Spazi multidimensionali per la rappresentazione semantica

Marcello Frixione e Antonio Lieto
Università di Genova e Università di Torino
marcello.frixione@unige.it - antonio.lieto@unito.it

Sommario

Nel campo delle scienze cognitive molti oggi condividono l'ipotesi che siano necessari differenti tipi di rappresentazioni per modellare i sistemi cognitivi sia naturali, sia artificiali. Si considerino le rappresentazioni basate su reti neurali, i formalismi simbolici e rappresentazioni analogiche quali rappresentazioni diagrammatiche o modelli mentali. Tutti questi metodi hanno successo nello spiegare e modellare alcune classi di fenomeni cognitivi, ma nessuno è in grado di rendere conto di tutti gli aspetti della cognizione.

A partire da queste considerazioni, riteniamo che sistemi intelligenti e architetture cognitive possano trarre vantaggio dalla combinazione di sistemi di rappresentazione diversi. Si pone allora il problema di fare interagire rappresentazioni di natura differente in maniera cognitivamente e teoricamente fondata. Esistono in fatti modelli ibridi che combinano in modi diversi i tipi di sistemi sopra menzionati. Tuttavia tale integrazione è di solito ad hoc e solo in parte soddisfacente.

La nostra ipotesi è che gli spazi concettuali, un modello di rappresentazione proposto da Peter Gärdenfors, possano offrire una sorta di lingua comune, che consentirebbe di integrare e generalizzare molti aspetti delle impostazioni sopra menzionate, superando i limiti delle varie proposte intese singolarmente.

Parole chiave: Rappresentazione della conoscenza,modelli cognitivi, architetture cognitive, spazi concettuali, modelli distribuzionali

1. Introduzione

Nel campo delle scienze cognitive molti oggi condividono l'ipotesi che siano necessari tipi differenti di rappresentazioni per modellare i sistemi cognitivi sia naturali, sia artificiali. Si considerino ad esempio la vasta classe delle rappresentazioni basate su reti neurali (incluse le cosiddette reti neurali profonde – *deep neural networks*), i molteplici tipi di formalismi simbolici (che includono la logica e i formalismi probabilistici e bayesiani) e le rappresentazioni di tipo analogico, quali immagini mentali, rappresentazioni diagrammatiche, modelli mentali. Esistono poi modelli ibridi che combinano in modi diversi i tipi di sistemi sopra menzionati. Ad esempio, nell'ambito delle architetture cognitive, alcuni sistemi (come ad esempio SOAR – Laird 2012) adottano un approccio simbolico classico; altri (ad esempio LEABRA – O'Reilly e Munakata 2000) sono basati su modelli puramente confessionisti; altri ancora (ad esempio CLARION – Sun 2006) adottano un'impostazione ibrida che combina livelli di rappresentazione simbolica e connessionista. Vi sono inoltre tentativi (ad esempio biSOAR) di estendere le capacità delle architetture cognitive mediante forme di rappresentazione e ragionamento diagrammatico (Kurup e Chandrasekaran 2007).

Tutti questi metodi hanno successo nello spiegare e modellare alcune classi di fenomeni cognitivi, ma nessuno è in grado di rendere conto di tutti gli aspetti della cognizione. Ciò risulta evidente anche se prendiamo in considerazione alcuni recenti sistemi artificiali di successo. Ad esempio, il sistema Watson è basato su tecniche probabilistiche in grado di gestire masse enormi di dati, ma fallisce in compiti banali di ragionamento di senso comune (si veda ad esempio Davis e Marcus 20015, p. 94). Analogamente, il sistema AlphaGo (Silver *et al.* 2016), basato sull'addestramento di *deep neural networks*, esibisce prestazioni impressionanti nel dominio ben definito del gioco Go, ma non è in grado di trasferire le sue abilità a domini di carattere più generale. Si tratta di un limite tipico delle reti neurali: per risolvere un determinato problema vengono addestrate su un vasto insieme di dati specifici; tuttavia, la capacità di trasferire a domini simili le capacità apprese è in genere difficile.

A partire da queste considerazioni, riteniamo che sistemi intelligenti e architetture cognitive possano trarre vantaggio dalla combinazione di sistemi di rappresentazione diversi. Si pone allora il problema di fare interagire rappresentazioni differenti in maniera cognitivamente e teoricamente fondata. Infatti, come si è detto, esistono molti sistemi e architetture ibride (Sun 2006) che combinano tipi diversi di rappresentazione

(si veda ad esempio la classe dei sistemi neuro-simbolici – D'Avila *et al*. 2008); tuttavia tale integrazione è di solito ad hoc (Chella *et al*. 2003) e, come vedremo, solo in parte soddisfacente.

La nostra ipotesi è che gli spazi concettuali, un modello di rappresentazione proposto da Peter Gärdenfors da oramai due decenni (Gärdenfors 1997, 2000) possano offrire una sorta di lingua comune, che consentirebbe di integrare e generalizzare molti aspetti delle impostazioni sopra menzionate, superando i limiti delle varie proposte intese singolarmente (Lieto, Chella e Frixione 2017).

Questo capitolo è organizzato come segue: nel par. 2 ritorniamo sul tema dell'eterogeneità delle rappresentazioni prendendo come esempio il problema specifico della rappresentazione dei concetti. Nel par. 3 consideriamo questo atteggiamento pluralistico nel contesto dell'intelligenza artificiale, concentrandoci su alcuni tra i tipi di rappresentazione più diffusi in letteratura. Il par. 4 fornisce una introduzione sintetica agli spazi concettuali e caratterizza tale tipo di rappresentazione rispetto ai modelli distribuzionali (oggi molto utilizzati nell'ambito della linguistica computazionale). Il par. 5 è dedicato ai vantaggi offerti dagli spazi concettuali nell'integrazione delle diverse forme di rappresentazione. Segue una breve conclusione.

Ringraziamo Antonio Chella, che ha contribuito all'elaborazione delle idee presentate in questo capitolo.

2. Eterogeneità delle rappresentazioni nelle scienze cognitive: il caso dei concetti

In questo paragrafo presenteremo alcuni elementi provenienti dalle scienze cognitive a favore dell'ipotesi dell'eterogeneità delle rappresentazioni nei sistemi e nelle architetture cognitive. In particolare, prenderemo in considerazione due tipi di evidenza che concerne le rappresentazioni concettuali: il problema della rappresentazione dei concetti non classici (par. 2.1) e l'applicazione alla conoscenza concettuale della dell'ipotesi dei processi duali (par. 2.2).

2.1 La rappresentazione dei concetti non classici

In scienza cognitiva sono state proposte diverse teorie su come gli esseri umani rappresentano, organizzano e utilizzano nel ragionamento la conoscenza concettuale. Secondo il punto di vista tradizionale, noto come teoria classica dei concetti, i concetti sono rappresentati nei termini di insiemi di condizioni necessarie e sufficienti. Tale teoria ha avuto un ruolo dominante sia in filosofia, sia in psicologia dall'antichità fino alla seconda metà del secolo scorso, quando le analisi filosofiche (Wittgenstein 1953) come anche i risultati empirici della ricerca psicologica (Rosch 1975) ne dimostrarono l'inadeguatezza. Fu chiaro che la maggior parte dei concetti ordinari esibiscono effetti prototipici: per caratterizzare moltii concetti risultano fondamentali informazioni che non costituiscono condizioni necessarie o sufficienti per la loro applicazione, ma corrispondono invece a tratti che valgono in casi tipici. I risultati ottenuti da Rosch ebbero una rilevanza cruciale per lo sviluppo di teorie dei concetti nuove, che aspiravano a spiegare gli aspetti di rappresentazione e di ragionamento legati alla tipicità. Usualmente, queste teorie sono raggruppate in tre grandi classi: (i) teorie dei prototipi, che derivano direttamente dai lavori di Rosch; (ii) teorie degli esemplari e (iii) teorie della teoria (si vedano ad esempio Murphy 2002 e Machery 2009 per una rassegna dettagliata di tali impostazioni). Tutte queste teorie aspirano a rendere conto di qualche aspetto degli effetti prototipici nella concettualizzazione. Secondo le teorie del primo gruppo, la conoscenza concettuale è rappresentata nei termini di prototipi, ossia rappresentazioni dell'istanza più rappresentativa di ciascuna categoria. Per esempio, la rappresentazione del concetto GATTO coinciderebbe con la rappresentazione di un gatto tipico. Nella versione più semplice, i prototipi sarebbero codificati come liste di caratteristiche tipiche (eventualmente pesate). Secondo le teorie degli esemplari, un concetto sarebbe memorizzato nei termini di un insieme di rappresentazioni di specifici esemplari: la rappresentazione mentale del concetto GATTO coinciderebbe con la rappresentazione dell'insieme (di alcuni) dei gatti che il soggetto ha incontrato nel scorso della sua esistenza. Le teorie della teoria adottano un punto di vista olistico rispetto alla rappresentazione concettuale. Secondo alcune versioni di questa impostazione, i concetti sarebbero analoghi a termini teorici in una teoria scientifica. Ad esempio, il concetto GATTO sarebbe individuato dal ruolo che esso gioca nella nostra teoria mentale della zoologia. Secondo altre versioni di questo approccio, i concetti stessi sarebbero da identificare con una sorta di micro teoria. Ad esempio, il concetto GATTO coinciderebbe con la rappresentazione mentale di una micro teoria sui gatti. Sebbene inizialmente queste impostazioni siano state proposte come tra loro alternative, i dati empirici sembrano suggerire (a partire da Barbara Malt 1989) che in realtà non siano mutualmente esclusive. Esse infatti sembrano avere successo nello spiegare fenomeni cognitivi differenti. In particolare, dati comportamentali quali la probabilità di categorizzazione o i tempi di reazione suggeriscono che i soggetti possano usare rappresentazioni differenti in compiti diversi. Alcuni soggetti sembrano preferire i prototipi, altri impiegano rappresentazioni basate sugli esemplari, altri ancora sembrano far ricorso a entrambi i tipi di rappresentazione. Inoltre alcune

rappresentazioni sembrano più adatte per certi compiti oppure per certi tipi di categorie. Infine, questa distinzione sembra avere anche plausibilità neurale, come è testimoniato da vari risultati empirici (a partire da Squire e Nolton 1995). Tutto ciò portò allo sviluppo della cosiddetta ipotesi eterogenea circa la natura delle rappresentazioni concettuali: i concetti non costituirebbero un fenomeno cognitivo unitario; viceversa, coesisterebbero nella mente tipi diversi di rappresentazioni concettuali.

2.2 Teorie dei processi duali e rappresentazioni concettuali

Una suddivisione ulteriore tra tipi differenti di rappresentazione concettuale fa riferimento alle teorie dei processi duali. Secondo tali teorie (Stanovich e West 2000, Evans e Frankish 2009, Kahneman 2011) esisterebbero due tipi di sistemi e di processi cognitivi distinti, indicati rispettivamente come Sistema 1 e Sistema 2. I processi relativi al Sistema 1 sono automatici; essi sono filogeneticamente più antichi e condivisi tra specie umana e altre specie animali; sono innati e controllano i comportamenti istintivi, e pertanto non dipendono dall'addestramento o da capacità individuali specifiche; sono cognitivamente poco dispendiosi; sono associativi e operano in modo parallelo e veloce. Inoltre, i processi di tipo 1 non sono direttamente accessibili alla coscienza dei soggetti. I processi relativi al Sistema 2 sono filogeneticamente più recenti e sono peculiari della specie umana; sono consci e cognitivamente penetrabili (cioè accessibili alla coscienza) e basati sull'applicazione esplicita di regole. Di conseguenza, se comparati a quelli del sistema 1, i processi del sistema 2 sono lenti e sequenziali, e cognitivamente impegnativi. Le prestazioni che dipendono dal sistema 2 risentono dell'apprendimento e di differenze nelle capacità individuali.

L'ipotesi dei processi duali è stata inizialmente proposta per spiegare gli errori sistematici di ragionamento. Tali errori (si considerino gli esempi classici del *selection task* e della fallacia della congiunzione) sarebbero ascrivibili ai processi di tipo 1, veloci, associativi ed automatici, mentre il sistema 2 sarebbe responsabile dell'attività lenta e cognitivamente impegnativa di produrre risposte corrette secondo i canoni della razionalità normativa. In generale, molti aspetti della psicologia dei concetti hanno presumibilmente a che fare con sistemi e processi di tipo 1, mentre altri possono plausibilmente essere ascritti al tipo 2. Ad esempio, la categorizzazione basata su tratti prototipici (siano essi rappresentati in termini di prototipi, di esemplari o di teorie), come ad esempio categorizzare Fido come un cane perché scodinzola, abbaia, eccetera, è presumibilmente un processo veloce ed automatico, che non richiede alcuno sforzo esplicito e che può essere facilmente attribuito al sistema 1. Viceversa, ci sono tipi di inferenze che di solito vengono inclusi tra le capacità concettuali, che sono lenti e cognitivamente impegnativi, e che potrebbero più plausibilmente essere attribuiti a processi di tipo 2. Si consideri la capacità di fare inferenze esplicite di alto livello che coinvolgono la conoscenza concettuale, e la capacità di giustificarle. O si consideri la classificazione: classificare un concetto comporta individuare i suoi superconcetti più specifici e i suoi sottoconcetti più generali, o, in altri termini, individuare relazioni di superconcetto e di sottoconcetto che sono implicite in una tassonomia. Per un soggetto umano si tratta di un compito impegnativo, lento, che è facilitato da un addestramento specifico. Pertanto, secondo la teoria dei processi duali, il compito inferenziale di classificare i concetti in una tassonomia sembra essere *prima facie* un processo di tipo 2, qualitativamente differente da un compito di categorizzazione basato su tratti tipici. Di conseguenza, è plausibile che le rappresentazioni concettuali nei sistemi cognitivi facciano riferimento ad (almeno) due tipi diversi di componenti responsabili di compiti diversi. Processi di tipo 2 sarebbero coinvolti in compiti complessi e cognitivamente impegnativi, e processi di tipo 1 veloci ed automatici sarebbero coinvolti in compiti quali la categorizzazione basata su conoscenza del senso comune. Una posizione teorica di questo tipo è stata difesa da Piccinini (2011), secondo cui esisterebbero due tipi di concetti: impliciti ed espliciti, che sarebbero da mettere in relazione rispettivamente con Sistema 1 e Sistema 2. Più recentemente è stato sostenuto (Frixione e Lieto 2013) che un modello artificiale cognitivamente plausibile della rappresentazione concettuale dovrebbe essere basato su un approccio duale, e pertanto formato da componenti differenti basate su rappresentazioni diverse. Alcuni sistemi proposti sono stati sviluppati sulla base di questa ipotesi (Lieto *et al.* 2015, 2017) e integrati in architetture cognitive quali ACT-R (Anderson *et al.* 2004) e CLARION (Sun 2006).

3. Intelligenza artificiale e formalismi di rappresentazione

Riteniamo che la pluralità di rappresentazioni eterogenee osservata nei sistemi cognitivi naturali, che abbiamo descritto nei paragrafi precedenti, sia anche proficua nella progettazione di architetture e di sistemi cognitivi artificiali. In intelligenza artificiale (IA) sono stati proposti approcci diversi al problema della rappresentazione della conoscenza, che possono essere utilizzati per modellare l'eterogeneità delle rappresentazioni concettuali sopra descritta. In quanto segue, prenderemo in considerazione tre esempi

principali: le rappresentazioni simboliche (Par. 3.1), quelle basate su reti neurali (par. 3.2), e le rappresentazioni analogiche e diagrammatiche (Par. 3.3).

3.1 Rappresentazioni simboliche

Le rappresentazioni simboliche, che in molti casi si basano su qualche tipo di formalismo logico, sono per lo più adatte per realizzare compiti di ragionamento complesso. Tali sistemi sono caratterizzati dal fatto di impiegare rappresentazioni composizionali: in un sistema di rappresentazione composizionale si può distinguere tra un insieme di simboli primitivi o atomici, e un insieme di simboli complessi, che sono generati a partire dai simboli primitivi per mezzo di un insieme di regole sintattiche ricorsive. Di solito, a partire da un insieme finito di simboli primitivi, viene generato un insieme potenzialmente infinito di simboli complessi. Il significato dei simboli complessi può essere determinato a partire dal significato dei simboli primitivi attraverso un insieme di regole semantiche che operano in parallelo alle regole di composizione sintattica.

La composizionalità è considerata da alcuni un carattere irrinunciabile della cognizione umana; in particolare, nel contesto delle scienze cognitive classiche si assume che le rappresentazioni mentali debbano essere composizionali. Una formulazione chiara ed esplicita di questa assunzione la dobbiamo a Fodor e Pylyshyn (1988), nel contesto di un'aspra critica alle reti neurali e ai sistemi connessionisti. Questi autori ritengono che le rappresentazioni mentali non possano che essere composizionali per poter spiegare alcuni fenomeni cognitivi fondamentali (quali il carattere generativo e sistematico della cognizione umana), e che le reti neurali non possano soddisfare questo requisito.

La composizionalità è anche un carattere dei sistemi simbolici artificiali, come è testimoniato da molti formalismi di rappresentazione della conoscenza. Nel campo delle architetture cognitive, ad esempio, SOAR è uno dei sistemi più noti che impiegano rappresentazioni simboliche composizionali. Questo sistema aderisce rigorosamente all'ipotesi del sistema di simboli fisico di Newell e Simon (1976), in base al quale l'elaborazione simbolica sarebbe condizione necessaria e sufficiente per il comportamento intelligente.

Tuttavia, la composizionalità non può essere facilmente conciliata con alcuni fenomeni cognitivi. Ad esempio, è ben noto che esiste un conflitto tra composizionalità e rappresentazione dei concetti in termini prototipici (Fodor 1981, Osherson e Smith 1981, Frixione e Lieto 2011, si veda oltre il par. 5.1). Questo problema non è circoscritto allo studio empirico dei sistemi cognitivi naturali; è di grande rilevanza anche per la progettazione di architetture e di sistemi cognitivi artificiali. Il conflitto tra composizionalità e tipicità nei sistemi di rappresentazione simbolica è evidente nei modelli artificiali per la rappresentazione dei concetti. Si considerino ad esempio le logiche descrittive e i linguaggi per la rappresentazione di ontologie (ad es. OWL), che sono sistemi composizionali ma che non ammettono la rappresentazione di tratti tipici.

In sintesi, rappresentare i concetti in termini di tipicità è rilevante anche per le applicazioni computazionali (in particolare per quelle di ispirazione cognitiva), ma Fodor e Pylyshyn hanno probabilmente ragione nel sottolineare che le reti neurali hanno problemi con la composizionalità. D'altro canto, nello sviluppo di architetture e sistemi cognitivi computazionali, non siamo disposti a rinunciare ai vantaggi offerti dalle reti neurali. È plausibile che la composizionalità abbia a che fare con la cognizione di alto livello e con compiti inferenziali complessi, di tipo 2, mentre le reti neurali risultano più adatte per modellare fenomeni di tipo 1. Rimane però aperto il problema dell'interazione di queste due classi di formalismi. Nella tradizione dell'IA simbolica, un tentativo di mitigare questi problemi è stato proposto con le reti bayesiane (Nielsen e Jensen 2009), che sono una classe di rappresentazioni simboliche, dove alle relazioni tra concetti è associato un peso calcolato in termini statistici. Nonostante il recente successo dell'impostazione bayesiana nel modellamento di molti compiti cognitivi (Griffiths *et al.* 2008), una spiegazione completa della cognizione sia naturale, sia artificiale in termini di macchine bayesiane è tutt'altro che scontata. Nell'ambito della cognizione umana, molte forme di conoscenza del senso comune non sembrano richiedere predizioni bayesiane o, in generale, forme di ragionamento probabilistico (Sloman 2014). Inoltre, anche in queste rappresentazioni simboliche più sofisticate, permane il problema di riconciliare la composizionalità con la rappresentazione in termini di tipicità come vedremo nel par. 5.1.

3.2 Rappresentazioni basate su reti neurali

Le reti neurali sono una classe di rappresentazioni che è stata impiegata con successo in molti problemi di modellamento cognitivo. In generale, nel campo delle architetture cognitive, questo tipo di rappresentazioni è stato ampiamente impiegato per affrontare i comportamenti "veloci" di un sistema, e per modellare aspetti relativi all'apprendimento e alla percezione. Le reti neurali sono particolarmente adatte per compiti di classificazione. Di conseguenza, sono ampiamente impiegate in IA per molti problemi di *pattern recognition*: un esempio classico concerne il riconoscimento della scrittura manoscritta. A differenza delle rappresentazioni simboliche, le reti neurali possono ricevere dati di ingresso direttamente dai sistemi

percettivi (immagini, segnali, eccetera), e, di conseguenza, risulta alleviato il problema dell'ancoramento (*grounding*) delle rappresentazioni nel mondo esterno (che è notoriamente problematico per i sistemi simbolici). L'importanza delle reti neurali per il *grounding* dei simboli è stato discusso da Steven Harnad in un articolo seminale (Harnad 1990). Da questo punto di vista, il principale vantaggio delle cosiddette reti neurali profonde (*deep neural networks*) e delle *convolutional neural networks* consiste nel fatto che esse risultano ancora più vicine ai dati sensoriali, e quindi necessitano di una fase di preelaborazione dei dati in ingresso minore o nulla (al proposito si veda la recente rassegna in LeCun *et al.* 2015). Tuttavia le rappresentazioni basate su reti neurali sono problematiche da altri punti di vista. Ad esempio, come abbiamo già detto, risulta difficile implementarvi la composizionalità (Fodor e Pylyshyn 1988, Frixione *et al.* 1989). Inoltre, non è chiaro come implementare compiti complessi di pianificazione e ragionamento, che invece possono essere modellati in modo naturale con formalismi simbolici. Di conseguenza, una mossa molto diffusa consiste nell'implementare sistemi ibridi neuro-simbolici. È il caso ad esempio dell'architettura ACT-R (Anderson *et al.* 2004), che impiega un'attivazione subsimbolica di blocchi di informazione simbolica. In casi come questo l'impostazione ibrida permette di superare in una prospettiva cognitivamente motivata certi problemi delle rappresentazioni simboliche e neurali considerate separatamente. In altri casi tuttavia tali approcci ibridi risultano *ad hoc* e non provvedono alcun modello esplicativo del fenomeno studiato. In ogni caso, rimane insoluto il problema ben noto che affligge le reti neurali: la loro opacità. Una rete neurale agisce come una sorta di scatola nera, e fornire un'interpretazione specifica del comportamento delle sue unità e connessioni è tutt'altro che banale. (su questo si veda più oltre il paragrafo 5.2).

3.3 Rappresentazioni analogiche e diagrammatiche

Negli ultimi decenni, sono stati proposti, sia nell'ambito dell'IA che delle scienze cognitive, vari tipi di rappresentazioni che condividono qualche caratteristica con immagini, o, più in generale, con diagrammi. Si consideri ad esempio il dibattito sulle immagini mentali che ha infiammato le scienze cognitive negli anni settanta (Kosslyn *et al.* 2006): secondo i sostenitori delle immagini mentali, alcune rappresentazioni mentali avrebbero il formato di "figure nella mente". Oltre alle immagini mentali in senso stretto, sono state proposti altri tipi di rappresentazioni che hanno caratteristiche in parte "pittoriche". Si consideri la nozione di modello mentale quale è stata proposta da Philip Johnson-Lard (Johnson-Laird 1983, 2006). Secondo Johnson-Laird, molte prestazioni cognitive umane (ad esempio nel campo del ragionamento deduttivo o della semantica del linguaggio naturale) possono essere spiegate ipotizzando l'elaborazione di rappresentazioni analogiche (i modelli mentali, appunto), piuttosto che in termini di manipolazione di rappresentazioni proposizionali, quali regole dichiarative o formule logiche. Secondo Johnson-Laird, ad esempio, quando i soggetti eseguono un compito deduttivo, creano un modello analogico delle premesse che poi ispezionano per derivare una conclusione.

In vari ambiti sono stati proposti modelli pittorici, analogici o diagrammatici, che, in qualche senso, "assomigliano" a ciò che rappresentano (si veda ad esempio Glasgow *et al.* 1995 e, nel caso della pianificazione, Frixione *et al.* 2001). Si tratta di una classe di rappresentazioni eterogenea, che non è certamente maggioritaria se confrontata con le linee predominanti dei sistemi simbolici o con le reti neurali. Inoltre, nonostante la loro attrattività intuitiva, manca una teoria generale di questo tipo di rappresentazioni. Tuttavia esse presentano numerosi vantaggi, soprattutto in domini di tipo spaziale. Al confronto con i modelli subsimbolici e neurali, esse sono molto più trasparenti; se comparate con le rappresentazioni simboliche, esse sono spesso più intuitive, ed evitano in molti casi la necessità di onerose assiomatizzazioni esplicite. Esistono anche varie proposte di incorporare rappresentazioni diagrammatiche nelle architetture cognitive (Kurup e Chandrasekaran 2007).

In conclusione dunque, dal punto di vista empirico, nessuna delle famiglie di rappresentazioni passate in rassegna sembra in grado di rendere conto dell'intero spettro di fenomeni della cognizione umana. Ciò sembra suggerire che anche nei sistemi artificiali sia richiesta una pluralità di approcci rappresentazionali. Tuttavia, il modo in cui queste rappresentazioni potrebbero interagire non sembra chiaro né dal punto di vista empirico, né dal punto di vista della progettazione di sistemi computazionali.

4. Il ruolo degli spazi concettuali

La nostra tesi è che una classe di rappresentazioni di tipo geometrico, gli *spazi concettuali* (Gärdenfors 2000), possa costituire una sorta di lingua comune per consentire l'interazione tra tipi diversi di approcci (questa è una delle tre proposte avanzate nell'ambito della integrazione di livelli rappresentazioni diversi nelle architetture cognitive. Per una analisi dettagliata si veda Lieto, Lebiere & Oltramari 2018).

Per un verso, gli spazi concettuali consentono di superare alcune limitazioni dei sistemi simbolici relative sia alla conoscenza di senso comune, sia al problema del *grounding* (si veda il par. 5.1). Per contro, essi

possono offrire una sorta di "blueprint" per modellare e progettare reti neurali artificiali che risultino meno opache. Inoltre essi mettono a disposizione un livello di interpretazione più astratto per i meccanismi neurali sottostanti (si veda il par. 5.2). Infine, grazie alla loro natura geometrica, possono offrire una cornice unificante per interpretare molti tipi di rappresentazione diagrammatica o analogica (si veda il par. 5.3).

La teoria degli spazi concettuali mette a disposizione un quadro generale per la rappresentazione della conoscenza negli agenti cognitivi che, negli ultimi due decenni, è stato applicato in un ampio spettro di ambiti dell'IA, che vanno dalla percezione visiva (Chella *et al.* 1997) alla robotica (Chella *et al.* 2003), dai sistemi di *question answering* (Lieto *et al.* 2015) alla percezione musicale (Chella 2015; si veda Zenker e Gärdenfors 2015 per una rassegna recente). Secondo Peter Gärdenfors, gli spazi concettuali costituiscono un livello di rappresentazione intermedio tra i livelli subsimbolico e simbolico. La principale caratteristica degli spazi concettuali è l'utilizzo di una impostazione geometrica per la rappresentazione della conoscenza: in sintesi, uno spazio concettuale è uno spazio metrico in cui entità e concetti sono descritti nei termini di un insieme di dimensioni (*quality dimensions*). In alcuni casi tali dimensioni sono direttamente correlate a informazioni percettive, come ad esempio temperatura, peso o luminosità. In altri casi le dimensioni possono essere di natura più astratta. L'idea centrale è che la rappresentazione della conoscenza tragga vantaggio dalla struttura geometrica dello spazio concettuale. Le dimensioni rappresentano proprietà dell'ambiente a prescindere da ogni tipo di descrizione linguistica. In questo senso, uno spazio concettuale precede ogni caratterizzazione simbolica delle informazioni. Un punto in uno spazio concettuale corrisponde a una entità epistemologicamente primitiva al livello di analisi considerato. Per esempio, nel caso della percezione visiva, il valore di un punto è ottenuto a partire da misurazioni effettuate sul mondo esterno per mezzo, poniamo, di una telecamera, attraverso le elaborazioni degli algoritmi di visione di baso livello. I concetti sono rappresentati come regioni in uno spazio. Un aspetto importante della teoria è costituito dal fatto che sugli spazi è definita una funzione metrica. Seguendo Gärdenfors, la distanza tra due punti calcolata in base a tale funzione corrisponde alla misura della somiglianza tra le entità corrispondenti quali esse sono percepite dal soggetto. Un aspetto importante della teoria ha a che fare con il ruolo che svolgono nella concettualizzazione gli insiemi convessi di punti. Secondo Eleanor Rosch (1975), le cosiddette categorie naturali rappresentano il livello più informativo di categorizzazione nelle tassonomie di oggetti ordinari. Esse sono le più differenziate tra loro, e costituiscono il livello preferenziale di riferimento per le espressioni del linguaggio naturale. Inoltre, sono le prime ad essere apprese dai bambini, e quelle che consentono un processo di categorizzazione più veloce. Gärdenfors propone il cosiddetto criterio P, in base al quale le categorie naturali corrisponderebbero a insiemi convessi di punti in uno spazio concettuale. Di conseguenza, nel caso di una categoria naturale, dati due punti che appartengono a un concetto C, tutti i punti che cadono tra loro appartengono a loro volta a C. In tale contesto prototipi ed effetti prototipici hanno una naturale interpretazione geometrica. I prototipi corrispondono al centroide geometrico della regione che rappresenta un concetto, e, dato un concetto, si può assegnare un grado di centralità ad ogni punto della regione corrispondente, che può essere interpretato come una misura di tipicità. Viceversa, dato un insieme di n prototipi rappresentati come punti in uno spazio, si possono determinare i concetti corrispondenti mediante una tessellazione dello spazio in n regioni convesse per mezzo dei cosiddetti diagrammi di Voronoi (Okabe *et al.* 2000). In sintesi, una delle caratteristiche principali degli spazi concettuali è costituita dal fatto che essi consentono di trattare conto in modo naturale gli effetti di tipicità nei concetti. La loro struttura geometrica fornisce un modo naturale di calcolare la similarità semantica tra oggetti sulla base di nozioni metriche o topologiche (ad esempio basate sul cosiddetto *Region Connection Calculus* – Gärdenfors e Williams 2001).

Gärdenfors si è concentrato principalmente sulla rappresentazione della tipicità mediante prototipi. Tuttavia, gli spazi concettuali consentono in modo naturale di rappresentare concetti non classici anche in termini di esemplari (Frixione e Lieto 2013) (come abbiamo detto nel par. 2.1, prototipi ed esemplari sono punti di vista complementari che consentono di spiegare aspetti differenti della tipicità).

4.*1 Spazi Concettuali vs Modelli Distribuzionali*

Sin dalla loro introduzione, negli anni 70, i modelli semantici a spazi vettoriali sono diventati una delle tecniche di rappresentazione e di elaborazione del linguaggio più note ed utilizzate. Trovano, infatti, applicazioni in una moltitudine di compiti: dalla classificazione di documenti, al recupero di informazioni, alla disambiguazione etc. Una delle assunzioni principali degli approcci linguistici basati su modelli vettoriali è l'ipotesi distribuzionale di Zelig Harris, secondo cui le parole che occorrono in un contesto simile hanno significati simili. Tale assunzione è alla base della costruzione di cosiddette risorse vettoriali di tipo distribuzionale, in cui il significato di una parola è rappresentato vettorialmente in uno spazio avente, come dimensioni, le parole che presentano, statisticamente, il maggior numero di co-occorrenze rispetto alla parola da descrivere. Un esempio: in tali risorse il vettore della parola CANE avrà, come dimensioni, GATTO,

CODA, ABBAIARE etc. Vale a dire le parole che, più frequentemente, ne costituiscono il contesto linguistico da un punto di vista distribuzionale.

Nonostante tali risorse vettoriali si siano rivelate utili in diverse applicazioni, tali modelli non hanno una nozione *built-in* di composizionalità (si veda Lenci 2008). A nostro avviso, una delle ragioni di tale problema riguarda il fatto che gli elementi caratterizzanti le dimensioni di tali vettori non sono realmente costituivi del contenuto di tali entità rappresentate. In altre parole: ne danno una caratterizzazione contestuale ma non forniscono una caratterizzazione intrinseca delle componenti lessicali che intendono descrivere.
Una possibilità per superare questo problema riguarda, a nostro avviso, l'integrazione di tali risorse linguistiche distribuzionali con gli spazi concettuali. Una prima proposta, in tal senso, è stata fatta in (Lieto et al 2016) in cui risorse linguistiche distribuzionali (come NASARI, Camacho-Collados et al. 2016) sono state mappate su spazi concettuali passando per risorse intermedie (CONCEPTNET, Havasi et al. 2007.). Tale mapping ha permesso di passare da risorse ad distribuzioni ad altissima dimensionalità a rappresentazioni concettuali a bassa dimensionalità e quindi facilmente trattabili dal punto di vista computazionale per compiti di similarità semantica (per i dettagli si rimanda a Lieto et al. 2016). La procedura di integrazione messa a punto per risorse distribuzionali può valere, in linea di principio, anche per risorse lessico-grammaticali mono o multilingua (di cui Annibale Elia è stato tra i principali sviluppatori in Italia).

5. Sui vantaggi degli spazi concettuali

Nei paragrafi seguenti evidenzieremo alcuni dei vantaggi degli spazi concettuali nell'affrontare i problemi posti dai sistemi di rappresentazione che abbiamo evidenziato in precedenza. Tale analisi confermerà la nostra affermazione secondo cui gli spazi concettuali possono contribuire a superare i limiti di tali sistemi.

5.1 Prototipi e composizionalità

Come abbiamo anticipato, la composizionalità è difficile da riconciliare con gli effetti prototipici. In questo paragrafo sosterremo che gli spazi concettuali possono riconciliare questi due importanti aspetti della rappresentazione dei concetti. Secondo un argomento ben noto (Fodor 1981; Osherson e Smith 1981), i prototipi non sono composizionali. In breve, l'argomento può essere esposto attraverso l'esempio seguente: si consideri un concetto come PET FISH (ossia, il tradizionale pesce rosso). Esso risulta dalla composizione per congiunzione del concetto PET e del concetto FISH. Tuttavia, il prototipo di PET FISH non può essere ottenuto dalla composizione dei prototipi di PET e di FISH: un PET tipico è caldo e peloso, un pesce tipico è grigiastro e vive in mare, ma un PET FISH non è né caldo, né peloso, né grigiastro e non vive in mare.

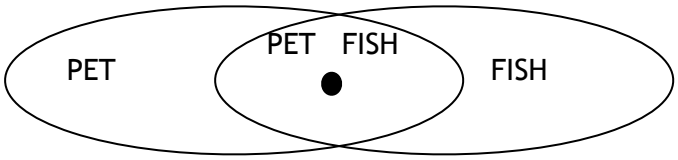

Fig. 1

La situazione risulta più promettente se, invece delle tradizionali rappresentazioni simboliche basate sulla logica, adottiamo una rappresentazione geometrica basata sugli spazi concettuali (va sottolineato, tuttavia, come di recente sia stato proposto un approccio simbolico - la logica Tcl - in grado di gestire tali aspetti. Si veda Lieto e Pozzato, 2019). Come abbiamo già detto, se noi rappresentiamo un concetto come una zona di spazio convessa in uno spazio concettuale appropriato, allora il grado di tipicità di un certo individuo può essere misurato come la distanza del punto corrispondente dal centroide del concetto. La congiunzione di due concetti è rappresentata dall'intersezione delle due porzioni di spazio corrispondenti, come in fig. 1. A questo punto, il prototipo di PET FISH potrà essere identificato con il centroide della zona di spazio che costituisce tale intersezione, anche se tale prototipo risulterà molto periferico sia rispetto al prototipo di PET, sia al prototipo di FISH. Questo tipo di rappresentazione cattura meglio le nostre intuizioni sulla tipicità. In conclusione, il trattamento della composizionalità e quello della tipicità sembrano richiedere forme di rappresentazione diverse, e, di conseguenza, una impostazione di tipo ibrido.

5.2 Interpretare le reti neurali

Abbiamo detto che le reti neurali sono particolarmente adatte a compiti di classificazione, e che, per le loro caratteristiche sono state ampiamente utilizzate nelle architetture cognitive. Tuttavia, uno dei problemi di questa classe di sistemi è costituito dalla loro opacità. Una rete neurale si comporta come una sorta di scatola nera: è estremamente difficile assegnare una interpretazione specifica alle strutture (unità, connessioni, pesi) che la costituiscono. In molti casi ciò può costituire un problema. In un dominio di tipo medico, ad esempio non è sufficiente produrre una classificazione dei sintomi, ma è richiesta anche una spiegazione dettagliata che motivi le risposte fornite. L'opacità di questa classe di rappresentazioni risulta inaccettabile anche nel caso delle architetture cognitive, che aspirano a fornire un modello perspicuo della cognizione umana e che, in quanto tali, dovrebbero consentire non soltanto di prevedere il comportamento di un agente, ma anche di spiegarlo. Questo problema diventa ancora più grave nel caso delle reti profonde (*deep neural networks*), dove, a causa del grande numero di strati nascosti, le unità e i pesi da interpretare sono molto più numerosi.

Una possibile soluzione consiste nel fornire una rappresentazione più astratta, in termini geometrici, delle reti. E' possibile avere una semplice interpretazione geometrica delle operazioni di una rete neurale: le operazioni di ogni strato possono essere descritte come uno strato geometrico funzionale le cui dimensioni sono correlate con le funzioni di trasferimento delle unità del livello stesso. Secondo questa interpretazione, i pesi delle connessioni tra gli strati possono essere descritti in termini di matrici di trasformazione da uno spazio a un altro. Tuttavia, mentre le interpretazioni degli spazi di input e di output dipendono dall'insieme di addestramento e dalla specifica progettazione della rete, l'interpretazione degli spazi che corrispondono agli strati nascosti di solito è difficile. Tuttavia, la letteratura riporta casi sporadici in cui una interpretazione parziale delle operazione delle unità è possibile: un esempio recente è riportato da Zhou (Zhou et al. 2015). Un tentativo più generale di interpretare l'attività di una rete neurale in termini geometrici è dovuto a (Amari e Nagaoka 2007).

Riteniamo che la teoria degli spazi concettuali possa essere considerata una sorta di stile di progettazione in grado di supportare la progettazione di reti neurali più trasparenti, e di facilitare l'interpretazione degli strati nascosti di unità: L'interpretazione di una rete neurale in termini di spazi concettuali Può provvedere un punto di vista più astratto e perspicuo sul comportamento sottostante della rete. Gärdenfors (2000) offer una semplice analisi delle relazioni tra spazi concettuali e reti auto-organizzanti. In seguito, Balkenius (1999) ha proposto una interpretazione più articolata delle reti RBF (*Radial Basis Function*), ampiamente adottate in letteratura, nei termini delle dimensioni di uno spazio concettuale opportune. Secondo questa proposta, una rete costituita da un insieme di unità RBF può essere interpretata come un semplice spazio concettuale descritto da un insieme di *quality dimensions* integrate. Di conseguenza, una rete neurale costituita da un insieme di insiemi di unità RBF può essere interpretata geometricamente come uno spazio concettuale formato da insiemi di dimensioni integrate.

Inoltre, in base alla proposta di Shimon Edelman (1995), le unità di una rete RBF possono essere lette come prototipi in uno spazio concettuale opportuno. Tale interpretazione permette di musurare la similarità tra l'input della rete e i prototipi corrispondenti alle unità. Ciò risulterebbe molto più problematico prendendo in considerazione la sola rete neurale, in quanto l'informazione in essa sarebbe nascosta e implicita. In aggiunta, diventa possibile trattare il caso di entità "chimeriche", che siano più o meno equidistanti tra due o più prototipi. Ad esempio, una sirena è una donna con ali e zampe di uccello, e risulterà pertanto equidistante tra il prototipo di uccello e quello di essere umano (di sesso femminile) (vedi Edelman 1995). Da questo punto di vista, la possibilità di conciliare composizionalità e tratti prototipici sembra una caratteristica cruciale degli spazi concettuali, in grado di arricchire sia le rappresentazioni simboliche che quelle subsimboliche.

Infine, un lavoro che va nella direzione di una interpretazione più astratta delle rappresentazioni neurali è stato ottenuto nell'ambito della *Semantic Pointers Perspective* adottata dal NEF (*Neural Engineering framework*) (Eliasmith e Anderson 2004); esso rappresenta il nucleo dell'architettura di ispirazione biologica SPAUN (Eliasmith *et al.* 2012). Le rappresentazioni neurali vengono interpretate come vettori ottenuti attraverso differenti tipi di operazioni (ad esempio compressione e *binding* ricorsivo per mezzo di convoluzioni circolari, si veda Crawford *et al.* 2015). Tale prospettiva è completamente compatibile con la nostra proposta di provvedere una interpretazione più astratta dei meccanismi delle neurali attraverso spazi concettuali multidimensionali.

5.3 Rappresentazioni analogiche, diagrammatiche e spazi concettuali

Le rappresentazioni analogiche e diagrammatiche consentono di trattare in modo efficiente e intuitivo tipi di informazione che comporterebbero rappresentazioni molto complesse e scomode se trattate esplicitamente per mezzo di formalismi dichiarativi di tipo simbolico. Consideriamo un semplice esempio che è stato discusso da Philip Johnson-Laird (1983). La relazione *essere alla destra di* usualmente è transitiva:

se A è a destra di B e B è a destra di C, allora A è a destra di C. Ma si consideri un caso in cui A, B e C sono collocati, ad esempio, attorno a un tavolino rotondo. In questo caso può succedere che C sia alla destra di B, B sia alla destra di A, ma C non sia alla destra di A: piuttosto è A ad essere alla destra di C (Fig. 2). Per rendere conto di questo semplice fatto per mezzo di regole o di un'assiomatizzazione simbolica esplicita occorrerebbe un gran numero di asserzioni complesse e dettagliate. Viceversa, l'adozione di qualche forma di rappresentazione analogica del tipo di un modello mentale, associata ad opportune procedure per la generazione, la revisione e l'ispezione dei modelli, potrebbe consentire di affrontare il problema in maniera più diretta e naturale.

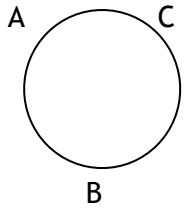

Figura 2

Come abbiamo già detto nel par. 3.3, molti tipi di modelli diagrammatici sono stati proposti, senza che si giungesse tuttavia a un quadro teorico unitario. Gli spazi concettuali, grazie alla loro natura geometrica, consentono di rappresentare questo tipo di informazioni e offrono, al tempo stesso, un inquadramento teorico solido e ben compreso, che potrebbe consentire di unificare molte delle rappresentazioni diagrammatiche esistenti sul mercato.

La natura geometrica degli spazi concettuali potrebbe essere utile anche per rappresentare problemi e fenomeni più astratti, di natura non specificamente spaziale. Un problema tipico dei sistemi di rappresentazione sia simbolici che neurali concerne la capacità di tenere traccia dell'identità di entità individuali nel corso del tempo. Nel tempo le proprietà di un oggetto cambiano. A quali condizioni possiamo re-identificare una data entità come la stessa, a prescindere da tali mutamenti? In molti casi la risposta non è facile. Gli spazi concettuali suggeriscono un modo di affrontare il problema. Abbiamo detto che in uno spazio concettuale gli individui sono rappresentati da punti. Tuttavia, in una prospettiva dinamica, gli oggetti possono piuttosto essere visti come traiettorie in uno spazio concettuale che disponga di un asse temporale. Un entità può spostarsi, può invecchiare, può cambiare forma o colore, e così via. Man mano che le proprietà di un oggetto si modificano, il punto che lo rappresenta si muoverà nello spazio concettuale seguendo una certa traiettoria. Poiché di solito questi cambiamenti sono graduali e non repentini, su tale traiettoria possono essere fatte numerose assunzioni, (continuità, uniformità, rispetto delle leggi fisiche) (Chella *et al.* 2004). Prevedere l'evoluzione di un oggetto (la sua posizione futura, o il modo in cui cambieranno le sue proprietà) può essere visto come l'estrapolazione di una traiettoria in uno spazio concettuale. Re-identificare un oggetto che per un certo intervallo di tempo non è stato percepibile equivale a interpolare la sua traiettoria passata e quella presente. In generale, Ciò può offrire una potente euristica per tener traccia dell'identità di un oggetto individuale. Anche in questo caso, aspetti cruciali delle rappresentazioni diagrammatiche possono trovare negli spazi concettuali una interpretazione generale e unificante.

6. Conclusioni

Abbiamo proposto gli spazi concettuali come una sorta di lingua franca che consenta di unificare e integrare a partire da una base comune le rappresentazioni di tipo simbolico, subsimbolico e diagrammatico, in modo da superare i noti problemi che affliggono ciascuna di queste impostazioni presa singolarmente. Abbiamo, inoltre, illustrato come le rappresentazioni a spazi concettuali possano essere integrate con risorse linguistiche distribuzionali, oggi molto utilizzate nell'ambito dell'elaborazione automatica del linguaggio naturale (NLP). Tale integrazione può valere, in linea di principio, anche per risorse lessico-grammaticali mono o multilingua (di cui Annibale Elia è stato tra i principali sviluppatori).

A partire dagli argomenti proposti da Gärdenfors (1997) per giustificare la necessità di un livello di rappresentazione intermedio tra il simbolico e il subsimbolico, abbiamo suggerito che gli spazi concettuali possano contribuire a risolvere il noto conflitto tra composizionalità e tipicità nella rappresentazione concettuale, possano consentire una interpretazione più trasparente per i sistemi di tipi neurali, e fornire una

sorta di “blueprint” per la progettazione di tali reti. Inoltre gli spazi concettuali possono fornire un quadro unificante per interpretare molti tipi di rappresentazione analogica e diagrammatica.

Riferimenti bibliografici

Shun-ichi Amari e Hiroshi Nagaoka. *Methods of Information Geometry*, volume 191. American Mathematical Soc., 2007.

John R. Anderson, Daniel Bothell, Michael D Byrne, Scott Douglass, Christian Lebiere, e Yulin Qin. An integrated theory of the mind. *Psychological review*, 111(4):1036, 2004.

Christian Balkenius. Are there dimensions in the brain? *Spinning Ideas, Electronic Essays Dedicated to Peterì Gärdenfors on His Fiftieth Birthday*, 1999.

José Camacho-Collados, Mohammad Taher Pilehvar e Roberto Navigli. "Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities." *Artificial Intelligence*, 240, pp. 36-64, 2016.

Antonio Chella. A cognitive architecture for music perception exploiting conceptual spaces. In *Applications of Conceptual Spaces*, pp. 187–203. Springer, 2015.

Antonio Chella, Silvia Coradeschi, Marcello Frixione, and Alessandro Saffiotti. Perceptual anchoring via conceptual spaces. In *Proceedings of the AAAI-04 Workshop on Anchoring Symbols to Sensor Data*, pp. 40–45, 2004.

Antonio Chella, Marcello Frixione, e Salvatore Gaglio. A cognitive architecture for artificial vision. *Artificial Intelligence*, 89(1):73–111, 1997.

Antonio Chella, Marcello Frixione, e Salvatore Gaglio. Anchoring symbols to conceptual spaces: the case of dynamic scenarios. *Robotics and Autonomous Systems*, 43(2):175–188, 2003.

Eric Crawford, Matthew Gingerich, e Chris Eliasmith. Biologically plausible, human-scale knowledge representation. *Cognitive science*, 2015.

Artur S. D’Avila Garcez, Luis C. Lamb, e Dov M. Gabbay. *Neural-symbolic cognitive reasoning*. Springer Science & Business Media, 2008

Ernest Davis e Gary Marcus. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, 58(9):92–103, 2015.

Shimon Edelman. Representation, similarity, and the chorus of prototypes. *Minds and Machines*, 5(1):45–68, 1995.

Annibale, Elia. "Dizionari elettronici e applicazioni informatiche." JADT. 1995.

Chris Eliasmith e Charles H Anderson. *Neural Engineering: Computation, Representation, and Dynamics in neurobiological systems*. MIT press, 2004.

Chris Eliasmith, Terrence C Stewart, Xuan Choo, Trevor Bekolay, Travis DeWolf, Yichuan Tang, and Daniel Rasmussen. A large-scale model of the functioning brain. *Science*, 338(6111):1202–1205, 2012.

Jonathan St BT Evans e Keith Ed Frankish. *In two minds: Dual processes and beyond*. Oxford UniversityPress, 2009.

Jerry A Fodor. The present status of the innateness controversy. In Jerry A Fodor, a cura di, *Representations: Philosophical Essays on the Foundations of Cognitive Science*, chapter 10, pp. 257 – 316. MIT Press, Cambridge, MA, 1981.

Jerry A Fodor e Zenon W Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988.

Frixione, Marcello, and Antonio Lieto. "Representing and reasoning on typicality in formal ontologies." Proceedings of the 7th International Conference on Semantic Systems. ACM, 2011.

Marcello Frixione e Antonio Lieto. “Exemplars, prototypes and conceptual spaces." Biologically Inspired Cognitive Architectures 2012. Springer, Berlin, Heidelberg, 131-136, 2013.

Marcello Frixione, Giuseppe Spinelli, e Salvatore Gaglio. Symbols and subsymbols for representing knowledge: a catalogue raisonné. In *Proceedings of the 11th international joint conference on Artificial intelligence* - Volume 1, pp. 3–7. Morgan Kaufmann Publishers Inc., 1989.

Marcello Frixione, Gianni Vercelli, e Renato Zaccaria. Diagrammatic reasoning for planning and intelligent control. *Control Systems, IEEE*, 21(2):34–53, 2001.

Peter Gärdenfors. Symbolic, conceptual and subconceptual representations. In *Human and Machine Perception*, pp. 255–270. Springer, 1997.

Peter Gärdenfors. *Conceptual spaces: The geometry of thought*. MIT press, 2000.

Peter Gärdenfors e Mary-Anne Williams. Reasoning about categories in conceptual spaces. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence* - Volume 1, IJCAI'01, pp. 385–392,San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

Janice Glasgow, N. Hari Narayanan, e B. Chandrasekaran. *Diagrammatic reasoning: Cognitive and computational perspectives*. Mit Press, 1995.

Thomas L. Griffiths, Charles Kemp, e Joshua B. Tenenbaum. *Bayesian models of cognition*. Cambridge University Press, 2008.

Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335–346, 1990.

Catherine Havasi, Robert Speer e Jason Alonso. "ConceptNet 3: a flexible, multilingual semantic network for common sense knowledge." Recent advances in natural language processing. Philadelphia, PA: John Benjamins, 2007.

Philip N. Johnson-Laird. Mental models: Towards a cognitive science of language, inference, and consciousness. Harvard University Press, 1983.

Philip N. Johnson-Laird. *How we reason*. Oxford University Press, USA, 2006.

Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

Stephen M. Kosslyn, William L. Thompson, e Giorgio Ganis. *The case for mental imagery*. Oxford University Press, 2006.

Unmesh Kurup e B. Chandrasekaran. Modeling memories of large-scale space using a bimodal cognitivearchitecture. In *Proceedings of the eighth international conference on cognitive modeling*, pp. 267–272. Citeseer, 2007.

John Laird. *The Soar cognitive architecture*. MIT Press, 2012.

Yann LeCun, Yoshua Bengio, e Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

Alessandro Lenci. Distributional semantics in linguistic and cognitive research. Italian journal of linguistics, 20 (1), pp. 1-31, 2008.

Lieto, A., Radicioni, D. P., & Rho, V. (2015, June). A common-sense conceptual categorization system integrating heterogeneous proxytypes and the dual process of reasoning. In Twenty-Fourth International Joint Conference on Artificial Intelligence.

Lieto, Antonio, Mensa, Enrico, Radicioni, Daniele P. "A resource-driven approach for anchoring linguistic resources to conceptual spaces." Conference of the Italian Association for Artificial Intelligence. Springer, Cham, 2016.

Lieto, Antonio, Chella Antonio, Frixione Marcello. "Conceptual spaces for cognitive architectures: A lingua franca for different levels of representation". *Biologically Inspired Cognitive Architectures*, 19, pp. 1-9, 2017

Lieto, A., Radicioni, D. P., & Rho, V. (2017). Dual PECCS: a cognitive system for conceptual representation and categorization. Journal of Experimental & Theoretical Artificial Intelligence, 29(2), 433-452.

Lieto, A., Lebiere, C., & Oltramari, A. (2018). The knowledge level in cognitive architectures: Current limitations and possible developments. Cognitive Systems Research, 48, 39-55.

Lieto, A., & Pozzato, G. L. (2018). A description logic framework for commonsense conceptual combination integrating typicality, probabilities and cognitive heuristics. arXiv preprint arXiv:1811.02366.

Edouard Machery. *Doing without concepts*. OUP, 2009.

Barbara C Malt. An on-line investigation of prototype and exemplar strategies in classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(4):539, 1989.

Gregory Leo Murphy. *The big book of concepts*. MIT press, 2002.

Allen Newell e Herbert A Simon. Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, 19(3):113–126, 1976.

Thomas Dyhre Nielsen e Finn Verner Jensen. *Bayesian networks and decision graphs*. Springer Science & Business Media, 2009.

Atsuyuki Okabe, Barry Boots, Kokichi Sugihara, e Sung Nok Chiu. *Spatial Tessellations - Concepts and Applications of Voronoi Diagrams*, Second Edition. John Wiley & Sons, Chichester, 2000.

Randall C O'Reilly e Yuko Munakata. *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. MIT press, 2000.

Daniel N Osherson e Edward E Smith. On the adequacy of prototype theory as a theory of concepts. *Cognition*, 9(1):35–58, 1981.

Gualtiero Piccinini. Two kinds of concept: Implicit and explicit. *Dialogue*, 50(01):179–193, 2011.

Eleanor Rosch. Cognitive representations of semantic categories. *J. Exp. Psychol. Gen.*, 104(3):192–233, 1975.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al.

Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587): 484–489, 2016.

Aaron Sloman. How can we reduce the gulf between artificial and natural intelligence? In *Proceegings of AIC 2014*, 2nd International Workshop on Artificial Intelligence and Cognition, volume 1315, pp. 1–13, 2014.

Larry R. Squire and Barbara J. Knowlton. Learning about categories in the absence of memory. *Proceedings of the National Academy of Sciences*, 92(26):12470–12474,1995.

Keith E Stanovich e Richard F West. Advancing the rationality debate. *Behavioral and Brain Sciences*, 23(05):701–717, 2000.

Ron Sun. The CLARION cognitive architecture: Extending cognitive modeling to social simulation. *Cognition and multi-agent interaction*, pp. 79–99, 2006.

Ludwig Wittgenstein. *Philosophische Untersuchungen*. Oxford, Blackwell, 1953.

Frank Zenker e Peter Gärdenfors. *Applications of conceptual spaces - The case for geometric knowledge representation.* Springer, 2015.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, e Antonio Torralba. Object detectors emerge in deep scene cnns. arXiv preprint arXiv:1412.6856, 2015.