Peer Review Journal
Rivista Italiana di Filosofia Analitica junior

Rivista Italiana di Filosofia Analitica Junior 4:1 (2013)
ISSN 2037-4445 © http://www.rifanalitica.it
Patrocinata dalla *Società Italiana di Filosofia Analitica*

UNA TEORIA DELLA RAZIONALITÀ: IL MODELLO BDI

Costanza Larese

ABSTRACT. In quest'articolo propongo un'analisi di una teoria della razionalità, il modello *Belief-Desire-Intention* (BDI), con l'obiettivo di stabilirne la fecondità teoretica. Interpreto il modello come il risultato dell'indebolimento di alcuni principi cardine della teoria della scelta razionale: se questa è di natura normativa e considera agenti altamente idealizzati, il modello BDI è invece motivato dallo scopo di dare una caratterizzazione cognitivamente plausibile delle azioni degli individui e inserisce nella definizione di razionalità aspetti non normativi. Per questa ragione, la teoria BDI introduce il concetto di intenzione e complica la propria ontologia: le intenzioni pongono dei vincoli di consistenza sulla componente motivazionale dell'individuo e fungono da filtro di ammissibilità sulla selezione di altre intenzioni (Bratman, 1987). Presento ed analizzo di seguito due formalizzazioni, sviluppatesi in due diverse aree di ricerca (logica e intelligenza artificiale), dei principi filosofici della teoria: il sistema BDICTL*-W3 (Georgeff e Rao, 1998) ed un esempio di *Agent Control Loop* (Wooldridge, 2000). La discussione vuole rilevare le peculiarità dei vari approcci alla teoria in oggetto, individuare i nodi concettuali comuni ma anche le specificità di ciascun apporto. Concludo quindi con alcune osservazioni di carattere epistemologico sui vantaggi di un approccio plurale.

KEYWORDS. Razionalità, Modello BDI, Intenzione, Plausibilità cognitiva, Normatività.

AUTORE. Costanza Larese. costanza.larese@sns.it.
RICEVUTO. 6 marzo 2013. **ACCETTATO**. 23 aprile 2013.

Il modello BDI è una teoria della razionalità che studia il ragionamento di individui razionali. Prima di cominciare l'analisi del modello, propongo subito due esempi del tipo di ragionamento che intendo discutere:

Esempio 1 (Bratman, 1990, p. 23) *Il terrorista e l'attentatore strategico hanno entrambi lo scopo di promuovere azioni militari per danneggiare il nemico. Entrambi intendono realizzare il proposito tramite bombardamenti. Il piano del terrorista prevede di bombardare la scuola nel territorio nemico, uccidendo i bambini, terrorizzando la popolazione, costringendo così il nemico alla capitolazione. Il piano dell'attentatore strategico prevede invece di colpire il magazzino delle munizioni nemiche, minando gli sforzi bellici dell'avversario. Tuttavia, l'attentatore strategico sa anche che accanto al magazzino di munizioni vi è una scuola; pur essendosi preoccupato dell'effetto disumano dell'azione, l'attentatore ritiene che il guadagno per l'esito della guerra dato dalla distruzione delle munizioni nemiche sia superiore al costo dell'operazione.*

Esempio 2 (Georgeff e Rao, 1998, p. 298) *Phil occupa un seggio alla Camera dei Rappresentanti e, in vista delle prossime elezioni, crede di avere le seguenti possibilità: candidarsi nuovamente per il seggio alla Camera, candidarsi per un posto al Senato oppure ritirarsi dalla politica. Non considera seriamente l'opzione di ritirarsi dalla vita politica, mentre è certo di poter mantenere il seggio alla Camera. Phil deve decidere se indire o meno un sondaggio, il cui esito sarà il consenso oppure il dissenso della maggioranza riguardo al suo passaggio al Senato. Sulla base del risultato del sondaggio, Phil deciderà se candidarsi alla Camera oppure al Senato.*

Nel primo esempio, i piani elaborati da entrambi gli agenti prevedono come effetto la strage dei bambini: tuttavia, mentre il terrorista intende colpire la scuola per danneggiare il nemico, l'attentatore strategico non intende uccidere i bambini, ma giudica la conseguenza del suo piano un mero effetto collaterale. Si vedrà come il potere espressivo del modello BDI permetta di rappresentare la distinzione tra gli effetti intesi e gli effetti collaterali presenti nello scenario possibile selezionato dall'agente. Poichè intuitivamente si può pensare ad una identificazione di terrorismo e irrazionalità, è opportuno notare subito che il modello BDI, come la teoria classica della scelta razionale, assume il Principio di Neutralità: la razionalità nella teoria è indipendente dal contenuto delle intenzioni.

Il secondo esempio analizza il ragionamento di un agente in condizione epistemica di incertezza: nel contesto accadono degli eventi su cui l'agente non ha un controllo diretto e a cui può solo assegnare una probabilità soggettiva che esprime il suo grado di convinzione del loro eventuale verificarsi. Phil non sa se la maggioranza approverà o meno il suo passaggio al Senato e neppure conosce l'esito delle elezioni.

1 Due modelli di razionalità

1.1 Teoria della scelta razionale: idealizzazione e normatività

Analizzo ora i principi cardine della teoria classica della scelta razionale: mi riferirò alle decisioni individuali in condizioni di certezza poichè i modelli di decisione razionale per problemi più complessi non fanno che estendere tali fondamenti.

Fissati un insieme A di alternative reali e un insieme E di esiti, la teoria considera il comportamento di scelta di un agente dotato di preferenze personali e capace di codificare

informazioni. Dato l'insieme non vuoto di alternative reali A, una *funzione di scelta* S ne restituisce un sottoinsieme non vuoto $S(A)$, detto l'insieme scelto: $\emptyset \neq S(A) \subseteq A$. Le *preferenze* di un agente i sono espresse da una relazione binaria sull'insieme non vuoto degli esiti E, $\succ_i \subseteq E^2$: posti $e, f \in E$, si scrive $e \succ_i f$ per indicare che l'agente i preferisce l'esito e a quello f. In condizioni di certezza, ogni alternativa reale $a \in A$ determina in modo univoco un esito $e \in E$. L'idea fondamentale della teoria classica della scelta razionale è data dal Principio di Massimizzazione, secondo cui un agente è razionale se e solo se si comporta in modo massimizzante rispetto agli esiti in E. Questo principio è normativo, perchè da un lato stabilisce una norma di decisione a cui gli agenti devono uniformarsi, dall'altro fissa un criterio con cui determina quali sono i comportamenti irrazionali.

A partire da una relazione di preferenza che rispetti determinati vincoli, si può definire una funzione di scelta che soddisfi il Principio di Massimizzazione. Il modello impone che $\succ_i$ sia una relazione d'ordine, cioè che rispetti i vincoli di asimmetria, completezza e transitività e di conseguenza esclude tutti quegli individui che, violando tali assiomi, sono detti irrazionali. In questo modo la teoria caratterizza come irrazionali, e dunque esclude, la gran parte degli individui concreti: questi infatti esibiscono spesso preferenze incomplete oppure preferenze acicliche ma non transitive. Si consideri un semplice esempio: Phil deve scegliere una bustina di tè fra tre possibilità, Earl Grey (G), tè verde (V) e tè Jasmine (J). L'unico criterio rilevante ai fini della sua scelta è costituito dall'aroma di ciascun infuso. Phil non preferisce V a J nè J a V: del resto, il tè Jasmine è preparato con tè verdi, per cui Phil non apprezza la differenza tra i due aromi. Tuttavia, Phil preferisce J a G e G a V. Gli attori della teoria classica della scelta razionale sono perciò agenti ideali.

Si chiama elemento ottimale rispetto all'ordine di preferenza l'esito $opt_E = \{e^* \in E | e^* \succ e, \forall e \in E\}$, cioè quell'esito preferito a tutti gli altri. Si dimostra che $S(E) = opt_E$ è una funzione di scelta che, come conseguenza immediata, soddisfa il Principio di Massimizzazione. Si noti come la consistenza delle preferenze sia una condizione essenziale per la soddisfazione del Principio di Massimizzazione: imporre come norma di decisione la selezione dell'esito ottimale richiede che la relazione di preferenza sugli esiti sia consistente. Il Teorema Fondamentale dell'Utilità stabilisce che se $\succ$ è un ordine, allora esiste una funzione $u : E \longrightarrow \mathbb{R}$ tale che $e \succ f \Leftrightarrow u(e) > u(f)$: in altre parole, i vincoli di consistenza imposti sulle preferenze sono sufficienti a garantirne una rappresentazione numerica, che si interpreta come utilità. Questo Teorema permette di catturare formalmente la definizione di decisione razionale: un individuo decide razionalmente se e solo se massimizza la propria funzione di utilità individuale, cioè sceglie quell'alternativa reale a^* che determina l'esito ottimale e^*, cioè l'argomento massimo della funzione di utilità.

In conclusione, la teoria assume che le preferenze dell'agente siano consistenti ed impone che la scelta sia massimizzante rispetto agli esiti: con la prima assunzione, il modello stabilisce di considerare soltanto *agenti ideali*; con la seconda, essa determina il proprio *status normativo*. I concetti di preferenza consistente e di scelta come massimizzazione sono i principi fondamentali di questo modello di razionalità: un agente che violi questi vincoli è caratterizzato come irrazionale ed è escluso dalla teoria. Come si è visto, la consistenza delle preferenze è una condizione necessaria ma non sufficiente affinchè il Principio di Massimizzazione possa essere soddisfatto: analogamente, l'idealizzazione dell'agente è una condizione necessaria ma non sufficiente per la normatività dell'approccio epistemologico del modello.

1.2 Per una teoria dell'azione cognitivamente plausibile

Il modello BDI è una teoria dell'azione razionale che considera *agenti cognitivamente plausibili* a differenza di quelli completamente idealizzati della teoria classica della scelta razionale. Proprio perché l'agente non è completamente idealizzato, l'approccio epistemologico del modello non può essere rigorosamente normativo: insieme ad una riduzione dell'idealizzazione, BDI inserisce nella definizione di razionalità alcuni principi di natura *descrittiva*. Abbassando il grado di idealizzazione dell'agente e inserendo caratterizzazioni descrittive della razionalità, BDI complica l'ontologia della teoria della scelta razionale creando un'ontologia dotata di maggior potere espressivo: nello specifico, analizza ulteriormente la componente motivazionale dei suoi attori in cui introduce il concetto di intenzione. La componente informativa di un individuo resta invece sostanzialmente invariata: come l'agente della teoria della scelta esprime le proprie informazioni sul contesto tramite un'assegnazione di probabilità sugli esiti, così l'agente BDI esprime le proprie convinzioni sul mondo.

Nella teoria della scelta razionale, il processo di idealizzazione dell'agente consiste nell'imposizione di determinati vincoli sulla relazione di preferenza. Viceversa, la plausibilità cognitiva del modello BDI permette di classificare come razionali agenti con *desideri inconsistenti*. Proprio perchè l'insieme delle alternative reali A è interpretato come un insieme di azioni A e un'azione richiede per il suo compimento una determinata quantità di risorse (tempo, informazioni, denaro, energie, memoria, benzina, ...), BDI assume il principio di natura descrittiva per cui gli agenti hanno *limitazioni di risorse*.

Sulla base dell'osservazione secondo cui gli agenti hanno il bisogno di *coordinare azioni* presenti e future tra di esse e queste con le attività degli altri individui, il modello BDI assume che gli agenti elaborino dei *piani*. I piani collegano le diverse intenzioni dell'agente e strutturano le relazioni tra di esse: «*Plans are intentions writ large*» (Bratman, 1987, p. 8). Il fatto che l'agente non abbia risorse infinite determina per Bratman due conseguenze sull'architettura dei piani: questi sono parziali e hanno una struttura gerarchica, nel senso che i piani che riguardano i fini contengono quelli sui mezzi per raggiungerli e gli obiettivi più generali vincolano quelli più specifici. Da un lato infatti piani molto dettagliati potrebbero risultare inutili nel momento in cui si dovessero verificare dei cambiamenti nel contesto e sarebbero difficili da formulare data la condizione di incertezza epistemica propria di un agente limitato; dall'altro lato, la gerarchizzazione dei piani permette di considerare intenzioni più specifiche tenendo ferme quelle più generali e quindi di instaurare rapporti di priorità tra i vari obiettivi. Nonostante la necessità di coordinazione inter e intrapersonale sia un'esigenza propria di attori cognitivamente plausibili, l'approccio epistemologico del modello non è descrittivo, perché assume che gli agenti strutturino dei piani ed impone normativamente dei vincoli di consistenza su questi. I piani di un agente razionale devono essere consistenti al loro interno: l'attentatore che intende indebolire l'avversario non può avere l'intenzione di sganciare una bomba e contemporaneamente non voler far uso di armi di distruzione di massa; inoltre, i piani devono essere consistenti rispetto alle convinzioni: l'attentatore non può avere l'intenzione di colpire il magazzino di munizioni nemiche pur essendo convinto che l'avversario non abbia un magazzino di munizioni; infine, i piani non devono rivelare un'incoerenza tra mezzi e fini: in altre parole, le intenzioni devono poter essere concretizzate attraverso la formulazione di piani riguardo i mezzi impiegabili.

Per poter esprimere sia l'inconsistenza dei desideri (aspetto descrittivo), che la consistenza dei piani (aspetto normativo), il modello BDI articola l'analisi della componente motivazionale dell'agente, complicando l'ontologia della teoria della scelta razionale, in cui le preferenze

consistenti esprimono le motivazioni dell'individuo. L'ontologia BDI assume quindi che un agente razionale sia dotato non soltanto di *desideri* e *convinzioni*, ma anche di *intenzioni*. L'introduzione dell'intenzione nell'analisi della razionalità migliora il potere espressivo del modello con cui diventa possibile considerare agenti cognitivamente plausibili. In ciò che segue, si mostrerà come le intenzioni di un agente razionale svolgano un ruolo centrale nella produzione di azioni e come la componente intenzionale di un individuo sia strutturalmente connessa a desideri e convinzioni.

2 Il modello BDI: relazioni tra convinzioni, desideri, intenzioni

***Intention is Choice with Commitment*: intenzioni, desideri, scelte.** Sia le intenzioni che i desideri esprimono le motivazioni di un individuo; tuttavia se le prime devono essere internamente consistenti, i secondi possono essere, e spesso lo sono, inconsistenti. Il modello impone normativamente che le intenzioni siano un sottoinsieme consistente dei desideri dell'agente.

Bratman (1990) nota che i desideri influenzano soltanto *potenzialmente* l'azione, al contrario dell'intenzione che invece controlla effettivamente la condotta successiva, interrompendo l'agente nel soppesare i pro e contro di un'opzione. Per esempio, il desiderio dell'attentatore di un cioccolatino durante le manovre militari interessa soltanto potenzialmente la sua condotta: sebbene questi possa adottare come intenzione l'acquisto di un cioccolatino, verosimilmente egli non agirà in conseguenza di questo suo desiderio, stabilendo come intenzioni delle motivazioni più rilevanti per orientare l'azione. Questa considerazione rafforza l'idea del ruolo fondamentale dell'intenzione nella produzione di azioni, tanto da indurre autori come Georgeff e Rao (1998) ad affermare che l'intenzione rappresenta lo stato *deliberativo* e non semplicemente quello motivazionale di un sistema.

Da queste due osservazioni, segue che le intenzioni sono il risultato di una *scelta* dell'agente nel dominio dei suoi desideri. Il modello impone normativamente che la scelta dell'agente restituisca un insieme consistente di desideri e che l'agente *si impegni* a realizzarli. A partire da questa analisi, Cohen e Levesque (1990, p. 220) stabiliscono un concetto di intenzione che non può prescindere dall'interazione di questo con scelte, desideri ed impegno:

> *Intention is choice with commitment*. Intention will be modeled as a composite concept specifying what the agent has chosen and how the agent is committed to that choice.

E' però importante notare che un individuo non intende tutto ciò che sceglie. Si consideri quindi un agente che ha scelto di realizzare un desiderio e, così facendo, ha scelto di raggiungere un certo stato di cose: se questi è convinto che le proprie azioni determinino certi effetti, selezionando quell'intenzione, egli ha scelto anche le conseguenze dei propri gesti. L'agente sceglie la globalità di uno tra gli scenari possibili. Tuttavia, egli non intende qualunque cosa in tale scenario, specie effetti indesiderati (ma previsti) oppure mezzi consapevolmente poco graditi. Nel primo esempio, se il terrorista intende colpire la scuola per raggiungere la vittoria, l'attentatore strategico non intende uccidere i bambini, ma giudica tale conseguenza un mero effetto collaterale di un piano più ampio. Come suggeriscono Cohen e Levesque (1990, p. 219), «*Expected side-effects are chosen, but not intended*». Bratman analizza il problema della distinzione tra effetti collaterali ed effetti intesi, chiamato *The Problem of the Package*

Deal, affermando che effetti collaterali e intenzioni non hanno lo stesso ruolo nella pianificazione delle azioni di un agente: ad esempio, se l'individuo non dovesse realizzare gli effetti indesiderati presenti nello scenario che ha scelto, non tenterà di eseguire un nuovo piano per concretizzarli.

***The Asymmetry Thesis*: intenzioni e convinzioni.** Dal momento che l'agente BDI è collocato in un contesto e gli scopi che si propone sono da questo dipendenti, è essenziale che l'individuo abbia delle informazioni sullo stato del mondo. In altre parole, il sistema è situato in un ambiente da cui acquisisce dei dati (input) per produrre azioni (output) che si ripercuotono sull'ambiente stesso. Data la caratteristica limitazione di risorse di un agente cognitivamente plausibile, le sue informazioni riguardo all'ambiente non saranno certezze, ma saranno soltanto convinzioni parziali, verosimili e orientative. Un aspetto fondativo del modello BDI, in cui si è visto che le intenzioni fungono da filtro di ammissibilità per la selezione di altre intenzioni, sarà quindi la relazione tra convinzioni ed intenzioni. L'analisi di Bratman (1987, p. 37), parte da questa idea centrale:

> There is a defeasible demand that one's intentions be consistent with one's beliefs. Violation of this demand is, other things equal, a form of criticizable irrationality.

Questo principio, eventualmente contraddetto da dati empirici sfavorevoli (*defeasible*), stabilisce normativamente un vincolo di consistenza tra intenzioni e convinzioni: esso fissa una regola a cui gli agenti devono uniformarsi per non essere caratterizzati come irrazionali ed esclusi dalla teoria. Tuttavia questa norma non è sufficiente a garantire che l'intenzione di un agente di a implichi la sua convinzione che a. Ad esempio, Phil intende fermarsi in libreria tornando a casa, ma, sapendo che mentre guida è spesso sovrappensiero, teme si dimenticherà di realizzare la sua intenzione e perciò non crede che si fermerà. Tuttavia, casi del genere non provano neppure che l'intenzione di compiere l'azione a non richieda la convinzione che a sia vera. Per questo motivo, Bratman non assume né che l'intenzione di compiere l'azione a implichi la convinzione che a sia vera, né la negazione di questa proposizione. Ciò che invece presenta è un'analisi più accurata della questione, l'*Asymmetry Thesis*, che consta di due argomenti. Con il primo Bratman (1987, p. 38) *ammette l'incompletezza di intenzioni e convinzioni*, con il secondo *respinge l'inconsistenza tra intenzioni e convinzioni*:

> [1.] An intention to a normally provides the agent with support for a belief that he will a. But there need be no irrationality in intending to a and yet still not believing one will.
> [2.] In contrast, there will normally be irrationality in intending to a and believing one will not a; for there is a defeasible demand that one's intentions be consistent with one's beliefs.

Supponiamo che l'attentatore intenda indebolire il nemico, ma che non creda all'efficacia dei suoi tentativi, perché l'avversario è molto potente: non è certo di fallire, ma neppure crede nella riuscita del suo intento. In questo caso, il terrorista intende nuocere al suo antagonista, ma non crede che riuscirà ad indebolirlo. Tuttavia l'attentatore sarebbe irrazionale se, data la sua intenzione, fosse convinto di non indebolire l'avversario: intendere qualcosa che si crede impossibile impedisce all'intenzione di svolgere la sua funzione di filtro di ammissibilità nei confronti delle selezioni successive. Il terrorista convinto dell'impossibilità della propria intenzione non procederà neppure al ragionamento mezzi-fini per trovare una strategia adatta ad indebolire l'avversario, esibendo un comportamento irrazionale. Con l'*Asymmetry Thesis*,

Bratman stabilisce che se un agente intende compiere a, allora: 1) l'agente crede che realizzare a sia possibile, 2) l'agente non crede che non riuscirà a realizzare a, 3) l'agente crede che realizzerà a sotto determinate circostanze.

3 Una formalizzazione logica: BDICTL*

Una prima formalizzazione logica delle nozioni di impegno ed intenzioni è data da Cohen e Levesque (1990), che adottano una struttura a mondi possibili in cui ciascun mondo è una struttura temporale lineare ed introducono le modalità di convinzione, scopo, scopo persistente e intenzione, analizzandone le relazioni. Un altro esempio di *famiglia* di logiche BDI è dato dal modello formulato da Georgeff e Rao (1998), BDICTL*, che combina una *logica temporale ramificata* (CTL*, Computational Tree Logic) con una *logica multi-modale* (dove gli operatori modali Bel, Des e Int rappresentano rispettivamente convinzioni, desideri e intenzioni di agenti). La semantica delle modalità BDI è data dalle strutture di Kripke. Inoltre si assume che i mondi stessi siano strutture temporali ramificate: ciascun mondo può essere visto come una struttura di Kripke per una logica temporale ramificata. Wooldridge (2000) estende il modello BDICTL* per definire $\mathcal{LORA}$ ($\mathcal{L}$ogic $\mathcal{O}$f $\mathcal{R}$ational $\mathcal{A}$gents), nel cui linguaggio confluiscono il modello BDI, la logica classica del prim'ordine, la logica temporale ramificata e una logica dell'azione. Presento e discuto ora il modello BDICTL*, seguendo le esposizioni di Georgeff e Rao (1998), Wooldridge (2000) e Van der Hoek e Wooldridge (2003).

3.1 Linguaggio e semantica

L'alfabeto di BDICTL* (Cfr. Tabella 1) è costituito da un insieme non vuoto Φ di lettere proposizionali; dai connettivi proprosizionali $\wedge$ e $\neg$; dagli operatori modali Bel, Des e Int, per convinzioni, desideri ed intenzioni degli agenti; dai connettivi temporali $\bigcirc$, $\Diamond$, $\Box$, $\mathcal{U}$, $\mathcal{W}$, **A** e **E**; e da variabili e costanti individuali per rappresentare gli agenti. Ci sono due tipi di formule ben formate: le *formule di stato*, che sono vere in determinati mondi in determinati punti temporali e le *formule di cammino*, che sono vere in determinati mondi lungo determinati cammini (ossia sequenze di transizioni di punti temporali). Le formule di stato sono definite ricorsivamente, per cui: ogni proposizione atomica φ è una formula di stato; se φ e χ sono formule di stato anche $\neg\varphi$ e $\varphi \wedge \chi$ sono formule di stato; se φ è una formula di cammino $\mathbf{A}\varphi$ e $\mathbf{E}\varphi$ sono formule di stato; se φ è una formula di stato allora $(\mathsf{Bel}_i\varphi)$, $(\mathsf{Des}_i\varphi)$, $(\mathsf{Int}_i\varphi)$ sono formule di stato, dove i è un termine (variabile o costante) che indica un agente. Anche le formule di cammino sono definite ricorsivamente, in questo modo: ogni formula di stato è anche una formula di cammino; se φ e χ sono formule di cammino allora anche $\neg\varphi$ e $\varphi \wedge \chi$ sono formule di cammino; se φ e χ sono formule di cammino allora anche $\bigcirc\varphi$, $\Diamond\varphi$, $\Box\varphi$, $\varphi\mathcal{U}\chi$, $\varphi\mathcal{W}\chi$ sono formule di cammino.
Una struttura di Kripke per BDICTL* è definita dalla settupla $\mathcal{M}$:

$$\mathcal{M} = \langle W, \{T_w : w \in W\}, \{R_w : w \in W\}, L, \mathcal{B}, \mathcal{D}, \mathcal{I}\rangle$$

- W è l'insieme dei mondi, T è l'insieme dei punti temporali, $R \subseteq T \times T$ è una relazione *totale* ($\forall t \in T, \exists t' | t' \in \boldsymbol{T}\ e\ (t, t') \in \boldsymbol{R}$) che rappresenta tutte le possibili evoluzioni del sistema.

- Un mondo $w \in W$ su T e R è una coppia $\langle T_w, R_w\rangle$, dove $T_w \subseteq T$ e $R_w \subseteq R$. Si noti che nel modello in analisi, i mondi non sono stati istantanei, ma strutture temporali ramificate:

$(\mathsf{Bel}_i\varphi)$	L'agente i crede φ
$(\mathsf{Des}_i\varphi)$	L'agente i desidera φ
$(\mathsf{Int}_i\varphi)$	L'agente i intende φ
$\bigcirc\varphi$	φ è soddisfatta nel punto temporale successivo
$\Diamond\varphi$	φ è soddisfatta "adesso" o in un punto temporale successivo
$\Box\varphi$	φ è sempre soddisfatta
$\varphi\mathcal{U}\chi$	φ è soddisfatta fino a quando χ è soddisfatta
$\varphi\mathcal{W}\chi$	φ è soddisfatta a meno che χ è soddisfatta
$\mathbf{A}\varphi$	φ è soddisfatta in ogni cammino
$\mathbf{E}\varphi$	φ è soddisfatta in qualche cammino

Tabella 1: Denotazione di alcuni elementi dell'alfabeto di BDICTL*

l'intuizione è che tali strutture rappresentino l'incertezza di un agente non solo sullo stato presente del mondo, ma anche sulla possibile evoluzione di questo.

- L è una valutazione proposizionale classica per ciascun mondo $w \in W$ in ciascun punto temporale $t \in T$: $L\langle w,t\rangle : \Phi \longrightarrow \{0,1\}$.

- $\mathcal{B}$ è una funzione che assegna ad ogni agente una *relazione di accessibilità per le convinzioni*, ovvero una relazione su mondi e punti temporali, come segue:

$$\mathcal{B}\text{: } \textit{Agenti} \longrightarrow \wp(W \times T \times W)$$

Con un abuso di linguaggio si dirà che $\mathcal{B}$ è una relazione di accessibilità per le convinzioni. Si scrive $\mathcal{B}^w_t(i)$ per indicare l'insieme dei mondi accessibili all'agente i a partire dal mondo w al tempo t. Dal momento che $\mathcal{B}$ dipende da un mondo w e un tempo t determinati, il risultato dell'applicazione di $\mathcal{B}$ su di una situazione differente può essere diverso: in questo modo si esprime il fatto che l'agente può cambiare le proprie convinzioni a proposito delle opzioni disponibili. Formalmente, $\mathcal{B}^w_t(i) = \{w' | \langle w,t,w'\rangle \in \mathcal{B}(i)\}$. $\mathcal{D}$ e $\mathcal{I}$ sono definite analogamente. Intuitivamente, i mondi $\mathcal{B}$-$\mathcal{D}$-$\mathcal{I}$-accessibili sono rispettivamente quelli che l'agente crede che siano possibili, che desidera realizzare e che intente concretizzare.

La soddisfacibilità di una formula è definita rispetto ad una struttura $\mathcal{M}$, un mondo w e un punto temporale t. L'espressione $\mathcal{M}, \langle w,t\rangle \vDash \varphi$ si legge "la struttura $\mathcal{M}$ nel mondo w e nel punto temporale t soddisfa φ". Un cammino $(t_0, t_1, \dots)$ in un mondo w si scrive $(\langle w,t_0\rangle, \langle w,t_1\rangle, \langle w,\dots\rangle)$.

Esempio 2* Formalizzo l'Esempio 2 relativamente al linguaggio e alle strutture semantiche fondamentali. Siano: *Sen*= concorrere per un seggio al Senato; *Sen(Win)*= vincere un seggio al Senato; *Rep*= mantenere il seggio alla Camera dei Rappresentanti; *Rit* = ritirarsi dalla politica; *Poll* = indire il sondaggio; *Poll(Yes)* = la maggioranza approva il passaggio al Senato.

Seguono alcuni esempi di formule espresse nel linguaggio BDICTL* con relativa interpretazione (l'indice P denota l'agente, Phil):

- $\mathsf{Bel}_P(Sen\mathcal{W}Rit)$: Phil crede di concorrere per un seggio al Senato a meno che non si ritiri dalla politica;

- $\mathsf{Bel}_P(\bigcirc Sen(Loss))$: Phil crede che in seguito non vincerà un seggio al Senato;

- $\mathsf{Bel}_P(Poll \rightarrow \mathbf{E}Poll(Yes))$: Phil crede che se indice il sondaggio allora è possibile che la maggioranza approvi il passaggio al Senato;
- $\mathsf{Bel}_P(\mathbf{E}Poll(Yes))$: Phil crede che sia possibile che la maggioranza approvi il passaggio al senato;
- $\mathsf{Bel}_P(\mathbf{A}\Diamond Rep)$: Phil crede che in ogni caso potrà mantenere il seggio alla Camera dei Rappresentati;
- $\neg\mathsf{Des}_P Rit$: Phil non desidera ritirarsi dalla politica;
- $\mathsf{Int}_P Poll$: Phil intende indire il sondaggio;
- $\neg\mathsf{Int}_P(\neg Poll)$: Phil non intende non indire il sondaggio.

Nella Tabella 2 è rappresentato l'insieme W degli otto mondi possibili. La Tabella 3 rappresenta invece le relazioni di accessibilità. La prima colonna mostra i quattro mondi $\mathcal{B}$-accessibili di Phil: questi corrispondono alla vittoria o meno del seggio al Senato sulla base dell'esito del sondaggio. Nella seconda colonna sono riportati i mondi $\mathcal{D}$-accessibili: si noti come l'opzione di ritirarsi dalla politica non sia presente nei mondi desiderati, ma soltanto in quelli creduti (Phil crede che ritirarsi dalla politica sia un'opzione, ma non la considera seriamente). Infine la terza colonna mostra i mondi $\mathcal{I}$-accessibili: questi sottomondi dei precedenti rappresentano la scelta di Phil e il suo impegno a realizzarla (Phil intende indire il sondaggio).

Seguono alcuni esempi di formule soddisfatte e di formule non soddisfatte:

- $\mathcal{M}, \langle w_3, t_0\rangle \vDash \mathbf{E}Rit$
  La struttura $\mathcal{M}$ nel mondo w_3 e nel punto temporale t_0 soddisfa $\mathbf{E}Rit$ perchè esiste un cammino $(\langle w_3, t_0\rangle, \langle w_3, t_1\rangle, \langle w_3, t_{13}\rangle)$ in cui la struttura $\mathcal{M}$ soddisfa Rit: $\mathcal{M}, (\langle w_3, t_0\rangle, \langle w_3, t_1\rangle, \langle w_3, t_{13}\rangle) \vDash Rit$.
- $\mathcal{M}, \langle w_3, t_0\rangle \nvDash \mathbf{A}Rit$
  La struttura $\mathcal{M}$ nel mondo w_3 e nel punto temporale t_0 non soddisfa $\mathbf{A}Rit$ perchè non è vero che in ogni cammino del mondo la struttura soddisfa Rit: ad esempio, $\mathcal{M}, (\langle w_3, t_0\rangle, \langle w_3, t_1\rangle, \langle w_3, t_2\rangle) \nvDash Rit$.
- $\mathcal{M}, \langle w_3, t_0\rangle \vDash \mathsf{Bel}_P\mathbf{E}Rit$
  La struttura $\mathcal{M}$ nel mondo w_3 e nel punto temporale t_0 soddisfa $\mathsf{Bel}_P\mathbf{E}Rit$ perchè in ogni mondo $\mathcal{B}$-accessibile a partire dal mondo w_3 e dal tempo t_0 la struttura soddisfa $\mathbf{E}Rit$; in altre parole, in ogni mondo che Phil nella situazione $\langle w_3, t_0\rangle$ crede che sia possibile esiste un cammino in cui la struttura soddisfa Rit. Formalmente, $\forall v \in \mathcal{B}_{t_0}^{w_3}, \mathcal{M}, \langle v, t_0\rangle \vDash \mathbf{E}Rit$.

3.2 Assiomatizzazione

Assiomi modali. Presento e discuto gli assiomi modali della famiglia di logiche BDICTL* e, tramite la teoria della corrispondenza, le condizioni imposte sulle relazioni di accessibilità. Gli assiomi modali, che considerano individualmente ciascuna delle tre modalità, esprimono alcune tra le norme di razionalità analizzate dall'indagine filosofica.

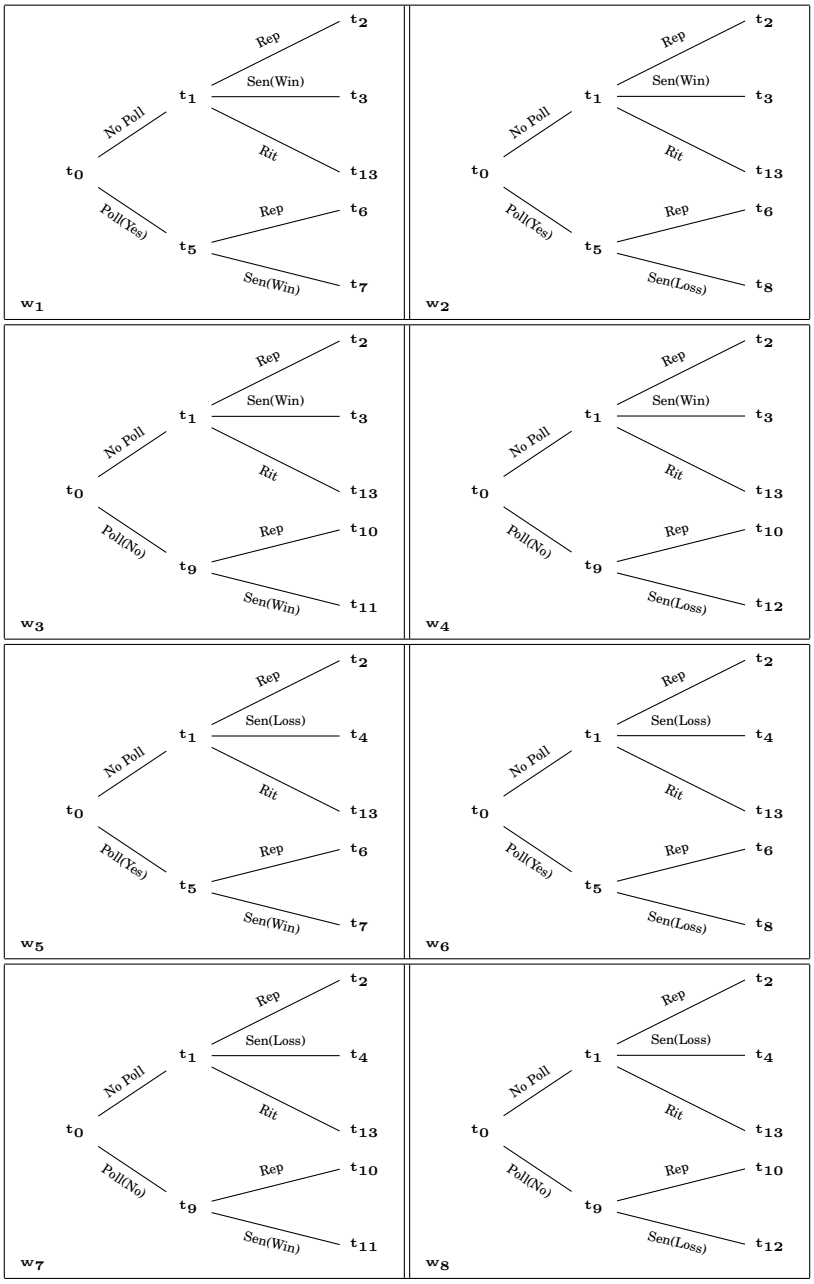

Tabella 2: L'insieme W dei mondi possibili dell'Esempio 2*.

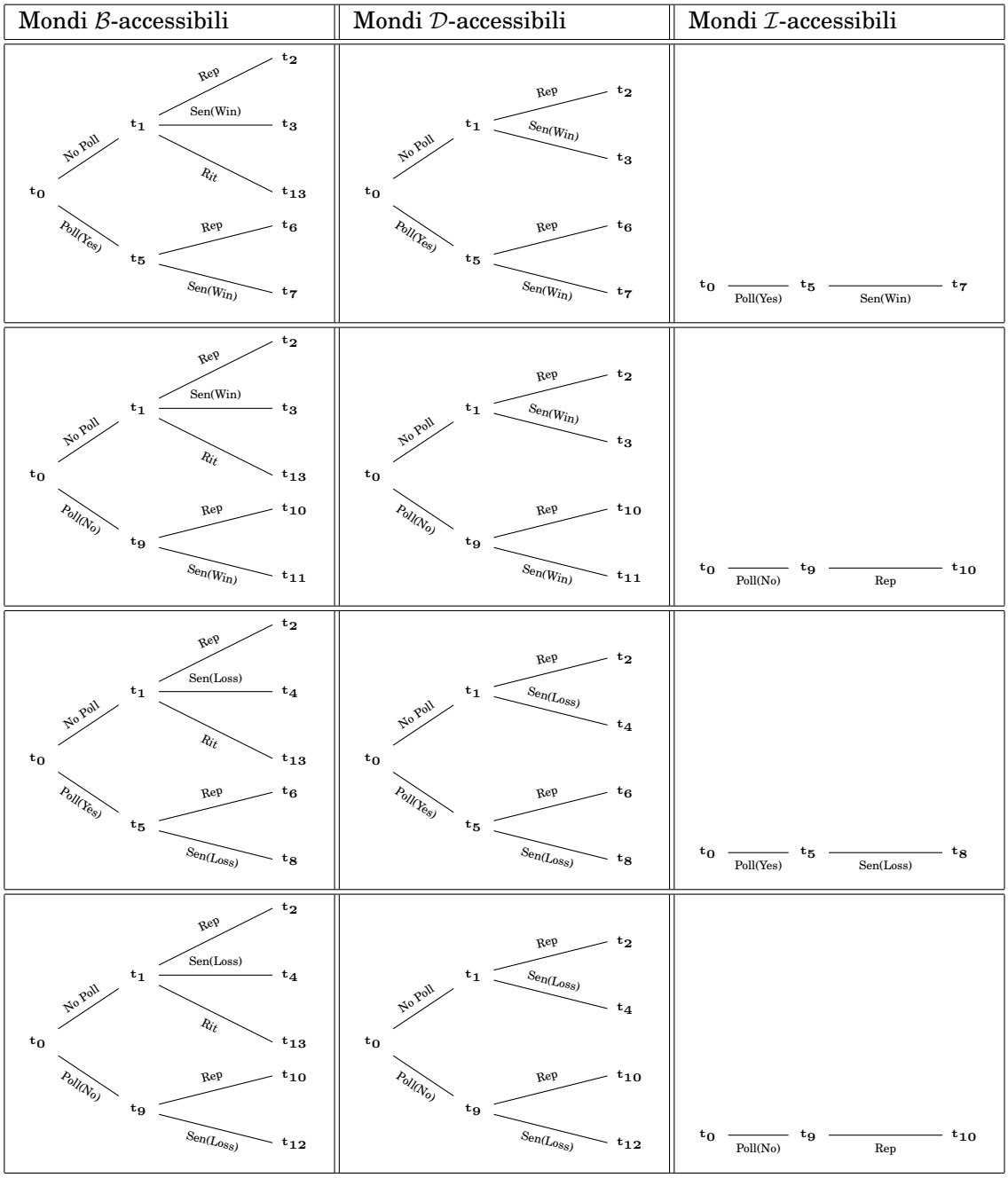

Tabella 3: I mondi $\mathcal{B}$-$\mathcal{D}$-$\mathcal{I}$-accessibili dell'Esempio 2*.

- $\mathbf{C}_B$: Se $w' \in \mathcal{B}_t^w(i)$, allora $t \in w$ e $t \in w'$
  $\mathbf{Gen}_B$: Se $\vdash \varphi$, allora $\vdash (\mathsf{Bel}_i\varphi)$
  $\mathbf{K}_B$: $(\mathsf{Bel}_i\varphi) \wedge (\mathsf{Bel}_i(\varphi \rightarrow \chi)) \rightarrow (\mathsf{Bel}_i\chi)$
  $\mathbf{D}_B$: $(\mathsf{Bel}_i\varphi) \rightarrow \neg(\mathsf{Bel}_i(\neg\varphi))$
  $\mathbf{4}_B$: $(\mathsf{Bel}_i\varphi) \rightarrow \mathsf{Bel}_i(\mathsf{Bel}_i\varphi)$
  $\mathbf{5}_B$: $\neg(\mathsf{Bel}_i\varphi) \rightarrow \mathsf{Bel}_i(\neg\mathsf{Bel}_i\varphi)$
  Questi sei assiomi equivalgono ad imporre che la semantica dell'operatore di convinzione corrisponda al sistema modale KD45. Infatti:
  $\mathbf{D}_B \Leftrightarrow \mathcal{B}$ è *seriale* (per ogni $\langle w, t\rangle$, esiste un w' tale che $w' \in \mathcal{B}_t^w(i)$)
  $\mathbf{4}_B \Leftrightarrow \mathcal{B}$ è *transitiva* (se $w' \in \mathcal{B}_t^w(i)$ e $w'' \in \mathcal{B}_t^{w'}(i)$, allora $w'' \in \mathcal{B}_t^w(i)$)
  $\mathbf{5}_B \Leftrightarrow \mathcal{B}$ è *euclidea* (se $w' \in \mathcal{B}_t^w(i)$ e $w'' \in \mathcal{B}_t^w(i)$, allora $w' \in \mathcal{B}_t^{w''}(i)$).

- $\mathbf{C}_D$: Se $w' \in \mathcal{D}_t^w(i)$, allora $t \in w$ e $t \in w'$
  $\mathbf{Gen}_D$: Se $\vdash \varphi$, allora $\vdash (\mathsf{Des}_i\varphi)$
  $\mathbf{K}_D$: $(\mathsf{Des}_i\varphi) \wedge (\mathsf{Des}_i(\varphi \rightarrow \chi)) \rightarrow (\mathsf{Des}_i\chi)$
  $\mathbf{D}_D$: $(\mathsf{Des}_i\varphi) \rightarrow \neg(\mathsf{Des}_i(\neg\varphi))$.
  Questi quattro assiomi equivalgono ad imporre che la semantica dell'operatore di desiderio corrisponda al sistema modale KD. Infatti:
  $\mathbf{D}_D \Leftrightarrow \mathcal{D}$ è *seriale* (per ogni $\langle w, t\rangle$, esiste un w' tale che $w' \in \mathcal{D}_t^w(i)$).

- $\mathbf{C}_I$: Se $w' \in \mathcal{I}_t^w(i)$, allora $t \in w$ e $t \in w'$
  $\mathbf{Gen}_I$: Se $\vdash \varphi$, allora $\vdash (\mathsf{Int}_i\varphi)$
  $\mathbf{K}_I$: $(\mathsf{Int}_i\varphi) \wedge (\mathsf{Int}_i(\varphi \rightarrow \chi)) \rightarrow (\mathsf{Int}_i\chi)$
  $\mathbf{D}_I$: $(\mathsf{Int}_i\varphi) \rightarrow \neg(\mathsf{Int}_i(\neg\varphi))$.
  Questi quattro assiomi equivalgono ad imporre che la semantica dell'operatore di intenzione corrisponda al sistema modale KD. Infatti:
  $\mathbf{D}_I \Leftrightarrow \mathcal{I}$ è *seriale* (per ogni $\langle w, t\rangle$, esiste un w' tale che $w' \in \mathcal{I}_t^w(i)$).

C, noto come assioma di compatibilità spazio-temporale, impone che se un mondo w' è accessibile per un agente a partire dalla situazione $\langle w, t\rangle$, allora t è un punto temporale sia in w che in w'. **C** è necessario poiché i mondi sono strutture temporali e le relazioni di accessibilità dipendono da una situazione determinata. **Gen**, nota anche come regola di necessitazione o di generalizzazione, impone che ogni formula valida sia creduta, desiderata o intesa, mentre **K**, detto assioma distributivo, è richiesto per ogni minimo sistema modale. **D** pone un vincolo di consistenza: come si è visto, convinzioni ed intenzioni di un agente razionale devono essere consistenti e non contraddittorie. Infatti se l'attentatore crede di indebolire il nemico, allora non crede di non indebolirlo e se intende sganciare una bomba non intende non sganciare una bomba. Consideriamo i desideri: mentre dall'analisi filosofica era emerso che i desideri possono essere inconsistenti, $\mathbf{D}_D$ impone che questi siano logicamente consistenti, cioè che se un agente desidera qualcosa allora non può desiderare la sua negazione. Il dissenso tra principi filosofici e formalizzazione logica è solo apparente: data la struttura temporale ramificata, desideri contrastanti possono portare l'agente lungo diversi cammini che non possono essere percorsi insieme. Nonostante i desideri siano logicamente consistenti, questi possono non essere tutti realizzabili, dal momento che un agente può eseguire soltanto un cammino tra le strutture ramificate delle esecuzioni possibili. Infine, $\mathbf{4}_B$ e $\mathbf{5}_B$ sono gli assiomi di introspezione positiva e negativa: insieme stabiliscono che, mentre un agente può avere soltanto convinzioni imperfette su ciò che crede vero nel mondo, egli ha convinzioni perfette riguardo le proprie convinzioni.

Assiomi intermodali. Come l'analisi filosofica studia le relazioni tra convinzioni, desideri e intenzioni, così la formalizzazione studia i nessi tra le tre modalità rappresentandoli come assiomi intermodali, generati cioè dai rapporti tra le relazioni di accessibilità. Georgeff e Rao (1998) individuano gli assiomi intermodali tramite un'analisi puramente combinatoria di rapporti, quali quelli di sottoinsieme o intersezione, tra le relazioni di accessibilità: ciascuna logica appartenente alla famiglia BDICTL* differisce dalle altre proprio per la selezione degli assiomi intermodali, i quali corrispondono ad una relazione tra $\mathcal{B}, \mathcal{D}, \mathcal{I}$. Per questo motivo, diversi tra questi assiomi intermodali non esprimono affatto i principi normativi del modello BDI e di conseguenza molti tra sistemi esposti dagli autori non formalizzano questo modello di razionalità. In ciò che segue, discuto gli assiomi intermodali della logica BDICTL*-W3, indicando come questi esprimano o permettano di derivare i principi normativi del modello.

I tre assiomi intermodali di BDICTL*-W3 equivalgono ad imporre che le seguenti intersezione tra le relazioni di accessibilità siano non vuote:

- **Ass1** : $(\mathsf{Des}_i\chi \Rightarrow \neg\mathsf{Int}_i\neg\chi) \Leftrightarrow \mathcal{I}_t^w(i) \cap \mathcal{D}_t^w(i) \neq \emptyset$
- **Ass2** : $(\mathsf{Bel}_i\chi \Rightarrow \neg\mathsf{Des}_i\neg\chi) \Leftrightarrow \mathcal{D}_t^w(i) \cap \mathcal{B}_t^w(i) \neq \emptyset$
- **Ass3** : $(\mathsf{Bel}_i\chi \Rightarrow \neg\mathsf{Int}_i\neg\chi) \Leftrightarrow \mathcal{I}_t^w(i) \cap \mathcal{B}_t^w(i) \neq \emptyset$

Ass1, noto come assioma di consistenza di desideri e intenzioni, affermando che un agente non intende la negazione di ciò che desidera, è in accordo con il principio per cui le intenzioni sono un sottoinsieme consistente dei desideri: la scelta dell'agente rispetta le proprie attitudini motivazionali. **Ass2** richiede che l'agente non desideri ciò che crede impossibile: questo vincolo impone che l'individuo non abbia desideri irrealizzabili, cioè che sia concreto. Come si è visto per l'assioma modale $\mathbf{D}_D$, anche **Ass2** non esclude che l'agente abbia desideri che lo portino lungo cammini divergenti e quindi non possa realizzarli tutti. Da **Ass3** discendono due importanti proprietà della teoria di Bratman: l'*Asymmetry Thesis*, di cui **Ass3** è la contrappositiva, e la soluzione a *The Problem of the Package Deal*. Si può dimostrare (Georgeff e Rao, 1998) come dagli assiomi di BDICTL*-W3 discendono entrambi gli argomenti, così formalizzati:

- *The Asymmetry Thesis*:
  - **AT1**, *Consistenza di intenzioni e desideri*:
    $\vDash (\mathsf{Int}_i\varphi) \Rightarrow (\neg\mathsf{Bel}_i\neg\varphi) \;\Leftrightarrow \nvDash (\mathsf{Int}_i\varphi) \wedge (\mathsf{Bel}_i\neg\varphi)$
  - **AT2**, *Incompletezza di intenzioni e desideri*:
    $\nvDash (\mathsf{Int}_i\varphi) \Rightarrow (\mathsf{Bel}_i\varphi) \;\Leftrightarrow \vDash (\mathsf{Int}_i\varphi) \wedge (\neg\mathsf{Bel}_i\varphi)$
  - **AT3**, *Incompletezza di desideri e intenzioni*:
    $\nvDash (\mathsf{Bel}_i\varphi) \Rightarrow (\mathsf{Int}_i\varphi) \;\Leftrightarrow \vDash (\mathsf{Bel}_i\varphi) \wedge (\neg\mathsf{Int}_i\varphi)$
- *Solution to The Problem of the Package Deal*:
  - **PD**: $\vDash (\mathsf{Int}_i\varphi) \wedge (\mathsf{Bel}_i(\varphi \rightarrow \chi)) \wedge (\neg\mathsf{Int}_i\chi)$
    $\Leftrightarrow \nvDash (\mathsf{Int}_i\varphi \wedge \mathsf{Bel}_i(\varphi \rightarrow \chi)) \;\Rightarrow\; (\mathsf{Int}_i\chi)$

4 Un esempio di *Agent Control Loop*

La teoria BDI è un modello interdisciplinare, dove i contributi forniti da diverse aree di ricerca interagiscono apportando prospettive differenti sullo stesso problema. Si considera

di seguito la componente informatica di programmazione di sistemi BDI capaci di agire in ambienti dinamici: nello specifico, esamino un esempio di *Agent Control Loop* (Wooldridge, 2000), cioè di un segmento di processo operato da un sistema. Una delle ragioni che giustificano i formalismi che seguono è la possibilità di numerose implementazioni: le applicazioni nel mondo reale spaziano infatti dagli assistenti automatici in rete al controllo del traffico aereo e alla gestione delle telecomunicazioni. La plausibilità cognitiva del modello BDI è confermata anche dalla costruzione di sistemi che, programmati secondo queste procedure, trovano numerosi impieghi nella realtà. La descrizione algoritmica delle diverse procedure di ragionamento di un agente razionale mette perciò in rilievo l'interesse pratico computazionale della teoria BDI.

Notazioni. Siano *Bel* l'insieme di tutte le convinzioni e *B* l'insieme delle convinzioni correnti di un agente; analogamente, siano *Des* l'insieme di tutti i desideri e *D* l'insieme dei desideri correnti di un agente. Infine, siano *Int* l'insieme di tutte le intenzioni possibili e *I* l'insieme delle intenzioni correnti di un agente. La *percezione* di un agente, ovvero le informazioni disponibili riguardo all'ambiente, è rappresentata da *impulsi percettivi* o *percepiti*. Si usa $\rho', \rho, \rho_1, \ldots$ e *Per* per indicare, rispettivamente, gli impulsi percettivi e l'insieme di tutti i percepiti. Siano $\pi', \pi, \pi_1, \ldots$ e *Plan*, rispettivamente, piani e l'insieme di tutti i piani. *execute*(π) è una procedura che prende come input un piano e lo esegue senza fermarsi; eseguire un piano significa eseguire ogni azione contenuta nel piano.

Procedure. A questo punto si possono discutere le formalizzazioni di tre tra le diverse procedure che portano un agente a compiere un'azione. Il *processo di aggiornamento delle convinzioni* è modellato dalla *funzione di revisione di convinzioni*, definita come:

$$\mathit{bfr}\colon \wp(\mathit{Bel}) \times \mathit{Per} \longrightarrow \wp(\mathit{Bel})$$

A partire dalle convinzioni correnti e dagli impulsi percettivi, la funzione di revisione di convinzioni stabilisce un nuovo insieme di convinzioni. Anche in BDICTL* l'agente può aggiornare le proprie convinzioni: il ruolo dinamico di *bfr* è sostituito dalla struttura temporale ramificata e dalla definizione di $\mathcal{B}$, il cui risultato dipende da precise coordinate temporali espresse dalla situazione.

Il *processo deliberativo* di un agente è descritto dalla funzione,

$$\mathit{deliberate}\colon \wp(\mathit{Bel}) \longrightarrow \wp(\mathit{Int})$$

che da un insieme di convinzioni, restituisce un insieme di intenzioni, quelle che l'agente vuole realizzare sulla base delle proprie convinzioni. Il processo di deliberazione è costituito da due fasi: nella prima, che chiamiamo *generazione di opzioni*, l'agente cerca di capire quali siano le opzioni disponibili; nella seconda, che indichiamo come *filtraggio*, l'agente sceglie una o più tra le opzioni appena selezionate, e si impegna a realizzarle. Formalmente:

$$\begin{aligned} \mathit{options}&\colon \wp(\mathit{Bel}) \times \wp(\mathit{Int}) \longrightarrow \wp(\mathit{Des}) \\ \mathit{filter}&\colon \wp(\mathit{Bel}) \times \wp(\mathit{Des}) \times \wp(\mathit{Int}) \longrightarrow \wp(\mathit{Int}) \end{aligned}$$

La funzione *options* a partire da convinzioni e intenzioni correnti di un agente, determina un insieme di opzioni, ovvero stabilisce un insieme di possibilità che, date le convinzioni sul mondo, si configura come adatto a raggiungere le proprie intenzioni. Queste opzioni saranno chiamate desideri, per sottolineare l'interpretazione intuitiva di un desiderio secondo cui, in un mondo ideale, un agente vorrebbe che tutti i propri desideri fossero realizzati. Tuttavia è

possibile che l'agente non sia grado di realizzare tutti i propri desideri, questo perchè spesso i desideri sono mutualmente esclusivi: perciò l'agente deve scegliere e la funzione *filter* rappresenta la selezione di un'opzione, quella cioè che l'agente si impegna a realizzare. La funzione *filter* interpreta il ruolo centrale delle intenzioni nel modello BDI emerso dall'analisi filosofica: queste condizionano e fungono da filtro di ammissibilità per le scelte successive dell'agente. Infatti *filter* esprime la relazione tra desideri, intenzioni e scelte secondo cui l'intenzione sarebbe ciò che è stato scelto con l'impegno per la sua realizzazione.

Il *ragionamento mezzi-fini* di un agente è rappresentato dalla funzione:

$$plan\colon \wp(Bel) \times \wp(Int) \longrightarrow Plan$$

che sulla base delle convinzioni e intenzioni correnti, seleziona il piano opportuno. Si vedrà, nell'*Agent Control Loop*, come *plan* segua *filter*: l'idea è la stessa esplicitata dal ragionamento di Bratman intorno ai piani come gerarchizzati, grazie ai quali quelli che mirano a fini più ampi condizionano quelli sui mezzi.

Agent Control Loop. Alla luce delle considerazioni precedenti esamino un esempio di *Agent Control Loop* (Wooldridge, 2000, p. 31), cioè un fragmento di un processo operativo.

```
Algorithm:  Agent Control Loop
1.
2.   B := B_0;
3.   I := I_0;
4.   while true do
5.       get next percept ρ;
6.       B := brf(B, ρ);
7.       D := options(B, I);
8.       I := filter(B, D, I);
9.       π := plan(B, I);
10.      execute(π);
11.  end-while
```

A partire da due insiemi rispettivamente di convinzioni (2) ed intenzioni iniziali (3), l'agente esamina l'ambiente in cui è immerso, tramite i sensi di cui è fornito, ricavandone un'osservazione (5). Da questo impulso percettivo è indotto a compiere una revisione delle proprie convinzioni rispetto al mondo: l'esito di questo processo può essere un insieme differente di informazioni sul contesto (6). A partire da queste nuove convinzioni, l'agente avvia il processo di deliberazione: innanzitutto vaglia le opzioni disponibili (7) e in un secondo istante decide quali di queste impegnarsi a realizzare (8). Quindi ragiona circa i mezzi per realizzare l'intenzione selezionata, stabilendo un piano tra quelli di cui dispone (9), ed esegue il piano (10).

5 Conclusione

Si è visto come il modello BDI, per fornire una caratterizzazione cognitivamente plausibile delle azioni degli agenti, riduca il grado di idealizzazione sugli individui proprio della teoria della scelta razionale. Questo passaggio determina due conseguenze. In primo luogo l'ontologia del modello, per migliorare il proprio potere espressivo, inserisce il concetto di intenzione

e ne analizza i rapporti con desideri e convinzioni. In secondo luogo l'approccio epistemologico del modello BDI presenta alcuni aspetti descrittivamente plausibili, come l'inconsistenza dei desideri, insieme ad altri di natura normativa, come i vincoli di consistenza sulle intenzioni.

L'analisi di questa duplice natura epistemologica della teoria è rafforzata dai contributi emersi dalle due formalizzazioni. Da un lato, infatti, il sistema logico che ho presentato sviluppa la componente normativa della teoria, già articolata nei termini di vincoli di consistenza sugli elementi dell'ontologia. I requisiti normativi per la razionalità dell'agente sono espressi dagli assiomi modali e intermodali: in altre parole, l'assiomatizzazione logica rappresenta e definisce le relazioni tra le norme di comportamento individuate dall'analisi filosofica. Dall'altro lato, l'implementazione del modello, attraverso una descrizione algoritmica della procedure di ragionamento di un agente razionale, ne dichiara l'interesse pratico computazionale e ne approfondisce gli aspetti di plausibilità cognitiva.

Si può quindi concludere affermando che logica teorica e implementazione reale approfondiscono i diversi aspetti dei principi filosofici del modello BDI: l'interazione e la pluralità di metodi di analisi potenzia la fecondità teoretica di questa concezione della razionalità.

Riferimenti bibliografici

Allen, James F. (1990). "Two Views of Intention: Comments on Bratman and on Cohen and Levesque". In: *Intentions in Communication*. Cambridge MA: The MIT Press, pp. 71–77.

Bratman, Michael E. (1987). *Intention, Plans, and Practical Reason*. Cambridge MA: Harvard University Press.

— (1990). "What is intention?" In: *Intentions in Communication*. Cambridge MA: The MIT Press, pp. 15–31.

Cohen, Philip R. e Hector J. Levesque (1990). "Intention is Choice with Commitment". In: *Artificial Intelligence* 42, pp. 213–261–339.

Georgeff, Michael P. e Anand S. Rao (1998). "Decision Procedures for BDI Logics". In: *Journal of Logic and Computation* 8.3, pp. 293–344.

Van der Hoek, Wiebe e Michael Wooldridge (2003). "Towards a Logic of Rational Agency". In: *Logic Journal of the IGPL* 11.2, pp. 135–159.

Wooldridge, Michael (2000). *Reasoning about Rational Agents*. Cambridge (MA): The MIT Press.