Ambiguità ed Autonomia negli agenti software

Matteo Bonifacio, Diego Ponte

***Abstract*— Il concetto di autonomia è la principale caratteristica che differenzia la programmazione ad agenti rispetto alla programmazione ad oggetti. Anche se in teoria tale concetto è stato descritto in modo dettagliato, in pratica non sembra essere stato sviluppato appieno. La tesi di questo articolo è che un agente software è autonomo solo se è capace, in un ambiente imprevedibile, complesso, ed ambiguo, di bilanciare due forme di razionalità: nella prima, l'agente deve essere in grado di sviluppare mezzi idonei a raggiungere un determinato fine; nella seconda, dati i mezzi a disposizione, l'agente deve essere in grado di adattare i propri fini a tali mezzi. La duplice modalità di ragionamento porta l'agente a considerare nella scelta non solo il valore degli obiettivi da raggiungere ma anche il valore delle risorse sviluppate non utilizzabili per raggiungere tali fini. In particolari condizioni, questo ragionamento porta l'agente a cambiare preferenze ed obiettivi in accordo agli stati del mondo raggiunti ed alle risorse disponibili. Si mostrerà infine come questo tipo di ragionamento offre una nuova prospettiva rispetto al concetto di autonomia.**

I. Introduzione

GLI agenti intelligenti[1] sono definiti come dei sistemi software capaci di porre in essere azioni autonome allo scopo di perseguire degli obiettivi prefissati [36]. L'importanza degli agenti risiede nel fatto che essi devono operare in ambienti complessi, dove l'incertezza e quindi la mancanza di informazioni è notevole. Russel e Norvig osservano che la complessità ambientale dipende da [24]:

- inaccessibilità ad informazioni aggiornate: grado in cui le informazioni necessarie alla scelta sono nascoste;
- indeterminatezza ambientale: grado in cui le azioni poste in essere hanno effetti imprevedibili;
- interrelazione degli eventi: grado in cui eventi attuali sono influenzati da eventi passati;
- dinamicità ambientale: grado in cui l'ambiente circostante cambia per eventi non dipendenti dall'agente;
- ambiente discreto o continuo: grado di granularità al quale i cambiamenti ambientali si manifestano.

Nell'ambito del software engineering il motivo per cui vengono implementati gli agenti risiede nel fatto che un ambiente con le caratteristiche sopraccitate non permette la progettazione di sistemi software che siano in grado di prevedere tutti gli eventi possibili. E' necessario progettare sistemi software che abbiano una conoscenza particolare ed isolata degli eventi e che quindi siano in grado di interagire in modo dinamico con l'ambiente. Tali sistemi software sono appunto gli agenti intelligenti [34]. Nel prossimo paragrafo si analizzerà la caratteristica peculiare degli agenti come elemento distintivo rispetto alla programmazione ad oggetti: l'autonomia. Nella terza sezione si approfondiranno le teorie della razionalità in quanto, come verrà successivamente presentato, l'autonomia è intrinsecamente legata al concetto di razionalità. Nella sezione quattro si presenteranno le linee guida per un modello di agente retrospettivo. Nella sezione cinque si presenterà un modello di ragionamento retrospettivo basato sul concetto di costo affondato. Infine, nella sezione sei si proporranno alcune conclusioni.

M. Bonifacio è Project Manager di EDAMOK (Enabling Distributed Knowledge Management) presso l'Università degli Studi di Trento (email: bonifacio@itc.it)

D. Ponte è un laureando in Economia e Commercio presso l'Università degli Studi di Trento (email: ponte@itc.it)

[1] Per brevità, nel resto dell'articolo quando si farà riferimento agli agenti software intelligenti si userà il termine agente se non diversamente specificato.

II. L'autonomia

Wooldridge afferma che la caratteristica principale degli agenti è l'autonomia. Tale concetto può essere considerato come la capacità di agire senza l'intervento esterno di umani o altri agenti [11] [22] [35]. Detto diversamente, gli agenti hanno un elevato controllo sulle proprie azioni. Tale tipo di controllo differenzia la programmazione ad agenti rispetto a quella ad oggetti. Infatti, gli oggetti non hanno un controllo pieno sulle proprie azioni. Si può ad esempio osservare che, quando un'entità richiama un metodo pubblico di un determinato oggetto, tale oggetto deve eseguire l'azione richiesta; sull'esecuzione del metodo l'oggetto non ha alcun controllo[2]. L'autonomia viene spesso descritta in modo più approfondito nei termini di 'proattività', detta anche 'orientamento al goal': un agente proattivo deve essere in grado - una volta ricevuto un obiettivo dall'esterno - di perseguirlo ponendo in essere quelle azioni non necessariamente previste dal progettista ma utili alla realizzazione del compito assegnato [6]. Questa caratteristica è assente negli oggetti. In altre parole, rispetto agli oggetti, gli agenti sono in grado di agire in autonomia al fine di soddisfare gli obiettivi assegnati dall'utente o dal progettista.

L'idea di autonomia appena presentata sembra ricadere in una delle due modalità di intendere l'autonomia proposte da Castelfranchi [5]. In particolare la prima concezione di autonomia ricade sotto il concetto di *executive autonomy*; essa rappresenta la discrezionalità posseduta da un soggetto relativamente al modo di perseguire un obiettivo. In modo critico Castelfranchi fa notare che tale tipo di agente non può essere considerato completamente autonomo. L'esempio che Wooldridge utilizza per descrivere gli agenti sembra rispecchiare esattamente questo tipo 'debole' di autonomia:

> "Imagine an autonomous automatic pilot controlling an aircraft, that we present with the goal of safely landing at some airport. We expect the system to plan how to achieve this goal (perhaps by making use of

[2] "The locus of control with respect to the decision about whether to execute an action is thus different in agent and objects systems. In the object-oriented case, the decision lies with the object that invokes the method. In the agent case, the decision lies with the agent that receive the request." [34]

pre-compiled plans, rather than reasoning from first-principles), and if necessary, we expect it to generate subsidiary goals (e.g., ascend to an altitude of 30,000 feet, then proceed due north at a speed of...).” [34]

In contrasto alla prima tipologia, Castelfranchi presenta anche la nozione di *goal autonomy* che esprime la discrezionalità dell'agente relativa alla scelta dell'obiettivo da perseguire. L'agente non è, come dice lo stesso Castelfranchi, uno 'schiavo razionale', ma piuttosto un sistema in grado di decidere come agire nei confronti di terze parti sulla base di un interesse generato endogenamente.

Come si nota, i concetti di *executive* e *goal autonomy* sono fortemente legati al rapporto tra mezzi e fini ed alla sua realizzazione. Tali concetti si distinguono tra di loro in quanto la *executive autonomy* agisce a livello di selezione dei mezzi più appropriati per soddisfare un dato goal, mentre la *goal autonomy* agisce direttamente sulla formazione endogena (“. . . non derivata dal volere di terzi” [5]) del goal da perseguire.

Il concetto di autonomia proposto da Castelfranchi è legato quindi alla discrezionalità che l'agente ha sugli elementi fondamentali del processo decisionale. Tali elementi, mezzi e fini, e il loro rapporto sono, così come proposto da Simon, le dimensioni sulle quali si articola il tema della razionalità, ovvero la capacità di un agente di decidere sulla base di una valutazione corretta del rapporto mezzi/fini [26]. Lo stesso Castelfranchi fa notare il legame stretto tra autonomia e razionalità essendo la prima determinata dalle assunzioni che si fanno circa la seconda. Verranno analizzati nei prossimi paragrafi i diversi concetti di razionalità.

III. La Razionalità

A. La Razionalità Prospettiva

L'approccio classico alla razionalità è chiamato 'mezzi-fini' in quanto prevede l'uso di mezzi per raggiungere determinati fini. Tale approccio è inoltre definito prospettivo perché la scelta deve essere fatta esclusivamente sulla base dei guadagni futuri e delle perdite future [14]. Questo processo prevede l'esistenza di due ipotesi importanti:

- la pre-esistenza degli obiettivi derivanti da una curva di preferenze coerente [18].
- la capacità del soggetto di svolgere un processo decisionale attraverso il quale scegliere il percorso più efficiente per ottenere tale obiettivo [27].

Come afferma March, la teoria classica della decisione non si pone il problema della scelta dell'obiettivo [17]. Egli, mostrando una certa perplessità al riguardo, si chiede come un elemento così importante possa essere ignorato in una teoria considerata come un punto di riferimento per prendere delle decisioni corrette[3].

L'unico elemento gestito direttamente dal decisore è il processo decisionale di ricerca delle strategie soddisfacenti per l'obiettivo. Simon, studiando tale processo decisionale, ha coniato il termine razionalità procedurale [28].

[3]“. . . it is reasonable to ask how something as conspicuous as the fluidity and ambiguity of objectives can plausibly be ignored in a theory that is offered as a guide to human choice behavior.” [17]

E' da notare la similarità esistente tra la descrizione del processo decisionale prospettivo, per quanto in condizioni di incertezza, ed il processo decisionale adottato per gli agenti. Attualmente gli agenti sembrano infatti essere descritti esclusivamente in termini di *executive autonomy*. Tali agenti non possono infatti scegliere gli obiettivi da perseguire sulla base di valutazioni endogene ovvero su valutazioni che nascono in base a specifici interessi personali [16]. Come afferma Steels, essi non sono in grado di cambiare radicalmente i propri comportamenti rispetto a quanto è stato previsto dal designer:

> “AI systems built using the classical approach are not autonomous, although they are automatic . . . these systems can never step outside the boundaries of what was foreseen by the designer because they cannot change their own behaviour in a fundamental way.” [31]

Lo stesso autore, ha quindi definito gli agenti come automatici piuttosto che autonomi[4].

Nel prossimo paragrafo si mostrerà un altro tipo di razionalità che apre la strada al concetto di ambiguità nel trattamento delle informazioni.

B. La Razionalità Retrospettiva

Anche la razionalità retrospettiva o ex-post come la razionalità prospettiva, agisce sul rapporto esistente tra mezzi ed obiettivi ma la natura di tale rapporto è inversa. Nella razionalità retrospettiva, infatti, gli obiettivi si costruiscono in base alle azioni poste in essere precedentemente [18]. Inoltre, l'interpretazione della situazione incide molto sull'intero processo:

> “Il concetto di razionalità a posteriori porta l'accento sulla scoperta delle intenzioni per effetto dell'interpretazione dell'azione piuttosto che quale posizione di premessa o anteriore. Le azioni sono viste come eventi esogeni e come fonti produttive di esperienze che una valutazione posteriore, a fatti avvenuti, si preoccuperà di organizzare. Tale valutazione regge sulle preferenze generate dall'azione e dai suoi effetti e le scelte trovano la loro giustificazione nella coerenza successiva che esse rivelano rispetto ad obiettivi, pure essi ricavati da una critica interpretazione della scelta. I modelli di razionalità a posteriori conservano, dunque, il criterio che l'azione debba essere compatibile o coerente con le preferenze, ma considerano l'azione un evento antecedente rispetto agli scopi.” [19]

Il primo elemento importante di questo processo risiede nel fatto che l'azione precede la decisione. In altre parole, gli obiettivi non sono scelti e, solo in un secondo tempo, perseguiti; tali obiettivi si possono formare durante o anche dopo che le azioni sono state poste in essere. Diverse motivazioni spingono i soggetti ad agire senza una ragione o senza uno scopo apparenti [20]. Tali azioni servono infatti a:

- diminuire la complessità ambientale: i soggetti pongono in essere azioni in modo tale da 'chiarirsi le idee';

[4]“. . . they do not make the laws that their regulatory activities seek to satisfy. These are given to them, or built into them.” [31]

- esplorare l'ambiente: in questo modo i soggetti cercano di scoprire nuove situazioni, nuovi obiettivi, ecc.;
- validare i goal scelti: i soggetti cercano di trovare conferme degli obiettivi scelti;
- forzare la realtà tramite le 'profezie che si autorealizzano': i soggetti forzano la realtà attraverso le loro azioni in modo tale da creare le condizioni di verità delle proprie credenze [33].

Il secondo punto importante della razionalizzazione a posteriori è l'interpretazione che si dà delle azioni poste in essere. Essa aiuta a costruire ed a modellare le preferenze del soggetto in modo che tali preferenze siano consistenti con le azioni passate. Gli obiettivi, di conseguenza, vengono scoperti in base alle preferenze generate dalle interpretazioni delle vicende. L'interpretazione delle situazioni porta all'idea di *enactment* [32]. Con tale concetto, Weick afferma che i soggetti, nel dare senso agli infiniti stimoli che ricevono, filtrano - in base al contesto, in base alle credenze, ecc. - da tali stimoli solo quella parte di elementi che ritengono più importanti o rilevanti di altri. Questo processo, chiamato *sensemaking*, permette ai soggetti di interpretare gli elementi filtrati come se fossero la realtà oggettiva sulla quale agire. In questo modo i soggetti 'strutturano' la realtà attraverso le proprie credenze e, con le azioni poste in essere in base a queste credenze, essi creano le condizioni in base alle quali la realtà costruita diventa 'oggettivizzata'.

L'interpretazione delle azioni 'a posteriori' è possibile a causa dell'ambiguità che pervade le situazioni nelle quali i soggetti si trovano [32]. L'ambiguità è considerata come:

> "...la mancanza di chiarezza o di coerenza nella realtà, nella causalità o nell'intenzionalità. Situazioni ambigue sono quelle che non si possono codificare in maniera precisa entro categorie reciprocamente esaustive ed esclusive. Finalità ambigue sono intenzioni che non possono essere definite chiaramente. ... Risultati ambigui sono quelli le cui caratteristiche o implicazioni risultano sfocate." [20]

Se l'ambiente risulta essere ambiguo, ovvero soggetto a molteplici interpretazioni, ambigui ed incerti risultano essere anche gli stimoli in base ai quali giudicare cosa è giusto e cosa è sbagliato [3]. L'interpretazione gioca quindi un ruolo fondamentale nella valutazione positiva o negativa delle situazioni ambigue. Essa può essere utilizzata come mezzo per dare senso alle azioni; è quindi *razionale* in un mondo confuso, equivoco, non strutturato, non predicibile - dove gli assiomi della razionalità prospettiva non vengono soddisfatti[5] - e del quale l'agente ha solo una conoscenza parziale e frammentata [25]. In questi casi le persone risultano essere maggiormente razionali quando sviluppano un processo retrospettivo piuttosto che prospettivo; esse cercano infatti di capire cosa sta avvenendo, di costruire obiettivi - che a priori non esistono o non sono considerati - capaci di dare senso agli eventi passati. Questa visione rappresenta quindi le decisioni come delle interpretazioni costruttive della realtà[6].

Quando un ambiente è sostanzialmente ambiguo, il ragionamento 'mezzi-fini' prospettivo spesso non funziona in quanto, come precedentemente visto, mancano molti presupposti basilari per tale ragionamento. Un metodo per chiarire la situazione è il ragionamento retrospettivo, tramite il quale, i soggetti costruiscono interpretazioni plausibili degli eventi passati. A questo punto, se un ambiente appare complesso e se l'equivocità è la ragione maggiore per essere razionali in senso retrospettivo, allora per un agente - che deve fronteggiare un ambiente simile - potrebbe essere corretto adottare una strategia decisionale retrospettiva. Per questo motivo, nel prossimo paragrafo si mostrerà come un processo decisionale che considera i costi affondati possa rappresentare la base economica sulla quale impostare un ragionamento ex-post. Tramite tale ragionamento, l'agente manipola endogenamente le proprie preferenze in modo da giustificare il proprio stato corrente. Come si proporrà alla fine, questo processo permetterà di selezionare e perseguire goal non prevedibili a priori poiché nati dalle nuove preferenze createsi.

IV. Linee guida per un modello retrospettivo

La progettazione di un modello economico basato sulla razionalità retrospettiva potrebbe fondarsi, secondo l'ipotesi di chi scrive, sul concetto di costi affondati. Essi sono definiti come investimenti passati irreversibili [23]; in teoria tali investimenti non dovrebbero rientrare nel calcolo razionale in quanto non possono più influenzare le decisioni future [14]. Ed è per questo motivo che essi sono definiti costi affondati, in inglese sunk costs. La particolarità di questi investimenti risiede nel fatto che è possibile recuperare il loro costo solo attraverso lo sfruttamento diretto degli stessi. Non possono essere infatti trasformati o scambiati in qualche altra risorsa o capacità [23].

Nella realtà gli individui contano nelle decisioni non solo i costi e i proventi futuri ma anche gli investimenti passati non recuperabili. Questo comportamento, considerato irrazionale proprio perché contrario ai postulati della decisione razionale, è stato provato da diversi esperimenti scientifici [29]. L'esempio successivo di influenzamento da costi affondati è tratto da Arkes e Blumer:

> "Immagina di essere il presidente di una compagnia aerea e di aver investito 10 miliardi di lire in un progetto di ricerca per la costruzione di un aereo non rilevabile da radar convenzionali, in altre parole un aereo invisibile. Quando il progetto è completato per il 90% un'altra compagnia inizia la promozione di un aereo analogo. Inoltre, appare evidente che l'aereo della compagnia concorrente è piu veloce e piu economico di quello progettato presso la tua compagnia. Investiresti il rimanente 10% dei fondi di ricerca per portare a termine il progetto del tuo aereo?
> L'85% del campione si dichiara favorevole al completamento del progetto." [1]

In questo esempio specifico, l'irrazionalità dei soggetti sussisterebbe nel fatto che essi continuano lo sviluppo dell'aeroplano in quanto hanno già speso una gran quantità di denaro nel progetto

[5] "L'ambiguità indica che i presupposti necessari per prendere razionalmente una decisione non sono soddisfatti." [32]

[6] "Decisions are often reached by focusing on reasons that justify the selection of one option over another. Different frames, context, and elicitation procedures highlight different aspects of the options and bring forth different reasons and considerations that influence decisions." [25]

non prestando attenzione al fatto che una ditta concorrente ha progettato un aeroplano più efficiente.

Diverse sono le giustificazioni, sia in campo psicologico che economico, date al comportamento influenzato da costi affondati; per una revisione dell'intero argomento si rimanda a [2]. Qui si elencano le principali spiegazioni:

- i decisori diventano 'risk-seeking' in quanto il costo affondato rappresenta una perdita [12];
- volontà eccessiva di non sprecare [1];
- autogiustificazione [30].

Nonostante i comportamenti osservati da questi studiosi siano considerati in termini negativi essi sono simili a quelli considerati in modo positivo da March e Weick. Ad esempio, in tutte e due le visioni si trovano argomenti quali:

- il guardare al passato: la prima visione considera questo comportamento utile per dare senso alle situazioni; la seconda visione invece, lo considera semplicemente come un comportamento volto a recuperare gli investimenti passati;
- la giustificazione di azioni passate: nella prima visione essa serve come metodo per rendere coerente le azioni e i risultati; la seconda visione invece, la considera come un modo per apparire razionali o per 'salvare la faccia';
- l'overcommitment: nella prima visione esso sta ad indicare che le persone perseguono delle azioni per motivi diversi; nella seconda invece, esso rappresenta l'effetto del comportamento giustificativo (le persone persistono in un corso d'azione per giustificare se stessi).

La differenza tra le due visioni risiede quindi nel come vengono interpretati questi comportamenti. Ciò che permette di collegare il comportamento influenzato da costi affondati con la razionalità retrospettiva è l'ambiguità. Essa, infatti, pervade gli elementi e le situazioni decisionali. A tal proposito è importante notare che esiste un filone di studi, chiamato "decision dilemma theory", il quale considera l'ambiguità delle informazioni come un elemento esplicativo dei comportamenti considerati irrazionali dai primi studi sui costi affondati[7]. Ad esempio, nel test riguardante l'areoplano, il decisore potrebbe lecitamente pensare di poter utilizzare l'investimento in usi alternativi. Nella domanda proposta mancano infatti le informazioni prospettive ovvero quelle informazioni necessarie ad effettuare delle scelte secondo il modello decisionale classico. In un articolo pionieristico in tal senso, Bowen afferma che molti comportamenti considerati irrazionali dagli studiosi dei costi affondati sono tali solo se visti attraverso una visione decontestualizzata e a posteriori. E' da notare che Bowen fa riferimento a concetti come 'ambiente attivato' ed ambiguità considerati nella prima concettualizzazione di razionalità retrospettiva da Weick. Bowen afferma infatti:

> "... decision makers will continue to invest in courses of action beyond the point where others, having enacted a different reality with possibly different decision standards, believe prudent. It is also understandable that resources may be reinvested in a course of action, assuming some degree of commitment to that course of action, because of the equivocality inherent in the situation and not because of an overcommitment to a failed investment." [3]

Bowen afferma inoltre che l'etichettatura di determinati comportamenti in situazioni di ambiguità come 'affetti da costi affondati' (etichettatura chiamata *fallacia retrospettiva*) è dovuta al fatto che chi giudica la situazione a posteriori prende come elemento di valutazione il risultato finale e non le previsioni iniziali sui risultati futuri [4].

Oltre alla "decision dilemma theory", si passeranno ora in rassegna altre due importanti spiegazioni dei comportamenti condizionati dai costi affondati. Gli studiosi della "reinforcement history" affermano che, il comportamento influenzato da sunk costs non è altro che un comportamento legato al fatto che i decisori imparano a perseguire una certa traiettoria di azioni sulla base dell'esperienza passata [13]. Detto diversamente, i decisori formano delle catene dei risultati passati; sulla valutazione complessiva di tali catene essi basano la scelta. Questo comportamento rispecchia il processo bayesiano di aggiornano delle credenze in modo tale da effettuare previsioni future in condizioni di incertezza [10]. I singoli risultati negativi non portano quindi le persone a pensare che i propri comportamenti siano scorretti.

Infine, gli studiosi dell'"attribution theory" puntano invece l'attenzione sulla mancanza di informazioni. I decisori rimarrebbero sul sentiero intrapreso (corrispondente ad un comportamento influenzato da costi affondati) non in quanto irrazionali ma a causa della mancanza di informazioni [21]. In questo modo i soggetti cercano di incamerare l'informazione necessaria[8]. Il comportamento persistente sarebbe quindi aderente ad un principio di prudenza, in quanto i decisori scelgono di sospendere la scelta in mancanza di dati sufficienti, piuttosto che di irrazionalità.

Nel prossimo paragrafo si proporrà un modello di ragionamento retrospettivo basto sul concetto di costo affondato che permetterà la produzione endogena di obiettivi e preferenze.

V. Ambiguità, Costi affondati e Goal Autonomy

Il ragionamento retrospettivo, che in questa sede si vuole collegare al concetto di costi affondati, è stato sintetizzato da Doyle:

> "When people realize they are in situations that they have never considered before, they do not judge themselves to be irrational. Instead, they simply try to decide what beliefs and preferences to adopt (if any)." [8].

Doyle afferma che quando un soggetto affronta una situazione inaspettata, egli prova a cambiare idea sull'obiettivo da perseguire - adattando credenze e preferenze - piuttosto che considerarsi irrazionale. E' da notare che in questa definizione sono impliciti i concetti di interpretazione e cambiamento delle

[7] I primi esperimenti sui costi affondati sono stati effettuati per dimostrare che l'irreversibilità degli investimenti influenza la decisione razionale. Maggiore è il grado di irreversibilità, maggiore è il grado di influenzamento. Gli autori sono invece attualmente coinvolti in un esperimento volto a dimostrare che non è l'irreversibilità degli investimenti a influenzare i decisori ma piuttosto il grado di ambiguità sull'irreversibilità degli investimenti. Tale ambiguità permette infatti molteplici interpretazioni sui possibili utilizzi futuri degli investimenti.

[8] "... delaying exit decisions under equivocal conditions is not necessarily erroneous; it may be the case that investors are waiting to gather more information about the situation" [15]

preferenze presentati nel paragrafo riguardante la razionalità ex-post. I soggetti si trovano quindi in situazioni ambigue dove le informazioni posso essere interpretate in modi diversi.

Come verrà presentato a breve, l'ipotesi di chi scrive è che, in un ambiente dove l'informazione è soggetta a molteplici interpretazioni, il cambiamento degli obiettivi e delle preferenze in un agente può avvenire sulla base delle risorse sviluppate in passato.

A. Generazione endogena di goal

Si consideri un agente a cui viene assegnato un obiettivo. Si consideri inoltre, che tale obiettivo possa essere scomposto in tanti stadi da raggiungere in sequenza. Il percorso che porta all'obiettivo principale avrà la tipica forma ad albero. E' probabile che, più l'obiettivo è complesso da raggiungere, più l'agente dovrà acquisire risorse (intese anche come capacità da dover sviluppare) necessarie a perseguire tale obiettivo. A questo proposito, a causa delle mutevoli e poco chiare condizioni ambientali, l'agente può arrivare a detenere una certa quota di costi affondati che non sono utili per il raggiungimento dell'obiettivo originale. Infatti, le caratteristiche ambientali, proposte ad inizio articolo, possono condurre l'agente su percorsi devianti rispetto all'itinerario prestabilito; tali percorsi possono portare verso stati del mondo nei quali le risorse a disposizione possono essere inutili per rientrare nella traiettoria prefissata. A questo punto, l'agente avrà la necessità di cambiare le proprie risorse con mezzi più adatti. Ma, se le risorse possedute rientrano nella categoria dei costi affondati - ovvero sono almeno in parte irreversibili - esse incontrano delle difficoltà nell'essere scambiate con altre risorse. Tali costi affondati spingono quindi l'agente ad operare in due modi distini: da una parte egli è incentivato a creare un mercato limitando in tal modo l'irreversibilità degli stessi; dall'altra l'agente è incentivato al riuso di tali risorse in quanto lo sfruttamento diretto risulta essere l'unico modo per utilizzarle [23]. Bisogna notare che tale sfruttamento permette di ammortizzare il loro costo storico in più periodi permettendo al costo medio unitario della risorsa di decrescere. La riduzione progressiva del costo medio unitario produce un risparmio di costo 'retrospettivo' - ovvero soltanto dopo che la risorsa è stata sfruttata. Si considerino ad esempio questi dati:
Costo storico della risorsa = 100
Costo variabile per utilizzo = 3
Il costo medio medio unitario al primo utilizzo sarà uguale al costo storico più il costo variabile:

$$100 + 3 = 103$$

Il costo medio medio unitario al secondo utilizzo sarà invece uguale a:

$$\frac{100}{2} + 3 = 53$$

Infine, il costo medio medio unitario al terzo utilizzo sarà:

$$\frac{100}{3} + 3 = 33,33$$

Come si può osservare, maggiore è l'utilizzo, maggiore è il decremento del costo medio unitario. In effetti, il risparmio di costo 'retrospettivo' è causato dall'ammortamento del costo storico per il numero di utilizzi effettuati.

Quello che si vuole affermare qui è che stati del mondo raggiunti in modo imprevedibile potrebbero essere considerati come obiettivi, se nel processo decisionale venissero contati i costi affondati. In tal modo l'agente sfrutterebbe le proprie risorse in quanto la razionalizzazione guidata dai sunk costs avrebbe la peculiarità di valorizzare tali costi affondati.

Il processo di razionalizzazione appena presentato risulta essere il primo tentativo di generazione endogena di goal. Infatti, lo stato del mondo raggiunto casualmente può diventare un obiettivo solo retrospettivamente ovvero dopo essere stato raggiunto. La prima manifestazione di *goal autonomy* è quindi la razionalizzazione di eventi imprevedibili in eventi preferiti. In questo caso, nonostante l'obiettivo sia stato scelto in modo endogeno, le preferenze dell'agente rimangono ancora immutate rispetto a quelle inizialmente date. Gli obiettivi infatti, vengono giudicati ancora con la vecchia curva di utilità nonostante sia modificata in modo tale da contare i costi affondati.

In determinate circostanze, l'agente può invece essere portato a cambiare preferenze quando, date le proprie risorse affondate, la curva di utilità non permette a nessun stato del mondo raggiunto e a nessun obiettivo visibile di avere un valore positivo. Questo caso verrà presentato nel prossimo paragrafo.

B. Generazione endogena delle preferenze e nuovi obiettivi

Quando l'agente si trova a dover affrontare situazioni che agli occhi della sua curva di utilità e delle sue risorse hanno tutte dei valori negativi, egli ha l'opportunità di cambiare le proprie preferenze. In tal modo l'agente può far coincidere gli stati del mondo raggiunti con le proprie risorse affondate. Questo cambiamento può essere effettuato adottando, ad esempio, un processo controfattuale che può essere espresso con una frase del tipo: "Cosa avrei dovuto preferire in modo da essere soddisfatto con lo stato del mondo raggiunto?" [9]. Tramite il ragionamento controfattuale, l'agente potrebbe chiedersi cosa preferire, date le sue risorse, in modo da essere soddisfatto con la situazione attuale nel quale si trova. Questo processo è molto più radicale di quello precedentemente visto in quanto l'agente deve agire sui valori che stanno alla base delle proprie scelte. Riconsiderando le parole precedentemente citate di Steels, l'agente non è più semplicemente automatico, ma diventa autonomo in quanto modifica le 'leggi' in base alle quali effettua le scelte.

Apparentemente il comportamento di quest'agente sembrerebbe conservativo. Infatti, quando il peso degli investimenti effettuati in passato sovrasta il valore dell'utilità futura degli obiettivi attuali, l'agente potrebbe fermarsi e giustificare retrospettivamente la situazione raggiunta. In realtà, da questo comportamento possono derivare nuovi comportamenti 'prospettivi'. In effetti, nuovi obiettivi possono derivare dall'idea di adozione sociale dei goal proposta da Castelfranchi [7]. Attraverso la lente delle preferenze createsi durante il processo controfattuale, l'agente è in grado di assegnare nuovi valori agli stati del mondo accessibili alla sua conoscenza. In questo modo, egli è in grado di riordinare tali stati del mondo sulla base della nuova scala di preferenze e di scegliere gli obiettivi che meglio soddisfano le preferenze. Ad esempio, osservando le situazioni nelle quali si trovano gli altri agenti, l'agente potrebbe scoprire stati del mondo preferibili rispetto alla sua

situazione attuale; potrebbe quindi decidere di adottare quello stato del mondo che apporta un beneficio netto maggiore. Infine, come sopra, questo processo porterebbe all'acquisizione di nuove risorse e capacità che, a causa della complessità ambientale, potrebbero rendere difficile il perseguimento di tale obiettivo e portare l'agente ad un nuovo processo di razionalizzazione.

VI. Conclusioni

Gli argomenti trattati portano a chiedersi se, in un ambiente ambiguo ovvero soggetto a molteplici interpretazioni e/o poco chiaro, gli agenti attualmente implementati (che sembrano rispecchiare un'autonomia di tipo esecutivo), siano veramente razionali o lo siano solo in parte. Essi sembrano essere dotati di un tipo di razionalità che funziona correttamente solo in ambienti stabili e poco complessi:

> "...it has been repeatedly observed that axioms of rational choice which are often violated in non-transparent situations are generally satisfied when their application is transparent." [25]

Inoltre, se la tecnologia software ad agenti considera solo questo tipo di razionalità, essa dimentica di prendere in esame un comportamento razionale basato sulla scoperta degli obiettivi in tempi successivi rispetto al momento in cui gli agenti agiscono: la razionalità retrospettiva.

A questo punto, se il concetto di autonomia è la caratteristica cardine che differenzia la programmazione ad agenti rispetto a quella ad oggetti, bisogna chiedersi, come fa Castelfranchi, se l'agente *executive autonomous* sia da considerarsi un agente potenzialmente irrazionale piuttosto che uno 'schiavo razionale' che persegue obiettivi dati dall'esterno [5]. Tali obiettivi infatti, potrebbero portare l'agente ad agire in modo irrazionale in quanto egli perseguirebbe obiettivi in conflitto con i propri interessi (ad esempio le risorse sviluppate). Noi crediamo, come fa lo stesso Castelfranchi, che se un agente non è autonomo sui suoi obiettivi è potenzialmente irrazionale. Queste considerazioni portano a chiedersi come progettare un agente *goal autonomous*, ovvero un agente capace di generare in modo endogeno obiettivi all'interno di un ambiente imprevedibile.

In questo articolo si è tentato di presentare un modello di *goal autonomy*, nel quale l'agente genera endogenamente preferenze ed obiettivi in base alle risorse sviluppate in passato. La considerazione di tali risorse nel processo decisionale non risulta essere irrazionale quando i soggetti si trovano a dover affrontare degli ambienti imprevedibili e delle circostanze equivoche. Infatti, in tali situazioni i risultati delle azioni non sono sempre determinabili a priori. In breve, in questo modello, l'agente sviluppa preferenze ed obiettivi in modo che tali obiettivi siano consistenti con le risorse, classificabili come costi affondati, da lui possedute.

References

[1] H. R. Arkes and C. Blumer, *The psychology of sunk cost*. Organizational Behavior and Human Performance, Vol. 35, 1985.

[2] H. R. Arkes, *Incremental Investment Decisions and The Diagnosticity of Feedback*. In pubblicazione, 2002.

[3] M. G. Bowen, *The Escalation Phenomenon Reconsidered: Decision Dilemmas or Decision Errors?* Academy of Management Review, Vol. 12, 1987.

[4] M. G. Bowen and F. C. Power, *The moral manager: Communicative ethics and the Exxon Valdez disaster*. Business Ethics Quarterly, Vol. 3, 1993.

[5] C. Castelfranchi, *Guarantees for Autonomy in Cognitive Agent Architecture*. Intelligent Agents: Theories, Architectures, and Languages. Springer-Verlag, 1995.

[6] C. Castelfranchi, *Engineering Social Order*. Societies in the Agent World, First International Workshop, Springer-Verlag, vol. 1972, 2000.

[7] R. Conte and C. Castelfranchi, *Cognitive and social Action*. UCL Press, 1995.

[8] J. Doyle, *Rationality and Its Roles in Reasoning*. Computational Intelligence, Vol. 8, 1992.

[9] R. Ferrario, *Conterfactual Reasoning*. Proceedings of the 3th International and Interdisciplinary Conference on Modelling and using Context, Vol. 2116, Springer-Verlag, 2001.

[10] B. O'Flaherty and J. L. Komaki, *Going beyond with Bayesian updating*. Journal of Applied Behavior Analysis, Vol. 25, 1992.

[11] S. Franklin and A. Graesser, *Is it an Agent, or just a Programm? A Taxonomy for Autonomous Agents*. Third International Workshop on Agent Theories, Architectures and Languages. Springer-Verlag, 1996.

[12] H. Garland and S. Newport, *Effects of Absolute and Relative Sunk Cost on the Decision to Persist with a Course of Action*. Organizational Behavior and Human Decision Processes, Vol. 48, 1991.

[13] S. M. Goltz, em Examining the Joint Roles of Responsibility and Reinforcement History in Recommitment. Decision Sciences, Vol. 24, 1993.

[14] M. L. Katz and H. S. Rosen, *Microeconomia*. McGraw-Hill, 1996.

[15] D. H. Hantula and J. L. DeNicolis Bragger, *The Effects of Feedback Equivocality on Escalation of Commitment: An Empirical Investigation of Decision Dilemma Theory*. Journal of Applied Social Psychology, Vol. 29, 1999.

[16] M. Luck and M. d'Inverno, *Autonomy: A Nice Idea in Theory*. Agent Theories, Architectures, and Languages, 2000.

[17] J. G. March and J. P. Olsen, *Ambiguity and choice in organizations*. Bergen, 1976.

[18] J. G. March, *How decisions happen in Organizations*. Human Computer Interaction, Vol. 6, 1991.

[19] J. G. March, *Decisioni e organizzazioni*. il Mulino, 1993.

[20] J. G. March, *Prendere decisioni*. Il Mulino, 1998.

[21] B. E. McCain, *Continuing investment under conditions of failure: A laboratory study of the limits to escalation*. Journal of Applied Psychology, Vol. 71, 1986.

[22] H. S. Nwana, *Software Agents: An Overview*. Knowledge Engineering Review, Vol. 11, 1995.

[23] B. di Bernardo and E. Rullani, *Il management e le macchine*. il Mulino, 1990.

[24] S. Russel and P. Norvig, *Artificial Intelligence: A Modern Approach*. Prentice-Hall, 1995.

[25] E. Shafir and I. Simonson and A. Tversky, *Reason-based Choice*. Cognition, Vol. 49, 1993.

[26] H. A. Simon, *A behavioral Model of Rational Choice*. The Quarterly Journal of Economics, Vol. 69, 1954.

[27] H. A. Simon, *Reason in human affairs*. Stanford University Press, 1983.

[28] H. A. Simon, *Dalla razionalità sostanziale alla razionalità procedurale*. Le ragioni delle Organizzazioni Economiche, Rosenberg e Sellier, 1994.

[29] D. Soman, *The Mental Accounting of Sunk Time Costs: Why Time is not Like Money*. Journal of Behavioral Decision Making, Vol. 14, 2001.

[30] B. Staw, *The Escalation of Commitment To a Course of Action*. Academy of Management Review, Vol. 6, 1981.

[31] L. Steels, *When are robots intelligent autonomous agents?* Journal of Robotics and Autonomous Systems, Vol. 15, 1995.

[32] E. K. Weick, *The social psychology of organizing*. McGraw-Hill, 1979.

[33] R. L. Daft and K. E. Weick, *Toward a Model of Organizations as Interpretation Systems*. Academy of Management Review, Vol. 9, 1984.

[34] G. Weiss, *Multiagent Systems*. The MIT Press, 1999.

[35] M. Wooldridge and N. R. Jennings, *Intelligent agents: Theory and practice*. The Knowledge Engineering Review, Vol. 10, 1995.

[36] N. R. Jennings and M. J. Wooldridge, *Applications of Intelligent Agents*. Agent Technology: Foundations, Applications, and Markets, Springer-Verlag, 1998.