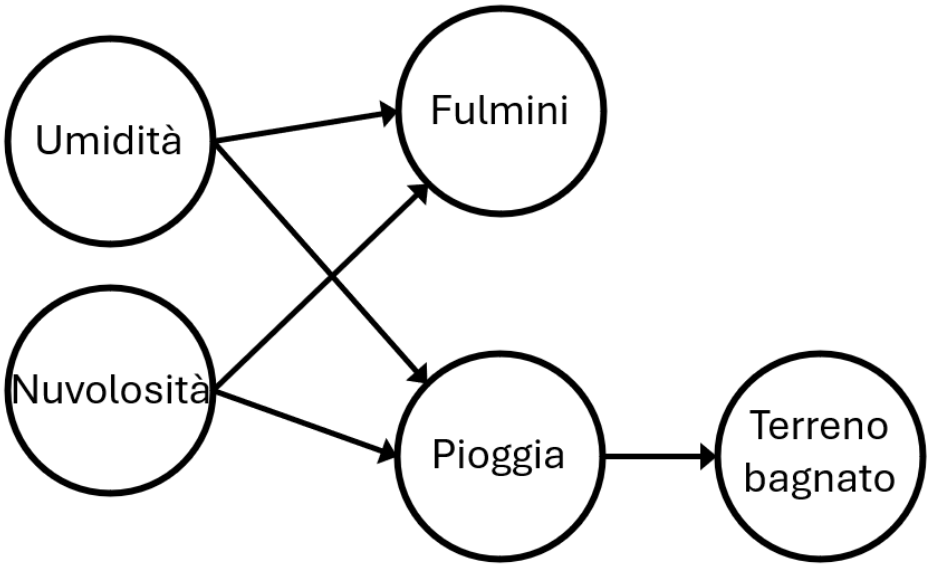

Figura 3.3

lo stato u è stato imposto per intervento. Importante notare, dato la natura causale del sistema, che le due probabilità, quella osservazionale $p(\bar{u}_i|u)$ e quella interventistica $p(\bar{u}_i|do(u))$ coincidono.[10, pp. 143–145]

Considerando nel caso peggiore di un grafo completo con $|U| = n$ unità con valori binari, la complessità di calcolare una singola probabilità viene determinata dal numero esponenziale dello spazio degli stati $|\Omega_U| = 2^n$. Anche nel caso in cui il sistema sia deterministico, dunque le probabilità in questione sono 0 oppure 1, la complessità di questa operazione rimane invariata.

Un aspetto fondamentale che non emerge in modo esplicito dalla pubblicazione è quella della probabilità marginale di u, $p(u)$. Può risultare banale che in questo caso essa sia uniforme ragionando sulla natura interventistica secondo la quale abbiamo imposto u in tutti i valori possibili, ma come vedremo, questa acquista enorme valore quando parleremo di transizioni di "causa", ossia dallo stato $t-1$ a t.

3.2.1 Identificazione dei substrati di coscienza

Il substrato U nello stato u con associata TPM T_U, rappresenta il punto d'inizio dell'analisi dell'IIT. In questa prima parte ci si chiede se U, oppure un qualsiasi suo sottoinsieme S, possa soddisfare tutti i postulati dell'IIT.

Durante lo sviluppo seguente U rappresenta "l'universo" di tutte le unità. Si parlerà dunque di universo, di substrato (potenziale) di coscienza, oppure semplicemente di "sistema". Ogni sottoinsieme S di U invece verrà chiamato anche "sottosistema", "candidato o potenziale complesso", oppure, laddove sia chiaro dal contesto, semplicemente "sistema". L'insieme di unità $W = U \backslash S$ [1] prende il nome di "condizioni di fondo" (background conditions in inglese).

Viene dunque proposta una ricerca esaustiva tra tutti i sottoinsiemi S di U, e per ognuno di essi si verifica che gli assiomi e i postulati della teoria vengono soddisfatti. Per valutarne l'integrazione verrà introdotta una metrica chiamata "informazione integrata" o φ. Quello con il valore più alto (per il principio uno dell'esistenza massimale) viene chiamato substrato massimale, oppure complesso. Per il postulato dell'esclusione che delimita i bordi dell'esperienza, anche il complesso deve avere bordi definiti, e quindi se più potenziali complessi concorrono sulle stesse unità, per il principio dell'esistenza massimale, esiste solo il complesso con il valore più alto dell'informazione integrata; mentre gli altri candidati complessi vengono esclusi dall'esistenza. La ricerca esaustiva procede dunque alla ricerca di altri complessi nella parte restante del sistema.

Spostiamo dunque la nostra analisi su come la teoria si sviluppa, considerando un sottosistema S, nello stato s parte di un universo di unità U nello stato u, e condizioni di fondo W nello stato w. Ricordiamo che S è un sottoinsieme di U e $W = U \backslash S$. Lo sviluppo della teoria analizza dunque il sottosistema S, chiedendosi in successione, se soddisfa le condizioni imposte in linea con gli assiomi e postulati:

Esistenza. Il sistema esiste solo se ha potere di causa-effetto. Potere di causa viene inteso come la quantificazione numerica di quanto uno stato precedente del sistema aumenti la probabilità dello stato attuale rispetto al caso; mentre potere di effetto, viene inteso come la quantificazione numerica di quanto lo stato attuale del sistema aumenta la probabilità di uno stato futuro rispetto al caso. La quantificazione numerica del potere di causa-effetto viene chiamata "informatività causale" (causal informativeness in inglese). Un concetto reminiscente della teoria classica dell'informazione, intesa come la riduzione dell'incertezza rispetto al caso. Dal punto di vista intuitivo, considerando per esempio solo le transizioni di effetto dello stato s dal tempo t a $\bar{s}$ a tempo $t + 1$, ci si chiede il valore del rapporto $\frac{p(\bar{s}|s)}{p(\bar{s})}$ dove $p(\bar{s}|s)$ è la probabilità dello stato futuro dato lo stato corrente, mentre $p(\bar{s})$ è la probabilità marginale dello stato futuro, ossia la sua probabilità in assenza di s. Questa è una componente di una misura di "distanza", o "differenza" tra due distribuzioni di probabilità, che verrà introdotta dopo, chiamata ii, "informazione intrinseca" (intrinsic information in inglese).

Intrinsecità. Ci si chiede se questo potere di causa-effetto è intrinseco al sistema S stesso. Questo significa concentrare l'analisi solo sul sistema S e rendere le unità di W, le condizioni

[1] L'operatore "\" denota la differenza insiemistica, che diventa il complemento quando, come nel caso presente, $S \subseteq U$.

di fondo, “causalmente inerti” in modo che non possano contribuire direttamente al potere di causa-effetto.

Informazione. Questo potere di causa-effetto deve essere informativo e specifico. Lo stato attuale deve scegliere uno tra gli stati futuri verso il quale esercita il maggior effetto, ed uno stato passato che esercita il maggior effetto su di esso.

Integrazione. Ci si chiede quanto il sistema sia “integrato”, ovvero quanto esercita il potere di causa-effetto come una sola unità, e non come somma delle sue parti. L’idea intuitiva dietro al concetto di integrazione può essere ricavata pensando al semplice esempio mostrato in figura.

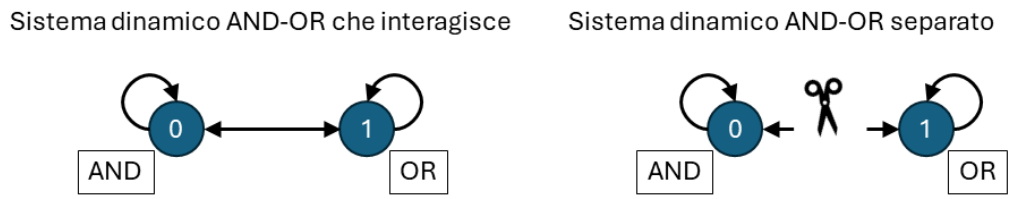

Figura 3.4

Se si pensa ad un sistema di unità come due ipotetiche porte logiche AND e OR, che si influenzano a vicenda, hanno effetto anche su loro stesse ed hanno uno stato interno $s = (\{AND,0\}, \{OR,1\})$; oppure $s = (0,1)$, per semplicità, il loro stato futuro $(0,1)$, sarà deterministicamente calcolato dallo stato corrente, ma ognuna di esse determina il proprio stato in modo indipendente dall’altra. Una volta che uno degli input di una porta AND è zero, l’altro input, proveniente dalla porta OR, non ha nessun effetto sullo stato futuro della porta AND. La stessa considerazione vale per la porta OR. Dunque il sistema “sceglie” con probabilità 1 il suo prossimo stato, ma esso è riducibile alle singole scelte delle porte AND e OR. In questo contesto si dice che il sistema non è integrato oppure è riducibile. Vedremo in seguito come formalizzare matematicamente il concetto di “taglio”, o “separazione”, e il concetto di “integrazione”.

Esclusione. Ci si chiede se sia S il sistema che esercita il maggior potere di causa-effetto sulle sue unità. Se un altro sottoinsieme S_1 di U esercita un maggiore potere di causa-effetto su parte delle unità di S, allora, basandosi sul principio di esistenza massimale, è S_1 ad esistere non S.

Composizione. Una volta che sono stati individuati tutti i complessi presenti nell’universo U mediante la ricerca esaustiva e la successiva comparazione in termini di informazione integrata, ci si chiede, per ogni complesso, come il potere di causa-effetto sia composto a

partire dai suoi sottoinsiemi. Questa operazione prende il nome di "dispiegazione" (unfolding in inglese).

Il postulato dell'intrinsecità viene implementato richiedendo che il potere di causa-effetto di un sistema, o un suo qualsiasi sottoinsieme sia esercitato su se stesso. Si indica con la dicitura T_U il TPM ottenuto per l'intero universo. Considerando un sottoinsieme S vengono calcolate altre due TPM: la matrice di transizione degli effetti T_e, e la matrice di transizione delle cause T_c.

L'idea è quella di catturare ogni possibile transizione da ogni stato attuale s di S a tempo t, verso un qualsiasi stato $\bar{s}$ futuro di S a tempo $t+1$ per quanto riguarda le transizioni di effetto, ed ogni possibile transizione da qualsiasi stato $\bar{s}$ di S a tempo $t-1$ verso un qualsiasi stato s corrente a tempo t. Si precisa che, in linea con la notazione usata dalla pubblicazione dove la dicitura può rappresentare delle ambiguità, s rappresenta sempre lo stato corrente a tempo t del sistema S, mentre $\bar{s}$ può rappresentare lo stato precedente, o il passato del sistema, a tempo $t-1$, oppure lo stato successivo, o futuro del sistema a tempo $t+1$ a seconda del contesto.

3.2.2 Transizioni di effetto

La matrice di transizione di effetto viene definita come

$$T_e = p_e(\bar{s} \mid s) = p(\bar{s} \mid u) = p(\bar{s} \mid s, w) \tag{3}$$

La scrittura delle tre eguaglianze serve per evidenziare e chiarire meglio, rispetto la pubblicazione, che la prima è semplicemente una forma abbreviata per enfatizzare l'effetto di s su $\bar{s}$, ma non corrisponde ad una probabilità nel senso tradizionale, da qua l'uso del simbolo p_e. Le altre due invece sono uguali e rappresentano la probabilità classica di $\bar{s}$ dato lo stato attuale dell'universo u, ossia specificando sia s che w. Queste probabilità possono essere calcolate grazie alla fattorizzazione che si ottiene assumendo la proprietà locale di Markov in accordo con la formula (2).

$$T_e = p_e(\bar{s} \mid s) = p(\bar{s} \mid u) = \prod_{i=1}^{|S|} p(\bar{s}_i \mid u)$$

Non risulta chiarissimo come questa formulazione rispecchi l'assioma dell'intrinsecità. La notazione $p_e(\bar{s}|s)$ dovrebbe in qualche modo "isolare" l'effetto che solo s ha sul suo stato futuro $\bar{s}$, ma come indicato in (3), l'effetto delle condizioni di fondo w è sempre presente. Di fatti dunque, quello che l'espressione (3) cattura, è l'effetto dello stato attuale dell'universo u sullo stato futuro di un suo sottoinsieme $\bar{s}$.

3.2.3 Transizioni di causa

La valutazione della matrice delle transizioni di causa risulta essere più complicata. Ora viene assunto che lo stato u dell'universo U viene osservato e non imposto. Ci si concentra su un sottoinsieme S in stato s e ci si chiede, per ogni valore degli stati precedenti $\bar{s}$, qual è la probabilità di s. La difficoltà in questo caso è che, rispetto al caso precedente nel quale u (e dunque anche w) erano imposti per intervento, in questo caso il sistema a tempo t "non sa" lo stato delle condizioni di fondo $\bar{w}$ a tempo $t-1$. La TPM di causa è definita come :

$$T_c = p_c(s \mid \bar{s}) = \prod_{i=1}^{|S|} \sum_{\bar{w}} p\left(s_i \mid \bar{s}, \bar{w}\right) \left(\frac{\sum_{\hat{s}} p(u \mid \hat{s}, \bar{w})}{\sum_{\hat{u}} p(u \mid \hat{u})} \right), \quad s, \bar{s} \in \Omega_s \qquad (4)$$

L'espressione viene calcolata a partire dalla transizione di una singola unità s_i

$$p(s_i|\bar{s}) = \sum_{\bar{w}} p(s_i|\bar{s},\bar{w}) q(\bar{w})$$

Non è chiaro l'uso di $q(\bar{w})$ rispetto a $p(\bar{w}|\bar{s})$ come dalla legge di totale probabilità. Inoltre si parla dell'uso dello stato attuale del sistema per calcolare la distribuzione di probabilità delle condizioni di fondo a tempo $t-1$, che non è necessariamente uniforme o deterministica. Questo motiva il seguente calcolo usando la regola di Bayes :

$$q(\bar{w}) = p(\bar{w}|u) = \frac{p(u|\bar{w})p(\bar{w})}{p(u)} = \frac{\sum_{\hat{s}} p(u \mid \hat{s}, \bar{w}) p(\hat{s}|\bar{w}) p(\bar{w})}{\sum_{\hat{u}} p(u \mid \hat{u}) p(\hat{u})}$$

I termini $p(\hat{s}|\bar{w})p(\bar{w}) = p(\hat{s},\bar{w})$ e $p(\hat{u})$ si semplificano in virtù del fatto che si assume implicitamente che essi siano entrambi stati precedenti del sistema, ed abbiano una probabilità uniforme. Combinando (moltiplicando) questi termini ottenuti per una singola unità s_i, per ogni i da 1 a $|S|$ si ottiene la formula (4).

Tuttavia, questa moltiplicazione dovrebbe poter essere giustificata, in linea con la proprietà locale di Markov, quando per ogni unità s_i vengono dati i suoi genitori, che sono $\bar{u}$, ossia $(\bar{s},\bar{w})$ e non solo $\bar{s}$ come viene indicato nella derivazione della formula.

Questa tesi propone un calcolo alternativo per la TPM di causa T_c. Innanzitutto si argomenta che una scelta ragionevole sulla distribuzione di probabilità dello stato precedente $\bar{u}$ dell'universo, è quella che rappresenta il valore più alto dell'incertezza di $\bar{u}$, la distribuzione con il più alto valore di entropia, la distribuzione uniforme. Questo viene usato dalla pubblicazione per il calcolo della formula (4), ma non viene esplicitato. Si definisce dunque :

$$T_c = p\left(s|\bar{s}\right) = \sum_{\bar{w}} p\left(s|\bar{s},\bar{w}\right) p\left(\bar{w}|\bar{s}\right)$$

mentre la seguente probabilità può essere fattorizzata come :

$$p\left(s|\bar{s},\bar{w}\right) = \prod_{i=1}^{|S|} p\left(s_i|\bar{s},\bar{w}\right)$$

mentre

$$p\left(\bar{w}|\bar{s}\right) = p\left(\bar{w},\bar{s}\right)/p\left(\bar{s}\right) = \left|\Omega_U\right|^{-1}\left|\Omega_S\right|^{-1} = \left|\Omega_W\right|^{-1}$$

Dunque la formula risulta essere

$$T_c = p\left(s|\bar{s}\right) = \left|\Omega_W\right|^{-1} \sum_{\bar{w}} \prod_{i=1}^{|S|} p\left(s_i|\bar{s},\bar{w}\right) \qquad (4')$$

Il calcolo di queste matrici di probabilità nel caso peggiore è esponenziale nel numero delle unità di U. Come nel caso del calcolo di T_U, il calcolo di T_c impone un ciclo attraverso i $|\Omega_W| =$

$2^{|W|}$ stati possibili delle condizioni di fondo, che sono nel caso peggiore, $2^{|U|-2}$, visto che il più piccolo sottoinsieme S può avere 2 soli elementi e dunque $|W| = |U| - 2$

3.3 Informazione

L'IIT definisce l'informazione come la misura di potere intrinseco di causa-effetto esercitato dal sistema nel suo stato attuale, selezionando uno stato specifico di causa ed uno specifico di effetto. Cercando di dare un'introduzione intuitiva a questa misura, consideriamo per esempio solo le transizioni di effetto. La $TPM\ T_e$ calcola la probabilità che ha il sistema di transitare in ognuno dei sui stati futuri $\bar{s}$ a partire dallo stato corrente s. Tuttavia, la probabilità $p(\bar{s})$ che $\bar{s}$ accada, potrebbe non dipendere da s, (s e $\bar{s}$ sono indipendenti), allora vuol dire che il potere esercitato dal sistema nello stato corrente verso quel determinato stato futuro è nullo. L'informazione in questo caso è zero.

L'IIT propone l'informazione intrinseca (ii) [2] come una misura di "differenza", o "distanza", tra due distribuzioni di probabilità tra la probabilità condizionata $p(\bar{s} \mid s)$ e la probabilità marginale (incondizionata) di $p(\bar{s})$ di $\bar{s}$. Nel caso delle transizioni di effetto essa viene definita come :

$$ii_e(s,\bar{s}) = D(p_e(\bar{s} \mid s) \mid\mid p_e(\bar{s})) \ = \ p_e(\bar{s} \mid s) \log \left(\frac{p_e(\bar{s} \mid s)}{p_e(\bar{s})} \right) \tag{5}$$

Il termine $p_e(\bar{s})$ viene calcolato come segue:

$$p_e(\bar{s}) = |\Omega_S|^{-1} \sum_{s \in \Omega_S} p_e(\bar{s} \mid s) \tag{6}$$

Come già anticipato durante il calcolo della matrice di transizione T_c, la notazione di probabilità p_e racchiude al suo interno la dipendenza dalle condizioni di fondo w, quindi non risulta chiaro l'attribuzione del potere di causa-effetto allo stato s del sottosistema verso $\bar{s}$, rispetto quello che tutto il sistema nello stato u verso $\bar{s}$.

L'informazione intrinseca assomiglia molto alla definizione della divergenza di Kullback–Leibler o misura di entropia relativa della teoria dell'informazione standard (per differenziare con l'IIT), ma in questo caso è riferita unicamente allo stato specifico in cui si trova il sistema.

Un modo alternativo dunque per riscrivere l'equazione 5 è il seguente:

$$ii_e(u,\bar{s}) = D(p(\bar{s} \mid u) \mid\mid p(\bar{s})) \;=\; p(\bar{s} \mid u) \log \left(\frac{p(\bar{s} \mid u)}{p(\bar{s})} \right) \qquad (5')$$

e il termine $p(\bar{s})$:

$$p(\bar{s}) = |\Omega_U|^{-1} \sum_{u \in \Omega_U} p(\bar{s} \mid u) \qquad (6')$$

Si noti come lo stato attuale del sottosistema è stato sostituito con quello del sistema universo, e sono state inoltre trasformate le probabilità p_e in probabilità tradizionali. Le due scritture sono equivalenti. Una nota importante che gioca un ruolo fondamentale, è la considerazione dell'uniformità dello stato corrente u nella formula (6'), e analogamente s nella formula (4). Nonostante questa sia una transizione da t a $t+1$, quando si tratta di valutare la probabilità marginale di uno stato futuro, incondizionato al presente, ci si ritrova nella medesima situazione descritta nella formula (4') sulle transizioni da $t-1$ a t. Valgono dunque le stesse considerazioni che portano ad una scelta di probabilità uniforme.

Per quanto riguarda le transizioni di causa, l'informazione intrinseca viene definita come :

$$ii_c(s,\bar{s}) = D(p_c(s \mid \bar{s}) \mid\mid p_c(s)) \;=\; p_c^{\leftarrow}(\bar{s} \mid s) \log \left(\frac{p_c(s \mid \bar{s})}{p_c(s)} \right) \qquad (7)$$

Non risulta chiaro in questo caso l'uso del termine $p_c^{\leftarrow}(\bar{s} \mid s)$ al posto di $p_c(s \mid \bar{s})$, ottenuto da quest'ultimo tramite l'utilizzo della regola di Bayes.

Si parla di informazione intrinseca quando si considera l'informazione rispetto a due stati fissati s e $\bar{s}$, mentre il postulato richiede di "scegliere" uno stato specifico s', quello verso il quale si ha il massimo valore dell'informazione. Nel caso delle transizioni di effetto (e similarmente per quelle di causa) abbiamo :

$$s_e{}' = \operatorname*{argmax}_{\bar{s} \in \Omega_S} ii_e(s, \bar{s}) \qquad (12)$$

Si parla in questo caso di distanza intrinseca (ID intrinsic distance [2]) e viene definita come il valore massimo dell'informazione intrinseca :

$$ID(s) = ii_e(s, s'_e) = \max_{\bar{s} \in \Omega_S} p_e(\bar{s} \mid s) \log \left(\frac{p_e(\bar{s} \mid s)}{p_e(\bar{s})} \right) \tag{13}$$

Importante notare come diverse nozioni di "distanza", o "differenza", siano state usate in diverse iterazioni dell'IIT; inizialmente usando la KLD (Kullback–Leibler divergence) in IIT 2.0, per poi optare per EMD (earth mover's distance) in IIT 3.0, ed infine all'intrinsic difference (ID)

3.4 Integrazione

Una volta determinati gli stati massimi di causa-effetto del sottosistema S, si procede all'analisi dell'integrazione del sistema. Intuitivamente ci si chiede se il sistema determina gli stati massimi di causa ed effetto come un tutt'uno, oppure è riducibile a suoi sottoinsiemi, o parti, che "esistono" separatamente l'uno dall'altro. Se il sistema ridotto alle sue parti determina almeno uno degli stati massimi tra causa ed effetto con la stessa probabilità dell'intero sistema, allora si conclude che il sistema non è integrato. Viene dunque definito il concetto di "partizioni direzionali del sistema". Esso considera ogni possibile partizione del sistema in $k >= 2$ parti disgiunte che "esistono" separatamente. Il termine "esistono" viene inteso come "esistenza" nell'ambito dell'IIT, ossia capacità di esercitare un potere di causa-effetto. Vengono dunque considerati tutti i possibili modi per ottenere questo scenario esercitando dei "tagli" che "separano causalmente" le parti tra di loro.

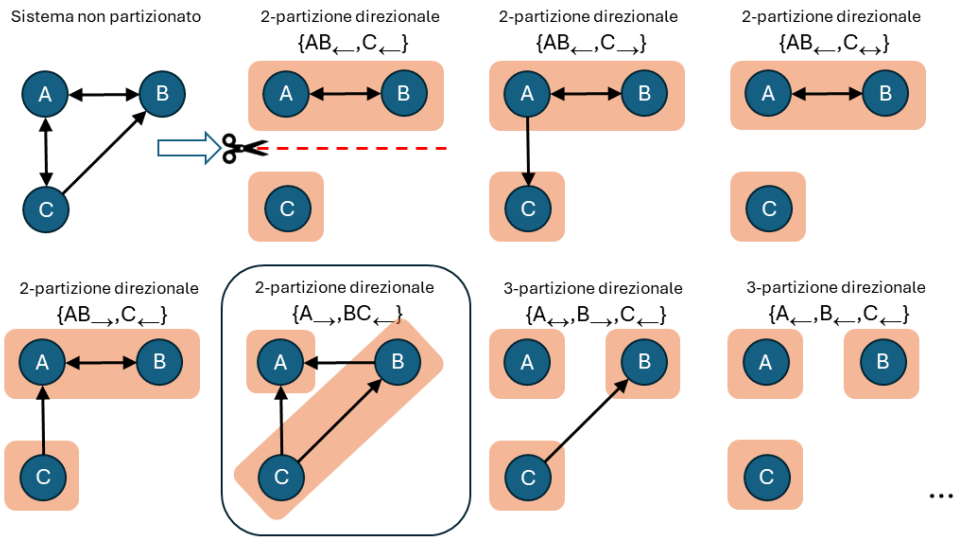

Figura 3.5: Esempio di partizioni direzionali

Nell'immagine vengono mostrate solo alcune delle partizioni possibili e alcuni dei tagli possibili del sistema $U = (A, B, C)$. In particolare per la partizione AB, C vengono indicati sette possibili tagli, per la partizione A, BC un solo taglio, e per la partizione A, B, C due tagli. Come si vede nell'esempio non tutti i tagli portano a configurazioni uniche.
Formalmente, l'insieme delle partizioni direzionali del sistema $\Theta(S)$, è l'insieme che si ottiene applicando "tagli direzionali" ad ogni elemento dell'insieme delle partizioni del sistema $P(S)$.

$$P(S) = \left\{ \left\{ S^{(i)} \right\}_{i=1}^{k} : S^{(i)} \neq \emptyset, \quad S^{(i)} \cap S^{(j)} = \emptyset, \bigcup_{i=1}^{k} S^{(i)} = S, \quad k \geq 2 \right\}$$

Chiameremo semplicemente "parte" un elemento della partizione. Per esempio, in figura 3.5, AB e C sono le due parti della partizione AB, C. Per "taglio direzionale" si intende che, per ogni partizione di cui sopra, ad ogni sua parte vengono tagliate tutte le "connessioni causali"[2]. Ulteriormente, tagliare le connessioni causali, implica che la parte non influenza il resto del sistema, il resto del sistema non influenza la parte, oppure entrambe le cose. Operativamente

[2]Le connessioni causali di una parte sono tutti gli archi uscenti, oppure tutti gli archi entranti

questo significa che ad ogni parte vengono "tagliati" rispettivamente tutti gli archi in uscita, tutti quelli in entrata, oppure entrambi. Questa scelta va fatta in modo indipendente per ogni parte della partizione. Un elemento $\theta \in \Theta(S)$ ha dunque la forma :

$$\theta = \{S^{(1)}_{\delta_1}, S^{(2)}_{\delta_2}, \ldots, S^{(k)}_{\delta_k}\} \tag{15}$$

dove $\delta_i \in \{\rightarrow, \leftarrow, \leftrightarrow\}$ e le frecce indicano la direzione del taglio. L'operazione di "tagliare" una relazione causale, corrisponde a sostituire l'input tagliato con un input determinato uniformemente a caso tra lo spazio degli stati dell'unità da dove l'input proviene. In riferimento all'immagine sopra, considerando solo la partizione direzionale $\{A_{\rightarrow}, BC_{\leftarrow}\}$, sono stati tagliati tutti gli archi in uscita da A e quelli in entrata in BC. L'insieme U può sembrare "connesso", ma in realtà A non può influenzare BC e BC non può essere influenzato da A. In questo modo, per ogni parte $S^{(i)}$ della partizione direzionale, possiamo identificare un insieme di unità $X^{(i)} \subseteq S$ l'input dei quali verso la parte è stato tagliato, e l'insieme complementare $Y^{(i)} = S \backslash X^{(i)}$ l'input dei quali verso la partizione è rimasto intatto. Nell'esempio sopra con riferimento alla parte BC, $X = \{A\}$ e $Y = \{B, C\}$. Vengono dunque costruite due matrici di transizione (causa ed effetto) per il sistema tagliato, in linea con quelli di cui formula (3) e (4). Nel caso delle transizioni di effetto abbiamo :

$$T^{\theta}_{e} = p^{\theta}_{e}(\bar{s} \mid s) = \prod_{j=1}^{|S|} p^{\theta}_{e}(\bar{s_j} \mid s) \qquad \bar{s}, s \in \Omega_S \tag{17}$$

mentre le singole probabilità delle unità vengono calcolate considerando la media su tutti i possibili stati delle unità tagliate, di fatti, rimpiazzando l'influenza causale con rumore bianco.

$$p^{\theta}_{e}(\bar{s_j} \mid s) = |\Omega_{X^i}|^{-1} \sum_{x^i \in \Omega_{X^i}} p^{\theta}_{e}(\bar{s_j} \mid x^i, y) \tag{18}$$

Viene dunque introdotto il concetto di informazione integrata del sistema (φ) come misura di quanto il sistema partizionato riduce il potere di causa-effetto verso lo stato di massimo valore dell'informazione intrinseca s', come da (12). Per le transizioni di effetto.

$$\varphi_e(T_e, s, \theta) = p_e(s'_e \mid s) \left| \log \left(\frac{p_e(s'_e \mid s)}{p^{\theta}_e(s'_e \mid s)} \right) \right|_+ \tag{19}$$

Si noti che la misura di "differenza" tra le due distribuzioni $p_e(s'_e \mid s)$ e $p^{\theta}_e(s'_e \mid s)$ è la stessa dell'informazione intrinseca. L'operatore $|.|_+$ rappresenta la parte positiva e impone a zero i valori negativi. Viene usato per assicurarsi che il sistema, come un tutt'uno, aumenti la probabilità dello stato s'_e rispetto alla probabilità nel sistema partizionato.
Intuitivamente se $p_e(s'_e \mid s) = p^{\theta}_e(s'_e \mid s)$ allora il sistema partizionato sceglie s'_e con la stessa probabilità del sistema non partizionato, dunque quest'ultimo non è integrato, quindi $\varphi_e = 0$. Nel caso $p_e(s'_e \mid s) < p^{\theta}_e(s'_e \mid s)$ allora il sistema partizionato aumenta probabilità di s'_e risultando in un valore negativo dell'informazione integrata, che in assenza di $|.|_+$, sarebbe difficile da interpretare.
Per le transizioni di causa l'informazione integrata viene definita come :

$$\varphi_c(T_c, s, \theta) = p^{\leftarrow}_c(s'_c \mid s) \left| \log \left(\frac{p_c(s \mid s'_c)}{p^{\theta}_c(s \mid s'_c)} \right) \right|_+ \tag{20}$$

Dove, come nel caso (7) l'uso di $p^{\leftarrow}_c(s'_c \mid s)$, non risulta chiaro. Si definisce informazione integrata del sistema rispetto alla partizione, il minimo tra l'informazione integrata per le transizioni di causa ed effetto sopra illustrate.

$$\varphi_s(T_e, T_c, s, \theta) = \min\{\varphi_c(T_c, s, \theta),\ \varphi_e(T_c, s, \theta)\} \tag{21}$$

La scelta del minimo tra le due quantità viene fatto in linea con il principio di esistenza minima. Nel caso uno dei due fosse zero allora il sistema non può essere considerato integrato. Nuovamente, invocando lo stesso principio, calcolando il valore di informazione integrata per ogni partizione direzionale del sistema è lecito pensare che, se esiste una per cui l'informazione integrata è zero, allora il sistema non è integrato. Indicheremo con $\theta' \in \Omega(S)$ la partizione che minimizza il valore dell'informazione integrata. Tale partizione si chiama MIP (minimum information partition). Si definisce dunque, semplicemente ommettendo la partizione tra i parametri, l'informazione integrata del sistema che equivale all'informazione integrata del sistema valutata sul MIP.

$$\varphi_s(T_e, T_c, s) = \varphi_s(T_e, T_c, s, \theta^{'}) \tag{22}$$

Tuttavia confrontare direttamente le partizioni, considerando il loro valore d'integrazione, può non essere la scelta migliore, in quanto esso dipende anche dal numero delle parti e dalla loro dimensione. Alcune partizioni possono essere peggiori di altre per integrare informazione in generale. Il confronto viene fatto dunque sulla misura relativa di informazione integrata, ossia il valore di integrazione calcolato rispetto al massimo valore di integrazione possibile per ogni $TPM\ T'_e$ e T'_c. Questo è il motivo per cui la partizione MIP viene definita come :

$$\theta^{'} = \underset{\theta \in \Theta(S)}{\text{argmin}} \frac{\varphi_s(T_e, T_c, s, \theta)}{\max_{T'_e, T'_c} \varphi_s(T'_e, T'_c, s, \theta)} \tag{23}$$

In [11] viene dimostrato che :

$$\max_{T'_e, T'_c} \varphi_s(T'_e, T'_c, s, \theta) = \sum_{i=1}^{k} |S^i|\ |X^i|$$

Generalmente aver trovato un sottosistema $S \in U$ con $\varphi > 0$, non garantisce che esso sia un substrato di coscienza. Molteplici sottosistemi possono concorrere su parte delle unità di S. Invocando il principio di esistenza massimale, il sottosistema S^*, con il valore più alto di integrazione, è quello che esiste, mentre gli altri no. In questo caso S^* è un complesso ed è definito :

$$S^* = \underset{S \in U}{\text{argmax}}\, \varphi_s(T_e, T_c, s) \tag{25}$$

Ulteriormente, invocando il postulato dell'esclusione, il complesso è composto esattamente dalle sole unità specificate al suo interno. Nell'immagine seguente viene illustrato un esempio dove vengono confrontati per semplicità solo tre sottosistemi : $\{(2,3), (1,2,3,4,6), (4,5,6)\}$. Il valore più alto dell'informazione integrata si ottiene per il sottosistema $S^* = (4,5,6)$, che è dunque un

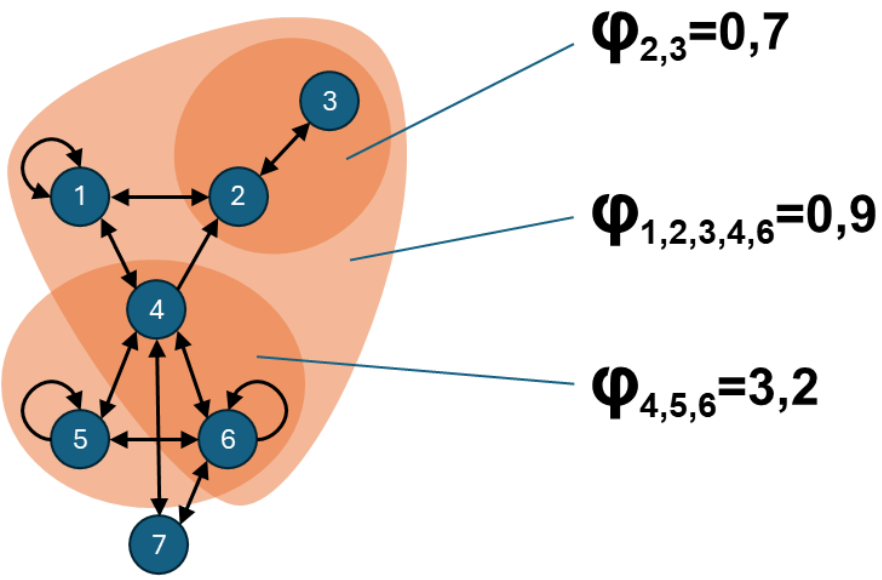

Figura 3.6
Un sistema dinamico di 7 nodi (in blu) e 3 candidati complessi (in arancione).

complesso. Le unità restanti $U \setminus S^*$ devono nuovamente essere considerate per la ricerca di altri complessi minori.

3.4.1 Complessità computazionale

Prima di proseguire ad analizzare la complessità computazionale della procedura proposta, si apre una parentesi sulla non unicità delle configurazioni risultanti apportando i tagli direzionali. Ricollegandoci alla figura 3.5 si nota che, pur operando tagli diversi, le configurazioni relative a $\{AB_{\leftarrow}, C_{\leftarrow}\}$, $\{AB_{\leftarrow}, C_{\leftrightarrow}\}$ e $\{AB_{\rightarrow}, C_{\leftarrow}\}$ risultano essere la stessa. Mediante una simulazione numerica, si può far vedere che il rapporto tra partizioni uniche e partizioni totali tende a uno al crescere del numero di unità.
Di seguito viene fornita una dimostrazione di questo fatto nel caso in cui il sistema corrisponda ad un grafo completo, ossia per ogni coppia di unità esiste l'arco diretto in entrambe le direzioni. Nella parte 1) della figura 3.7 vengono mostrati tutti i possibili tagli dell'unica partizione $\{A, B\}$. Si noti che esistono 9 possibili tagli dei quali le configurazioni 1b) e 1c) sono uniche, mentre la configurazione 1d) risulta da 7 diversi tagli.
Si può argomentare, in generale, che se due tagli producono la stessa configurazione, essa è necessariamente la configurazione che elimina tutti gli archi. Basti pensare ad una configurazione dove esiste almeno un arco non tagliato. Questo significa che per la parte da dove l'arco

esce è stata necessariamente fatta la scelta di tagliare tutti quelli in entrata. Nel grafo completo questa partizione è collegata a tutte le altre, quindi vincola la scelta su di esse ad aver tagliato gli archi in uscita. Per ogni configurazione dove almeno un arco non è stato tagliato, viene dunque identificata una scelta unica dei tagli direzionali.
Per quanto riguarda invece il numero di tagli che possono risultare in una combinazione senza archi, si usa il seguente argomento. Consideriamo il grafo in forma matriciale, come indicato nella sezione 2) della figura 3.7 per sistema di 4 unità (A, B, C, D), ma l'argomento rimane valido per un numero arbitrario di nodi. Gli uni nella matrice indicano un arco diretto tra l'indice di riga a quello di colonna. Viene considerata solo la partizione dove ogni nodo è anche una parte, in quanto ogni altra partizione è analoga a questa, considerando tutti i nodi all'interno di una parte come se fosse un unico nodo. Sempre con riferimento a 2) in figura 3.7 è ovvio che per avere una situazione di non unicità, come dimostrato in precedenza, la matrice deve contenere solo zeri una volta apportati tutti i tagli. Procedendo dal nodo A e facendo riferimento anche all'immagine 3), si hanno 3 scelte →,← oppure ↔, che nella matrice corrispondono ad azzerare rispettivamente la riga, la colonna, oppure entrambe. Consideriamo per A la scelta →, di azzerare le righe (la scelta ← è analoga). In questo caso, per tagliare l'arco diretto $(B \to A)$, il numero delle scelte di B è ridotto a due: tagliare gli archi in uscita (→), oppure quelli in entrata e uscita (↔). Una volta fatta la scelta iniziale su A, questa condiziona le scelte delle altre parti, fissandole a due. Per ogni parte (come nell'esempio di B) si deve scegliere tra tagliare tutti gli archi in uscita, oppure in entrata e uscita perché l'arco diretto $(P \to A)$ è presente per ogni parte P . Se invece si sceglie ↔ per A, ora B ha tre scelte. Nel caso di scelta su B tra →, o ← la situazione è analoga a quella quando si era fatta la stessa scelta su A. Analizzando l'albero in 3) possiamo dunque scrivere la seguente equazione ricorsiva per contare il numero di nodi, che corrisponde al numero di tagli che producono una configurazione senza archi.
Indicando $S(k)$ il numero totale di nodi e notando che l'albero, senza considerare la parte centrale, quella generata dalla scelta ↔ su A, è un albero binario di 2^k nodi. L'albero generato dalla scelta ↔ su A rispecchia quello più grande, ma con una profondità in meno. Possiamo dunque scrivere la seguente equazione ricorsiva :

$$\begin{cases} S(k) = 2^k + S(k-1) \\ S(0) = 1 \end{cases}$$

che da come risultato $S(k) = 2^{k+1} - 1$. Il numero totale di possibili tagli è semplicemente 3^k visto che per ogni parte abbiamo 3 scelte. Dunque il numero totale di partizioni direzionali

uniche è $3^k - 2^{k+1} + 2$, aggiungendo 1 per includere la partizione senza archi nel conteggio finale.

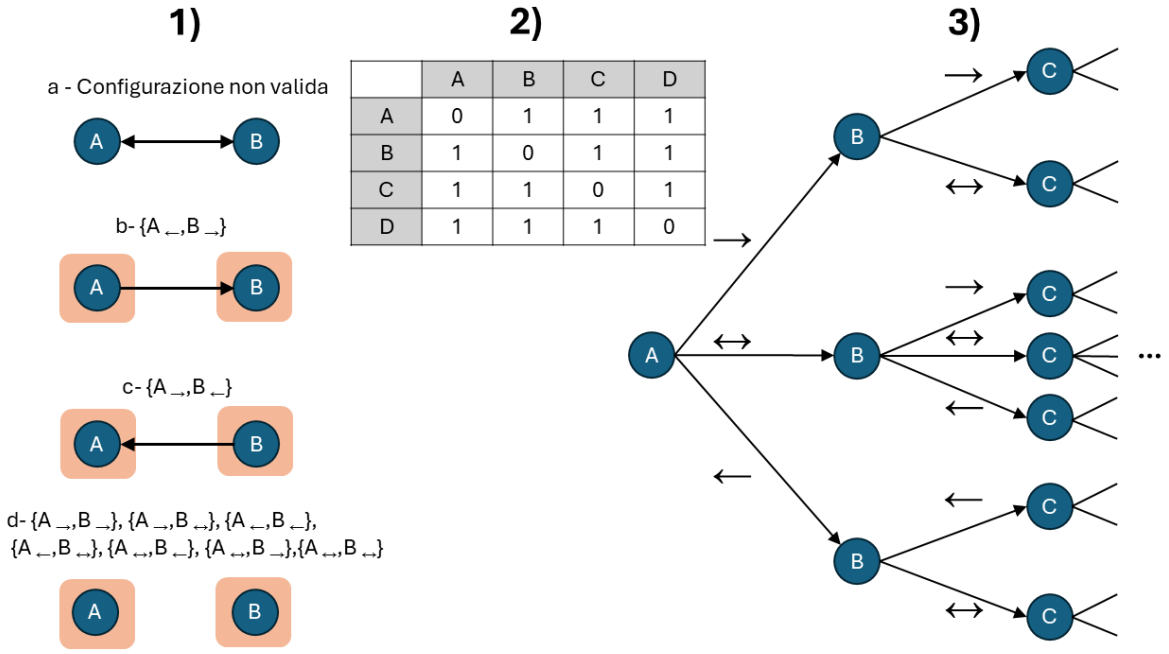

Figura 3.7
1) Partizioni direzionali di un sistema di 2 unità. 2) Matrice di connettività per un sistema di 4 unità. 3) Rappresentazione ad albero delle scelte fatte per ogni parte/unità.

Dal punto di vista della complessità computazionale questa operazione risulta molto onerosa. Considerando nel caso peggiore un sistema dinamico rappresentato da un grafo completo, il calcolo esatto delle operazioni procede come segue. Ponendo $|U| = n$, dobbiamo considerare ogni suo sottoinsieme di m elementi. Questo numero è dato da $\sum_{m=1}^{n} \binom{n}{m}$ dove $\binom{n}{m}$ è il coefficiente binomiale. Esso conta tutti i modi in cui possiamo scegliere m elementi su n, di fatti contando tutti i sottoinsiemi di m elementi su n. Per ognuno di essi, per valutare l'integrazione, dobbiamo considerare tutte le possibili k-partizioni. Questo numero è dato da $\sum_{k=1}^{n} \left\{ {n \atop k} \right\}$, dove $\left\{ {m \atop k} \right\}$ o $S(m,k)$ sono i numeri di Stirling del secondo ordine e contano il numero di partizioni di un insieme di m elementi in k sottoinsiemi disgiunti. Per ogni partizione di k elementi dobbiamo considerare $3^k - 2^{k+1} + 2$ possibili tagli unici, per un totale di :

$$\sum_{m=1}^{n} \binom{n}{m} \sum_{k=1}^{m} \left\{ {m \atop k} \right\} (3^k - 2^{k+1} + 2)$$

Per avere un' idea approssimativa dell'andamento asintotico, e ricavare un ragionevole limite inferiore possiamo considerare le seguenti approssimazioni.

Consideriamo inizialmente solo le bipartizioni del sistema e non tutte le possibili k-partizioni. Per $k = 2$ dunque, abbiamo $S(m, k) = S(m,2) = 2^{m-1}-1$ che unito al termine $3^k-2^{k+1}+2 = 3$ e ignorando le costanti moltiplicative possiamo approssimare a 2^m. Usando questo valore si ottiene l'espressione

$$\sum_{m=1}^{n} \left(\begin{array}{c} n \\ m \end{array} \right) 2^m$$

la quale, ricordando la seguente espansione binomiale di

$$(1+2)^n = \sum_{m=0}^{n} \left(\begin{array}{c} n \\ m \end{array} \right) 1^{n-m}\, 2^m$$

si semplifica in $3^n - 1$ [3] e rivela la natura esponenziale del procedimento.

D'altro canto, considerando tutte le k-partizioni, ma approssimando $3^k - 2^{k+1} + 2$ con 3^k, l'espressione risulta essere :

$$\sum_{m=1}^{n} \left(\begin{array}{c} n \\ m \end{array} \right) \sum_{k=1}^{m} \left\{ \begin{array}{c} m \\ k \end{array} \right\} 3^k = \sum_{m=1}^{n} \left(\begin{array}{c} n \\ m \end{array} \right) T_m(3)$$

dove il termine $T_m(3)$ è conosciuto come polinomio di Touchard. L'approssimazione di tale espressione, pur rivelando un interessante natura combinatoria del problema, va oltre lo scopo di questa tesi.

Tuttavia un limite inferiore molto più rappresentativo si può ricavare non considerando tutti i possibili tagli direzionali del sistema, ma solo un numero costante di essi; costante che per i fini di un' approssimazione asintotica può essere ignorata.

Otteniamo dunque la seguente espressione :

[3] Il fattore "-1" è dato dal fatto che la sommatoria comincia da $m = 1$.

$$\sum_{m=1}^{n} \left(\begin{array}{c} n \\ m \end{array} \right) \sum_{k=1}^{m} \left\{ \begin{array}{c} m \\ k \end{array} \right\} = \sum_{m=1}^{n} \left(\begin{array}{c} n \\ m \end{array} \right) B_m = B_{n+1}$$

dove $B_m = \sum_{k=1}^{m} \left\{ \begin{array}{c} m \\ k \end{array} \right\}$ è il m-esimo numero di Bell e conta il numero totale delle k-partizioni di un insieme di m elementi. L'ultima eguaglianza è ricavata dalle proprietà di questi numeri.

Un interessante limite inferiore, e superiore alla loro crescita, è dato da Berend & Tassa (2010) [12, p. 4]

$$\left(\frac{n+1}{e \ln(n+1)} \right)^{n+1} \leq B_{n+1} \leq \left(\frac{0{,}792(n+1)}{\ln(n+2)} \right)^{n+1}$$

il quale, approssimando per difetto $\frac{1}{e}$ con 0.367 risulta :

$$\left(\frac{0.367(n+1)}{\ln(n+1)} \right)^{n+1} \leq B_{n+1} \leq \left(\frac{0{,}792(n+1)}{\ln(n+2)} \right)^{n+1}$$

Il costo computazione cresce dunque almeno in maniera esponenziale rispetto al numero di unità.

3.5 Composizione

Consideriamo adesso l'ultimo passo verso l'analisi completa del sistema. L'ultimo assioma della composizione ci dice che, la coscienza è strutturata, è composta da distinti elementi e relazioni tra di essi. Dunque il potere di causa-effetto esercitato dal complesso su se stesso deve essere strutturato, ossia sottoinsiemi delle sue unità esercitano potere di causa ed effetto su altri suoi sottoinsiemi che sovrapponendosi danno luogo alla struttura di causa-effetto, oppure Φ-struttura. La Φ-struttura descrive qualitativamente l'esperienza, mentre la metrica Φ, che introdurremo, la descrive quantitativamente.

3.5.1 Meccanismi

Indicheremo sempre con la lettera S il sottoinsieme complesso di U che stiamo analizzando. Ora esso non è più un candidato, ma un substrato massimo di coscienza, un complesso. Considereremo ogni suo sottoinsieme $M \subseteq S$, detto "meccanismo" (mechanism in inglese), e ogni

sottoinsieme $Z \subseteq S$ detto "ambito", o "competenza" (purview in inglese, che manterremo in uso come termite tecnico). Ci chiediamo dunque, per ogni M nello stato m, derivato dallo stato s di S, verso quale sottoinsieme Z ha il maggior potere di causa-effetto. Ci concentreremo solo sul lato effetto delle transizioni, da t a $t+1$ e non su quelle di "causa", da $t-1$ a t visto che la trattazione è simile a quella fatto in precedenza. Saranno dunque oggetto di analisi coppie (M, Z), elementi del prodotto cartesiano $S \times S$. Per z in questa analisi è sottinteso z_e, ossia lo stato futuro di Z verso il quale M esercita un potere di effetto. Per concentrare l'analisi solo sull'effetto di m su z, e non quella delle unità $X = S \backslash M$, esse vengono "causalmente marginalizzate". Operazione che è formalmente rappresentata dalla sostituzione di tutti gli input da X verso Z con rumore bianco, lasciando intatte quelle da M. Questo implica impostare una distribuzione uniforme $p(X)$ e fare la media su tutti i possibili stati Ω_X.

$$p_e(z_i \mid m) = |\Omega_X|^{-1} \sum_{x \in \Omega_X} p(z_i \mid m, x) \quad z_i \in \Omega_{Z_i} \tag{28}$$

Le probabilità cosi ottenute per ogni unità in Z, in linea con l'assunzione di indipendenza di Markov, possono essere moltiplicate per avere la probabilità condizionata di tutte le unità in Z. Il simbolo π viene usato per enfatizzare questo fatto.

$$\pi_e(z \mid m) = \prod_{i=1}^{|Z|} p_e(z_i \mid m) \qquad z \in \Omega_Z \tag{29}$$

Allo stesso modo in in cui abbiamo ricavato le probabilità incondizionate in (6), ricaviamo :

$$\pi_e(z \; ; \; M) = |\Omega_M|^{-1} \sum_{m \in \Omega_M} \pi_e(z \mid m) \quad z \in \Omega_Z \tag{31}$$

con la notazione $\pi_e(z \; ; \; M)$ vuole ricordare che le probabilità marginali di z in questo caso sono state calcolate fissando il meccanismo M.
Come fatto in precedenza (formula 5), per calcolare la differenza tra le due distribuzioni di probabilità, allo stesso modo viene calcolata l'informazione intrinseca che il meccanismo s specifica sul purview z :

$$ii_e(m, z) = \pi_e(z \mid m) \log \left(\frac{\pi_e(z|\ m)}{\pi_e(z\ ;\ M)} \right) \tag{34}$$

Basandoci nuovamente sul principio di esistenza massima, m deve scegliere lo stato futuro z'_e verso il quale esercita maggior effetto. Dunque :

$$z'_e = \underset{z \in \Omega_Z}{\text{argmax}}\, ii_e(m, z) \tag{36}$$

e

$$ii_e(m, z'_e) = \max_{z \in \Omega_Z} ii_e(m, z) \tag{37}$$

Inoltre, in linea con i postulati dell'IIT si richiede che anche i meccanismi soddisfino esistenza, intrinsecità, informazione, integrazione; mentre si fa eccezione per la composizione in quanto non sono essi stessi substrati di coscienza.
L'esistenza nel senso dell'IIT è soddisfatta in quanto deve essere possibile per qualcosa cambiare lo stato di un meccanismo, e deve essere possibile per il meccanismo cambiare lo stato di qualcosa. I sottoinsiemi $Z \subseteq S$ costituiscono il repertorio di causa-effetto cercato. Inoltre intrinsecità e informazione sono soddisfatti in quanto il meccanismo influenza altri elementi all'interno di S stesso, e mediante l'informazione intrinseca, sceglie uno stato specifico su unità di S, z_e e z_c in questo caso. Per soddisfare il postulato dell'integrazione, dobbiamo calcolare quanto il potere di effetto esercitato dal meccanismo verso il purview sia irriducibile.
Viene dunque introdotto, diversamente da quanto fatto in formula (15), il concetto di "taglio" tramite il concetto di partizione "disintegrante" di (M, Z) come segue:

$$\Theta(M, Z) = \left\{ \{(M^{(i)}, Z^{(i)})\}_{i=1}^{k} : k \in \{2,3,4 \dots\} \right\} \tag{38}$$

dove $\{M^{(i)}\}$ e $\{Z^{(i)}\}$ sono partizioni di M e Z. Una nota aggiuntiva rispetto la pubblicazione è la seguente precisazione. In figura figura 3.8 d) è permessa la partizione $\{(ADB, \emptyset), (\emptyset, FE)\}$,

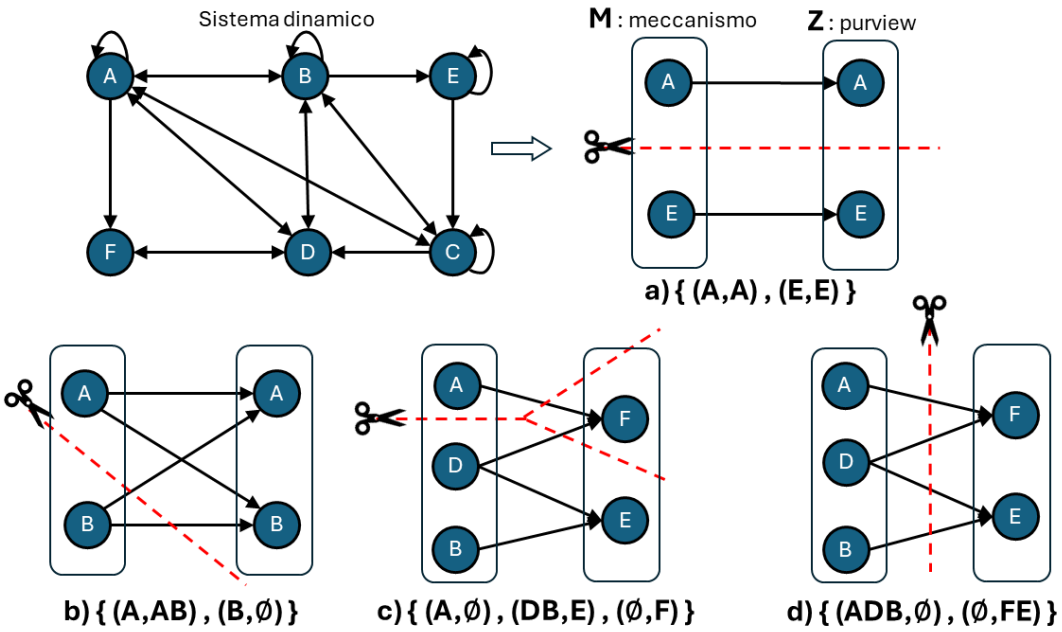

Figura 3.8
Illustrazione di possibili partizioni disintegranti delle coppie meccanismo-purview.

che rappresenta un'eccezione rispetto la definizione (38). Nell'esempio b) in figura 3.8 c) il meccanismo (AB) viene partizionato in $\{(A), (B)\}$ mentre il purview non viene partizionato, o meglio partizionato in una parte $\{(A, B)\}$, che risulta in contrasto con la definizione. Si deduce dunque che M deve essere sempre partizionato in $k \geq 2$ parti, mentre Z in $k \geq 1$ parti.
Come fatto in precedenza possiamo ricostruire le probabilità di z condizionate ad m nel sistema partizionato (coppia meccanismo-purview). Dato una partizione $\theta \in \Theta(M, Z)$ e lo stato di massimo effetto, ossia lo stato z'_e che massimizza l'informazione intrinseca, le probabilità condizionate nel sistema partizionato vengono date da:

$$\pi_e^{\theta}(z'_e \mid m) = \prod_{i=1}^{k} \pi_e(z_e'^{(i)} \mid m^{(i)}) \tag{39}$$

Questa fattorizzazione è possibile perché le componenti di z'_e presenti nelle partizioni i-esime sono condizionamene indipendenti dati i propri genitori, in linea con l'assunzione di indipendenza di Markov.
Inoltre nel caso di partizione disintegrante dove M e Z vengono separati, ossia $\{(M, \emptyset), (\emptyset, Z)\}$ allora :

$$\pi_e^\theta(z'_e \mid m) = \pi_e^\theta(z'_e \mid \emptyset) = \pi_e^\theta(z'_e) = |\Omega_S|^{-1} \prod_{i=1}^{|Z|} \sum_{s \in \Omega_s} p_e(z'_{e_i} \mid s) \tag{40}$$

ossia il valore medio di z'_e su tutti gli stati di s di S.
Si definisce dunque, in linea con la formula (19), l'informazione integrata del meccanismo m sul purview Z sullo stato z'_e per la particolare partizione θ :

$$\varphi_e(m, Z, \theta) = \pi_e(z'_e \mid m) \left| \log \left(\frac{\pi_e(z'_e \mid m)}{\pi_e^\theta(z'_e \mid m)} \right) \right|_+ \tag{41}$$

L'effetto di m su z'_e è riducibile se almeno una partizione non fa differenza nella probabilità condizionate tra sistema integro e partizionato, oppure aumenta le probabilità rispetto al sistema non partizionato. In linea con il principio di esistenza minimale, l'informazione integrata deve essere valutata rispetto al MIP (minimum information partition) e, come nel caso dell'informazione integrata del sistema S, abbiamo analogamente :

$$\varphi_e(m, Z) := \varphi_e(m, Z, \theta') \tag{42}$$

che richiede una ricerca tra tutte le $\theta \in \Theta(M, Z)$

$$\theta' = \underset{\theta \in \Theta(M,Z)}{\text{argmin}} \frac{\varphi(m, Z, \theta)}{\max_{T'} \varphi(m, Z, \theta)} \tag{43}$$

La ricerca deve essere fatta su ogni possibile sottoinsieme $Z \in S$ e per il principio di esistenza massimale e per il postulato dell'esclusione, deve individuare due stati z_e^* e z_c^* (di causa ed effetto) tra i sottoinsiemi di Z verso i quali viene massimizzata l'informazione intrinseca. Dunque l'informazione integrata di causa ed effetto è data da:

$$\varphi_c(m) = \varphi_c(m, z_c^*)$$

$$\varphi_e(m) = \varphi_c(m, z_e^*)$$

Per il principio di esistenza minima l'informazione integrata del meccanismo è il minimo tra i due:

$$\varphi_d(m) = min(\varphi_c(m, z_c^*)\ ,\ \varphi_c(m, z_e^*)) \tag{47}$$

La tripla

$$d(m) = (m\ ,\ \{z_c^*, z_e^*\}\ ,\ \varphi_d) = (m\ ,\ z^*\ ,\ \varphi_d)$$

prende il nome di "distinzione" ed è una delle due componenti della Φ-struttura.
In riferimento all'immagine sotto, per un sistema di due unità e spazio di stati binario, sono illustrati le tre distinzioni esistenti. Il sistema (A, B) si trova nello stato $(0,1)$. Si usano le lettere minuscole e maiuscole per evidenziare sia la variabile aleatoria, sia il suo stato. Si precisa quindi che in un contesto binario di stati, quando si parla dell'elemento, unità, o variabile aleatoria a per esempio, si intende l'unità A nello stato 0, mentre se si parla dell'unità B si intende l'unità B nello stato 1. In questo stato, il sistema aB ha come purview di causa l'elemento B in stato b (indicato in minuscolo), ed ha come purview di effetto Ab, ossia (AB) in stato $(1,0)$ e valore in informazione integrata 0.07. Gli elementi di (AB) presi singolarmente, sono dei meccanismi ed esercitano il loro potere di causa ed effetto nel modo illustrato nella parte inferiore della figura. Il calcolo computazione della composizione non viene affrontato esplicitamente, in quanto molto simile al calcolo affrontato in precedenza per l'integrazione del sistema S. Inoltre non vengono introdotte, cosi come nel caso delle partizioni direzionali, i tagli direzionali che influivano con un fattore $3^k - 2^{k+1} + 2$, quindi per quanto complessivamente il costo computazionale aumenti, esso non cambia a livello asintotico i limiti proposti dai numeri di Bell.

3.5.2 Relazioni

Nella filosofia dell'IIT l'esistenza è sinonimo di causa ed effetto, dunque una volta considerato ogni "fonte" di questo, allora tutto è considerato e non c'è più spazio per altro da includere. Le relazioni causali catturano il concetto di come le cause e gli effetti di un insieme di distinzioni si

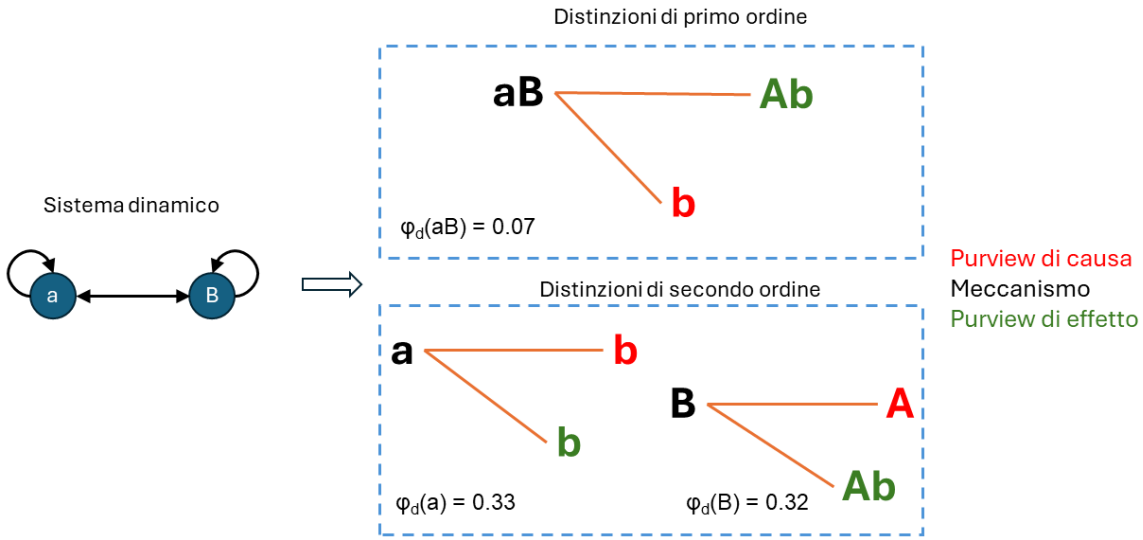

Figura 3.9
Illustrazione grafica delle tre distinzioni presenti nel sistema di due unità.

sovrappongono, e riflettono come il potere di causa ed effetto sia "legato insieme" all'interno del complesso. L'irriducibilità di questo legame viene misurato dall'informazione integrata delle relazioni φ_r.

Con riferimento all'esempio sopra elencato, sono possibili le seguenti relazioni. Tre auto-relazioni (self-relations) $r(aB)$, dove i purview di causa ed effetto hanno in comune l'elemento b, la relazione $r(a)$ che ha in comune l'elemento b, e $r(B)$ che ha in comune l'elemento A. Tre relazioni tra due distinzioni (viene indicato solo il meccanismo per identificare la distinzione) : $r(aB, a)$, $r(aB, B)$ e $r(a, B)$. Una relazione tra tre distinzioni $r(aB, a, B)$, per un totale di 7. Nell'immagine sotto viene resa un'idea intuitiva, prima delle definizioni formali, delle relazioni che legano queste distinzioni. Consideriamo per semplicità solo la relazione $r(aB, a)$. Gli elementi di questa relazione sono le tuple con almeno un'unità in comune tra i purview di causa ed effetto delle distinzioni. In figura 3.10 vengono indicate esattamente le 11 tuple, che sono state divise in tre immagini per poterle individuare meglio. Queste tuple sono le "facce" della relazione. Il grado di una relazione è il numero delle distinzioni che lega nell'esempio 2. Il numero di purview legati da una faccia rappresenta il suo grado. Nell'esempio ci sono 6 facce di grado 2 (i segmenti blu nell'immagine 1), 4 facce di grado 3 (i triangoli azzurri nell'immagine uno e due), ed una faccia di grado 4 (il parallelogramma azzurro nell'immagine 3). L'unità b è l'intersezione di tutte le facce e viene chiamato "face purview".

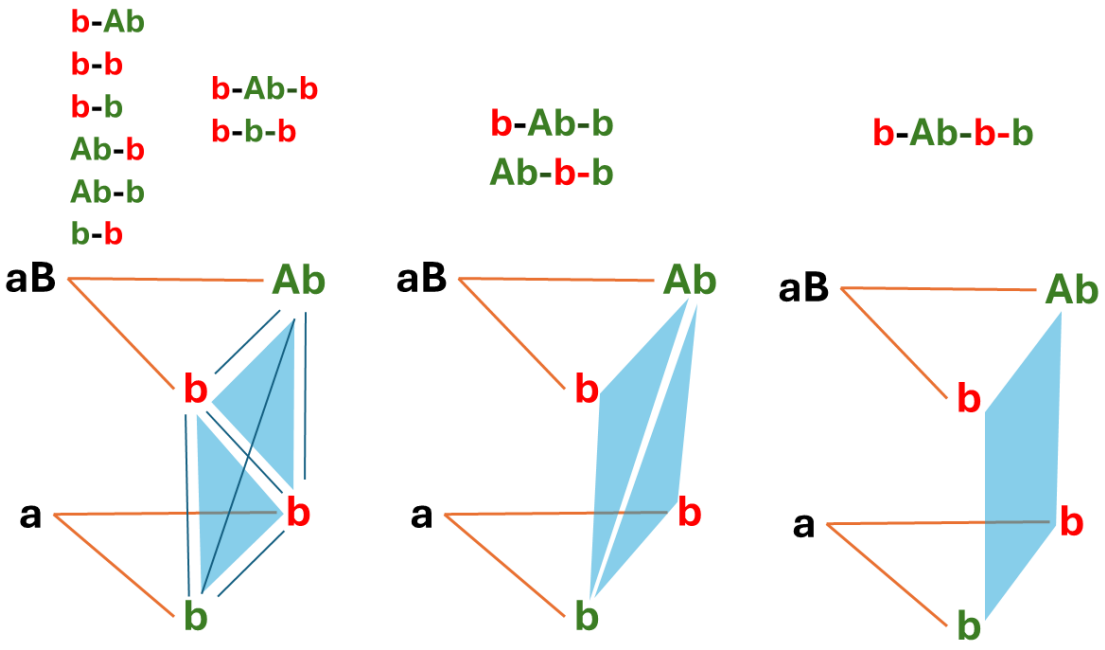

Figura 3.10
Rappresentazione grafica delle facce della relazione $r(aB, a)$.

Formalmente, dato un insieme di distinzioni $\boldsymbol{d}$ [4], con relativi purview di causa ed effetto $\{z_c^*(d), z_e^*(d)\}$, ci sono più insiemi $\boldsymbol{z}$ [5], dette facce, che li legano tali che :

$$\boldsymbol{z} : \boldsymbol{z} \cap \{z_c^*(d), z_e^*(d)\} \neq \emptyset \;\; \forall d \in \boldsymbol{d}, \bigcap_{z \in \boldsymbol{z}} z \neq \emptyset, |\boldsymbol{z}| > 1 \tag{49}$$

$\boldsymbol{z}$ dunque è un insieme di purview di causa o effetto che hanno almeno un' unità in comune. Viene inoltre definito il "face purview"

$$o^*(\boldsymbol{z}) = \bigcap_{z \in \boldsymbol{z}} z \neq \emptyset \tag{50}$$

come l'insieme delle unità in comune. Una faccia della relazione consiste nella tupla :

$$f(\boldsymbol{z}) = (\boldsymbol{z}, o^*(\boldsymbol{z})) \tag{52}$$

[4] d rappresenta una distinzione, mentre $\boldsymbol{d}$ rappresenta un'insieme di distinzioni.
[5] z rappresenta un singolo purview di causa o effetto, mentre $\boldsymbol{z}$ rappresenta un'insieme di purview.

mentre $\boldsymbol{f}(\boldsymbol{d}) = \{f(\boldsymbol{z})\}_{\boldsymbol{d}}$ rappresenta l'insieme di tutte le facce della relazione.
Una relazione $r(\boldsymbol{d})$ consiste dunque nella tupla contenente $\boldsymbol{d}$, l'insieme di facce $\boldsymbol{f(d)}$ ed il valore di irriducibilità $\varphi_r > 0$[6].

$$r(\boldsymbol{d}) \ = \ (\boldsymbol{d}, \boldsymbol{f(d)}, \varphi_r) \tag{51}$$

L'irriducibilità di una relazione si misura "slegando" le distinzioni dai loro purview comuni. Per determinare l'irriducibilità di una relazione assumiamo che il valore dell'informazione integrata di una distinzione sia distribuita uniformemente tra gli elementi unici dei suoi purview. Ossia, per una distinzione, il valore medio trasmesso alle unità uniche tra purview di causa ed effetto è dato da:

$$\frac{\varphi_d}{|z_c^*(d) \ \cup \ z_e^*(d)|} \tag{53}$$

Per fare un esempio di ciò, nell'immagine in figura 3.9 e 3.10 il meccanismo aB (intendendo la distinzione che esso forma) contribuisce il suo valore di informazione integrata $\varphi_d(aB)$ su $b \cup Ab$ ed essendo$A$ e b gli elementi unici dell'unione , abbiamo $|aB| = 2$, inteso come la cardinalità dell'insieme di 2 elementi. Siccome le distinzioni distribuiscono il loro φ_d sugli elementi unici dell'unione tra purview di causa ed effetto, allora il loro "scollegamento" deve essere proporzionale numero di unità uniche di tutte le facce della relazione :

$$\left| \bigcup_{f \in \boldsymbol{f(d)}} o_f^* \right| \tag{54}$$

Le o_f^* sono quelle definite in formula (50) e la loro unione si chiama il "purview relazione", o "purview unione" ("relation purview" e "joint purview" in inglese). Si faccia riferimento alla figura 3.11 per un esempio.

[6]Viene richiesto $\varphi_r > 0$ perché quando si considerano potenziali relazioni tra insiemi di distinzioni con *joint purview* vuoto, esse non rappresentano delle relazioni.

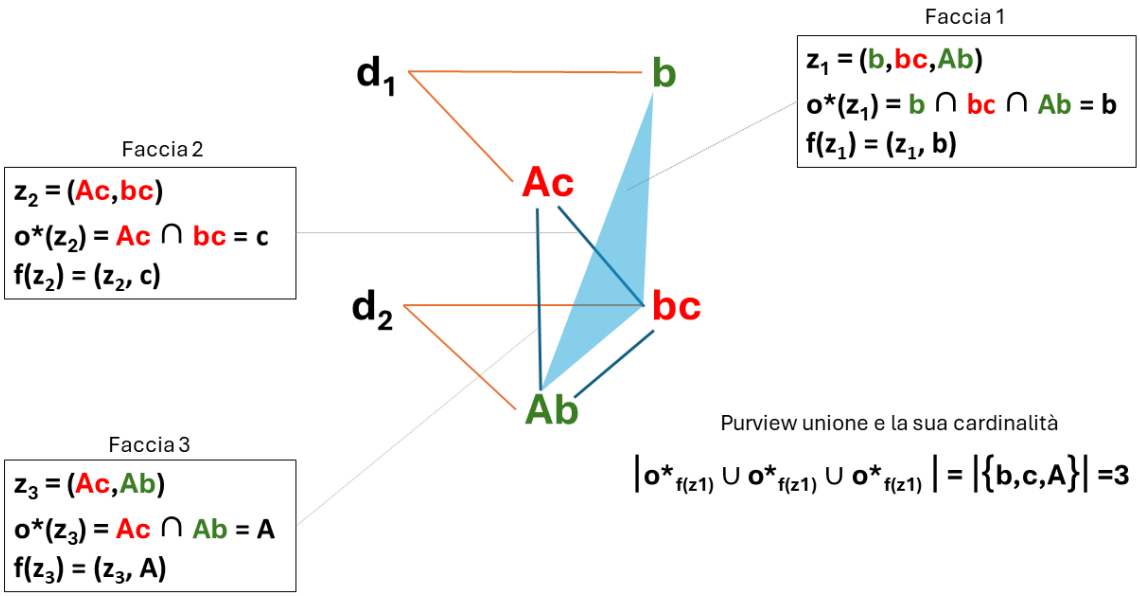

Figura 3.11
Esempio di relazione $r(d_1, d_2)$ tra due ipotetiche distinzioni d_1 e d_2. L'esempio non rappresenta il sistema in figura 3.10, ma un ipotetico sistema creato per illustrare meglio il concetto di *face purview* e *joint purview/relation purview*. La faccia (Ab, bc) appartiene all'autorelazione $r(d_2)$ e dunque non viene considerata.

Per il principio dell'esistenza minima si definisce come informazione integrata di una relazione la minima quantità di informazione integrata tra tutte quelle che ogni distinzione specifica:

$$\varphi_r(\boldsymbol{d}) = \min_{d \in \boldsymbol{d}} \left| \bigcup_{f \in \boldsymbol{f}(\boldsymbol{d})} o_f^* \right| \frac{\varphi_d}{|z_c^*(d) \ \cup \ z_e^*(d)|} \tag{55}$$

Considerando dunque D l'insieme delle distinzioni in un complesso, possiamo definire l'insieme delle relazioni tra di esse :

$$R(D) = \{r(\boldsymbol{d}) : \varphi_r(\boldsymbol{d}) \ > \ 0\}, \ \forall \boldsymbol{d} \subseteq D \tag{56}$$

Si considerano dunque tutti i sottoinsiemi $\boldsymbol{d}$ di D per catturare tutte le relazioni di grado $|\boldsymbol{d}|$, e tra di esse, considerare solo quelle con informazione integrata maggiore di zero.

Questa struttura di causa ed effetto, indicata con la lettera C, prende il nome di Φ-struttura. Viene definita come l'unione tra distinzioni e relazioni di un complesso S^* in stato s^*, dato anche la sua completa descrizione in termini delle matrici di transizione T_c e T_e :

$$C = \{\{d(m) = \{m, z^*, \varphi_d\} \in D\} \cup \{r(\boldsymbol{d}) = \{\boldsymbol{d}, \boldsymbol{f(d)}, \varphi_r\} \in R(D)\}\} \quad (58)$$

L'operazione di calcolare e "costruire" di questa struttura prende il nome di "dispiegamento" (unfolding in inglese) di uno substrato di coscienza.

3.5.3 Complessità computazionale

Per avere un'idea approssimativa della complessità computazionale del calcolo delle relazioni partiamo dal calcolare il loro numero. Dal punto di vista combinatorio può essere calcolato come segue, nel caso peggiore. Le relazioni sussistono su sottoinsiemi delle distinzioni, come da formula (56). Il numero di distinzioni di un complesso S, che nel caso peggiore può essere l'universo stesso U di n unità, è dato da $|D| = 2^n - 1$ visto che ogni sottoinsieme di S può essere una distinzione, tranne il sottoinsieme vuoto. Dobbiamo dunque calcolare le auto-relazioni (relazioni di grado uno), le relazioni di grado due, e cosi via per un numero totale di relazioni :

$$|R(D)| = 2^{|D|} - 1 = 2^{(2^n - 1)} - 1$$

Tuttavia, se siamo interessati solo a calcolare l'informazione integrata delle relazioni, non necessariamente ci troviamo di fronte ad un'esplosione combinatoria di questo tipo. Nel supplemento [13, p. 4] della pubblicazione viene illustrato un modo per calcolare φ_r in $\mathcal{O}(n^2 2^n)$, mentre il costo per il calcolo di tutte le relazioni rimane dell'ordine di $\mathcal{O}(2^{2^n})$

3.6 Φ-struttura

La somma dei valori dell'informazione integrata delle distinzioni e relazioni viene chiamato grande Φ (big Phi in inglese), e corrisponde all'informazione integrata strutturale della Φ-struttura.

$$\Phi = \sum_{d \in D} \varphi_d \; + \sum_{r \in R(D)} \varphi_r = \sum_{c \in C} \varphi_c \tag{59}$$

Ogni elemento $c \in C$ rappresenta una distinzione oppure una relazione. Il loro rispettivo valore di informazione integrata è definito in (47) e (55). Questa definizione cattura il concetto di "quantità" di coscienza presente nel sistema. Con riferimento alla figura 2.1 la Φ-struttura viene rappresentata graficamente come la collezione di distinzioni (nel senso di meccanismi e purview), che sono i suoi vertici e le facce delle relazioni, rappresentate come linee o superfici tra i vertici. Nel caso semplice di un sistema come quello descritto in figura 3.9, una visualizzazione parziale della Φ-struttura risulterebbe quella data dall'unione delle due immagini rappresentate in figura 3.10. L'identità esplicativa centrale dell'IIT può dunque essere riassunta dall'immagine seguente:

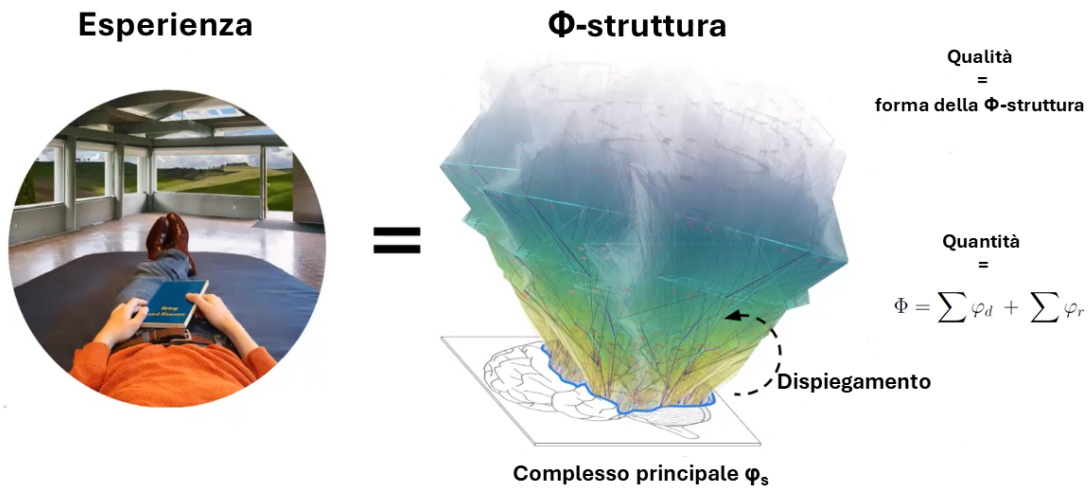

Figura 3.12
Rappresentazione dell'identità centrale dell'IIT.

Capitolo 4
Conclusione

La Integrated Information Theory (IIT) rappresenta un passo concreto in avanti nel panorama mondiale della ricerca sulla coscienza. Dal punto di vista metodologico parte dallo studio della fenomenologia stessa della coscienza. Propone cinque proprietà fondamentali sotto forma di assiomi che si assumono immediatamente e indubbiamente vere per ogni esperienza cosciente. Formula delle ipotesi su come tradurre questi assiomi in postulati operativi e li descrive con una rigorosa struttura matematica. Formula delle previsioni e le verifica attraverso esperimenti neuroscientifici. La teoria nasce nel contesto dello studio del cervello, ma estende la sua applicazione anche a sistemi artificiali seguendo il principio che la coscienza emerge dalla capacità di un sistema interconnesso di integrare l'informazione.

L'IIT ambisce sicuramente tanto in alto, cercando di spiegare non solo quanta coscienza contiene un determinato sistema, calcolato dalla misura Φ, ma anche la qualità di essa. Qualità che è identica alla struttura causa-effetto, chiamata Φ-struttura, dispiegata da un complesso. C'è una corrispondenza uno a uno tra il modo in cui l'esperienza viene percepita e il modo in cui la Φ-struttura è composta. Dunque ogni proprietà dell'esperienza è contenuta all'interno della struttura, quale combinazione di distinzioni e relazioni. Non risulta chiaro però quali siano queste combinazioni, dove siano esattamente, e perché siano fatte in quel particolare modo.

Con l'ammissione degli stessi autori della pubblicazione[3, p. 41] le sfide che la teoria deve superare sono molteplici. È essenziale garantire che il quadro assiomatico sia solido, che le assunzioni risultino fondate e le sue predizioni siano compatibili con le evidenze scientifiche esistenti e future. È importante riconoscere che la storia della scienza offre molteplici esempi di come assunzioni, una volta universalmente accolte, abbiano subito revisioni o siano state completamente rovesciate. Per esempio, prima dei teoremi di incompletezza di Gödel, molti matematici e filosofi assumevano che la matematica fosse completa, nel senso che ogni affermazione matematica vera potesse essere dimostrata. Altri esempi significativi includono la credenza nell'esistenza dell'etere, l'idea di un universo statico e di un tempo assoluto, il modello geocentrico del sistema solare ecc. Questo non promuove una critica alla teoria, ma vuole dimostrare che la possibilità di essere messa in discussione ha il potere di far progredire la nostra comprensione sul campo.

L'IIT ha subito notevoli variazioni e miglioramenti negli anni, ma non è ancora noto un metodo stabile per approssimare il calcolo di Φ e della Φ-struttura. Il calcolo sulla complessità computazionale presentato in questa tesi e l'uso della libreria python dedicata *pyphi* [14], evidenzia i

limiti operativi per sistemi con più di una quindicina di unità.
Nonostante le sfide evidenziate, la ricerca nell'IIT è molto attiva, ad esempio nello sviluppo di una versione compatibile con la meccanica quantistica [15] o, nel determinare la presenza o meno di coscienza nei computer e in sistemi di intelligenza artificiale [16][17].

Bibliografia

[1] D. Chalmers, «The Hard Problem of Consciousness,» in *The Blackwell Companion to Consciousness*. John Wiley Sons, Ltd, 2017, cap. 3, pp. 32–42, isbn: 9781119132363. doi: https://doi.org/10.1002/9781119132363.ch3. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119132363.ch3. indirizzo: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119132363.ch3.

[2] A. Seth e T. Bayne, «Theories of consciousness,» *Nature Reviews Neuroscience*, vol. 23, mag. 2022. doi: 10.1038/s41583-022-00587-4.

[3] L. Albantakis, L. Barbosa, G. Findlay et al., «Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms,» *PLOS Computational Biology*, vol. 19, n. 10, pp. 1–45, ott. 2023. doi: 10.1371/journal.pcbi.1011465. indirizzo: https://doi.org/10.1371/journal.pcbi.1011465.

[4] J. R. Hanson, «Falsification of the Integrated Information Theory of Consciousness,» tesi di dott., Arizona State University, 2021.

[5] M. A. Cerullo, «The Problem with Phi: A Critique of Integrated Information Theory,» *PLOS Computational Biology*, vol. 11, n. 9, pp. 1–12, set. 2015. doi: 10.1371/journal.pcbi.1004286. indirizzo: https://doi.org/10.1371/journal.pcbi.1004286.

[6] «The Integrated Information Theory of Consciousness as Pseudoscience,» doi: 10.31234/osf.io/zsr78. indirizzo: https://doi.org/10.31234/osf.io/zsr78.

[7] «Causal Models,» indirizzo: https://plato.stanford.edu/entries/causal-models/.

[8] L. Fenton-Glynn, «A Proposed Probabilistic Extension of the Halpern and Pearl Definition of 'Actual Cause',» *The British Journal for the Philosophy of Science*, vol. 68, n. 4, p. 23, 2017. doi: 10.1093/bjps/axv056. eprint: https://doi.org/10.1093/bjps/axv056. indirizzo: https://doi.org/10.1093/bjps/axv056.

[9] «Probabilistic causation,» indirizzo: https://en.wikipedia.org/wiki/Probabilistic_causation.

[10] J. Pearl e D. Mackenzie, *The Book of Why: The New Science of Cause and Effect*, 1st. USA: Basic Books, Inc., 2018, isbn: 046509760X.

[11] W. Marshall, M. Grasso, W. G. P. Mayner et al., «System Integrated Information,» *Entropy*, vol. 25, n. 2, 2023, issn: 1099-4300. doi: 10.3390/e25020334. indirizzo: https://www.mdpi.com/1099-4300/25/2/334.

[12] D. Berend e T. Tassa, «Efficient Bounds on Bell Numbers and on Moments of Sums of Random Variables,» *Probability and Mathematical Statistics*, vol. 30, gen. 2010.

[13] «Analytical results for the number and integrated information of relations.,» doi: 10.1371/journal.pcbi.1011465.s003. indirizzo: https://doi.org/10.1371/journal.pcbi.1011465.s003.

[14] W. G. P. Mayner, W. Marshall, L. Albantakis, G. Findlay, R. Marchman e G. Tononi, «PyPhi: A toolbox for integrated information theory,» *PLOS Computational Biology*, vol. 14, n. 7, pp. 1–21, lug. 2018. doi: 10.1371/journal.pcbi.1006343. indirizzo: https://doi.org/10.1371/journal.pcbi.1006343.

[15] L. Albantakis, R. Prentner e I. Durham, «Computing the Integrated Information of a Quantum Mechanism,» *Entropy*, vol. 25, n. 3, 2023, issn: 1099-4300. doi: 10.3390/e25030449. indirizzo: https://www.mdpi.com/1099-4300/25/3/449.

[16] G. Findlay et al., «Dissociating Intelligence from Consciousness in Artificial Systems - Implications of Integrated Information Theory,» in *Proceedings of the 2019 Towards Conscious AI Systems Symposium, AAAI Spring Symposium Series*, AAAI SSS19, 2019.

[17] Tiny Blue Dot Foundation. «Estimate Integrated Information for Computer Architectures and Demonstrate a Dissociation Between Artificial Intelligence and Consciousness.» Accessed: 2024-04-08, University of Wisconsin-Madison / Department of Psychiatry / Center for Sleep e Consciousness. (2023), indirizzo: https://www.tinybluedotfoundation.org/uwm/estimate-integrated-information-for-computer-architectures-and-demonstrate-a-dissociation-between-artificial-intelligence-and-consciousness.