Università degli Studi di Roma "Tor Vergata"

Facoltà di Lettere e Filosofia

Dottorato di Ricerca in Scienze Filosofiche e Sociali

Ciclo XXIV

Elaborato finale

Agenti (autonomi) normativi in società artificiali

Uno studio simulativo sull'emergenza di artefatti sociali

Marco Campennì

A.A. 2009/2010

Docente Guida/Tutor: Prof. Giovanni Iorio Giannoli

Coordinatore: Prof. Alessandro Ferrara

INDICE

Premessa

1. Introduzione
    1. Simulazione e Scienze Sociali: storia, problemi, prospettive
    2. Artefatti socio-culturali
    3. Le nuove scienze sociali
    4. Autonomia degli Agenti
        1. Economia artificiale
        2. Cultura artificiale
        3. Moralità artificiale
        4. Etica applicata
2. Le norme
    1. Il problema dell'autonomia normativa
    2. L'approccio convenzionalistico
    3. L'approccio logico-giuridico
    4. Rapporto fra convenzioni e norme
3. Ontologia Normativa (cognitiva)
4. Un approccio cognitivo allo studio delle norme: emergenza ed immergenza
    1. Generazione ed emergenza
    2. *Downward causation*
    3. Vantaggi del presente approccio
5. Realizzare un modello computazionale di agenti (autonomi) normativi
    1. Verso un'Architettura di Agente Normativo
    2. Innovazione Normativa
    3. EMIL-A: un'architettura normativa
    4. Simulare l'emergenza normativa
    5. Discussione dei risultati
6. Teoria dei giochi per lo studio delle convenzioni
    1. Introduzione
    2. *Replicator Dynamics*

3. Geometria dei giochi di popolazione
4. Il modello
5. Risultati delle Simulazioni
6. Risultati Analitici
7. Discussione dei risultati

7. Interiorizzare una norma
    1. Introduzione
    2. Dinamica degli Scopi
    3. Dinamica mentale delle norme
    4. Tipi e gradi di interiorizzazione
    5. Fattori
    6. L'interiorizzatore: un'architettura BDI
    7. Ipotesi e questioni supplementari
    8. Discussione dei risultati
8. Evoluzione della mente normativa
    1. *Evolutionary Game Theory*
    2. Antropologia Evolutiva
    3. Antropologia Cognitiva ed Approccio Epidemiologico
9. Considerazioni finali
10. Bibliografia

PREMESSA

Lo studio degli artefatti sociali, in particolare delle norme sociali, ha suscitato e suscita l'interesse di numerose e diverse comunità scientifiche.

Tradizionalmente, filosofi, sociologi, giuristi e psicologi si sono interessati allo studio dell'*origine* delle norme e ai *meccanismi socio-psicologici* coinvolti nella adozione o trasgressione delle stesse; tuttavia, nell'ultimo ventennio questo interesse si è notevolmente allargato, influenzando la ricerca di settori scientifici più "duri" quali l'intelligenza artificiale, lo studio dei sistemi multi-agente, la teoria dei giochi e la robotica.

Ognuno di questi ambiti disciplinari ha applicato il proprio approccio allo studio delle norme tentando di gettare luce sui meccanismi che sottendono:

i) *comportamento coordinato* che può emergere dall'adozione collettiva di una norma (o convenzione),

ii) al *processamento delle rappresentazioni mentali* relative a regole sociali (quali le credenze normative)

iii) all'influenza che il possesso o meno di norme può avere sul *comportamento degli individui*.

Ciascuno di questi ambiti ha iniziato più o meno recentemente ad occuparsi dello studio delle norme; di conseguenza, alcuni di essi si trovano in una fase più matura, altri sono ancora in una fase iniziale (come senza dubbio la robotica, che da questo punto di vista rappresenta l'ultima arrivata).

In questo lavoro, cercherò di mostrare come possa essere utile affrontare lo studio delle norme (di come esse si affermano socialmente, di come esse si insediano nella mente degli agenti, di come esse incidono sui comportamenti), utilizzando un approccio che per certi versi potremmo definire *ibrido*. Cercherò di conciliare due tradizioni contrapposte, provando a trarre il meglio da entrambe: la tradizione dell'approccio legato alla *game theory*, da un lato, e quello legato ai *sistemi multi-agente*, dall'altro.

Nel capitolo introduttivo di questo lavoro presenterò il contesto teorico-sperimentale di riferimento: quello dell'utilizzo della *simulazione* nell'ambito delle scienze sociali

(essenzialmente, l'uso di un *software* che gira su un calcolatore, fornendo un modello artificiale di un fenomeno reale); in particolare, cercherò di ricordare un aspetto essenziale del fertile connubio fra scienze sociali e scienze cognitive.

Nel capitolo successivo, presenterò i due approcci esistenti, quello che utilizza la *teoria dei giochi* e quello che adotta i *sistemi multi-agente*. Questi due approcci sono concepiti tradizionalmente come contrapposti ed alternativi, sia perché fanno uso di strumenti molto diversi, sia perché sono interessati ad aspetti diversi.

Il primo, quello che opera nel contesto della teoria dei giochi, usufruisce di una importante tradizione analitica e può sfruttare i numerosi vantaggi di un approccio matematico (per esempio, gli strumenti di previsione garantiti da un modello analitico, insieme all'analisi della qualità e robustezza dei risultati ottenuti). In questo caso, l'interesse principale è rivolto al modo in cui una norma (sociale) può emergere in un gruppo di individui; in questo contesto, non esiste una vera e propria differenza fra norma e convenzione (quest'ultima, nelle scienze socio-economiche può essere intesa come *soluzione ad un problema di coordinamento*); sempre in questo contesto, gli individui (o "agenti" come si è soliti dire in gergo tecnico) non processano *mentalmente* le norme, ma queste rappresentano sostanzialmente gli stati di equilibrio più o meno stabili che possono essere raggiunti nell'interazione fra strategie o comportamenti differenti. L'approccio game-teorico non è interessato agli aspetti strettamente cognitivi relativi all'adozione e all'uso delle norme e d'altro canto non potrebbe essere altrimenti, non disponendo degli strumenti concettuali necessari allo scopo. I tradizionali ambiti disciplinari di riferimento di tale approccio sono le scienze sociali, la matematica, la filosofia (soprattutto quella morale).

Il secondo approccio, quello che utilizza i sistemi multi-agente, opera in un contesto decisamente differente; le discipline di riferimento sono in questo caso l'intelligenza artificiale classica, la logica e le scienze giuridiche. Nel caso dei sistemi multi-agente le norme sono oggetti mentali *già acquisiti*, *built-in*, che influenzano in diversa misura il comportamento e le decisioni degli agenti. L'interesse non è rivolto in questo caso a come gli agenti possano formarsi nuove credenze normative; d'altro canto, gli strumenti formali di cui si dispone, essenzialmente quelli della logica, possono essere estremamente utili per trarre inferenze e definire l'ordine (logico) delle azioni da eseguire, ma poco utili per affrontare meccanismi

mentali dinamici che non operino esclusivamente su basi di conoscenze date ma facciano i conti con *input* nuovi che giungono *runtime*.

Fino ad oggi, queste due prospettive sono state considerate inconciliabili e scarse sono state le occasioni di confronto e contaminazione fra le due. In questo lavoro, presenterò un terzo approccio, che si potrebbe definire in un certo senso *multi-agente cognitivo*: attingerò al robusto contributo teorico concernente gli agenti (cognitivi) normativi, sviluppato nell'arco degli ultimi quindici anni.

Dopo aver discusso l'ontologia relativa ai concetti fondamentali che verranno utilizzati nel corso del lavoro, cercherò di chiarire il complesso rapporto che esiste fra i concetti di *emergenza* ed *immergenza*.

Se il concetto di emergenza è un concetto ben consolidato e fa riferimento esattamente alla stessa classe di fenomeni cui si riferisce la teoria dei giochi (per la quale un fenomeno può *emergere* globalmente grazie alle micro interazioni locali degli elementi che costituiscono il sistema), il concetto di immergenza è decisamente nuovo e fa riferimento al processo dinamico grazie al quale, ad esempio, nuove credenze possono immergersi nelle menti degli agenti, andando ad arricchire il loro repertorio di rappresentazioni mentali.

In questo senso, i due concetti di emergenza ed immergenza non vengono visti soltanto come complementari e caratterizzati da direzioni opposte (rispettivamente, dal micro al macro e dal macro al micro), ma sono anche necessari entrambi per spiegare il processo attraverso il quale gli agenti possano formarsi nuove credenze (normative), agire di conseguenza e generare attraverso l'interazione con gli altri l'emergenza di norme sociali.

Un punto cruciale rispetto alle assunzioni teoriche alla base di questo lavoro è che *le norme sono artefatti ben diversi dalle convenzioni.* Se le ultime, seguendo la tradizione della teoria dei giochi, sono essenzialmente *soluzioni a problemi di coordinamento* (e perciò sono artefatti sociali *esterni* agli agenti ed alle loro menti), le prime presentano per così dire due facce: una *privata*, relativa alla vita mentale degli agenti (per cui esistono nella mente dell'agente rappresentazioni mentali - quali credenze e scopi normativi - che si riferiscono ad una norma) ed una pubblica o *sociale* (per cui una norma può essere trasmessa ad altri agenti attraverso la comunicazione - esplicita e verbale o implicita e comportamentale - di messaggi normativi).

A questo punto del nostro lavoro diventerà necessario chiarire quali sono i requisiti che una architettura di agente normativo deve possedere per poter:

i) processare rappresentazioni mentali quali credenze e scopi normativi,
ii) comunicare agli altri agenti il contenuto di tali rappresentazioni,
iii) osservare il comportamento altrui, in modo da poter effettuare inferenze normative.

Nel capitolo 5 presenterò dunque alcuni risultati sperimentali ottenuti testando quali effetti possa avere l'utilizzo di un *modulo normativo*: confronterò cioè batterie di simulazioni diverse in cui, da una parte, gli agenti sono dotati di un *modulo di riconoscimento delle norme* (cioè sono in grado di formarsi nuove credenze normative ed inferire l'esistenza di norme sociali a partire dall'osservazione del comportamento altrui); dall'altra invece, gli agenti saranno essenzialmente *imitatori sociali* (cioè decidono quale azione compiere, imitando un certo numero di vicini ed eseguendo l'azione che viene maggiormente eseguita).

Lo scenario è qui rappresentato da un mondo in cui gli agenti interagiscono scambiandosi messaggi oppure osservando il comportamento altrui, convergendo in maniera più o meno stabile su una fra diverse azioni possibili.

Osserverò le differenze che si rilevano a livello della popolazione quando si adotta una tipologia di agenti o l'altra; mostrerò l'opportunità intrinseca di una architettura cognitiva, se si vuole implementare un modello di *obbedienza intelligente*.

Nel capitolo successivo presenterò un modello game-teorico relativo al seguente nodo concettuale: l'*affermazione sociale* di una fra diverse possibili convenzioni/strategie. L'idea alla base di questo capitolo è che se si vuole realizzare un modello computazionale dell'emergenza di una convenzione (e non di una norma), l'approccio game-teorico risulta estremamente efficace, in quanto permette un utile confronto tra modelli analitici.

A differenza di quanto accade nei tradizionali modelli di emergenza di convenzioni (nel quadro della teoria dei giochi), nel modello qui adottato i *payoff* (cioè i guadagni che gli agenti ottengono nell'adottare un comportamento piuttosto che un altro) non sono stabiliti a priori (in base ad una matrice dei *payoff* statica), ma vengono calcolati *runtime*, in un processo imitativo della strategia comportamentale che permette un guadagno superiore. Farò

vedere quali e quanti stati di equilibrio possono essere raggiunti in un *congestion game*, confrontando i risultati simulativi con quelli ottenuti usando differenti modelli analitici (essenzialmente matematici o geometrici).

Nel capitolo 7 avanzerò alcune ipotesi in merito ai meccanismi di *interiorizzazione delle norme*. Mi riferirò, da una parte, agli automatismi che possono esistere nell'esecuzione di un particolare comportamento (immaginiamo, per esempio, l'istintivo gesto di frenare quando siamo in macchina, non appena il semaforo diventa rosso; oppure l'automatico gesto di coprirsi la bocca con la mano quando sbadigliamo); questi automatismi non prevedono un processamento mentale completo degli *input* che riceviamo dall'ambiente, da altri o da noi stessi. Dall'altra, mi riferirò al fatto che alcune credenze normative possono radicarsi in noi al punto tale da raggiungere una salienza così elevata da trasformarsi in *scopi* (per altro non necessariamente normativi: pensiamo per esempio alla norma sociale di "non fare male ad altri"; probabilmente, essa nasce evolutivamente come una norma sociale per diventare poi uno scopo non-normativo).

Infine, nell'ultimo capitolo cercherò di analizzare le ipotesi esistenti in merito all'origine degli oggetti culturali ed in particolare delle norme (sociali); passerò in rassegna tre grandi famiglie di approcci: quello legato alla *teoria dei giochi evolutiva*, quello relativo alla *antropologia evolutiva* ed infine quello concernente l'*antropologia cognitiva* e la *prospettiva epidemiologica* (con uno sguardo anche al ruolo giocato dalle emozioni).

Questo lavoro è il frutto di tre anni di ricerche svolte presso il Laboratorio di Simulazione Sociale Basata su Agente (LABSS) dell'Istituto di Scienze e Tecnologie della Cognizione (ISTC) di Roma, nell'ambito del progetto europeo EMIL (EMIL: *Emergence In the Loop* – FP6 EC Contract No. 033841).

I risultati di queste ricerche sono stati già oggetto di diverse pubblicazioni su riviste specialistiche, alle quali io stesso ho potuto dare il mio contributo[1].

---

[1] Si vedano in particolare:

2010

- Campennì, M. (Eds.) (2010) "Emergence of Social Norms in Artificial Societies", Special Issue of *International Journal of Agent Technologies and Systems (IJATS)*, 2 (1), IGI Global.

Sono grato ai colleghi Giulia Andrighetto, Federico Cecconi, Gennaro di Tosto, Francesca Giardini, Federica Mattei, Mario Paolucci, Stefano Picascia, Walter Quattrociocchi ed in particolare alla Dott.ssa Rosaria Conte, direttrice del LABSS, per l'opportunità che mi hanno dato in questi anni e per gli insegnamenti dei quali ho potuto fare tesoro; in particolare, li ringrazio per le numerose occasioni di confronto e discussione che in questi anni hanno caratterizzato l'attività del laboratorio. Ovviamente, il lavoro di revisione, integrazione, elaborazione del materiale che segue ricade esclusivamente sotto la mia responsabilità.

- Cecconi, F., Campennì, M., Andrighetto, G., Conte, R. (2010) "What Do Agent-Based and Equation-Based Modeling Tell Us About Social Conventions: The Clash Between ABM and EBM in a Congestion Game Framework", *Journal of Artificial Societies and Social Simulation (JASSS)*, 13, (1), 6.
- Conte, R., Andrighetto, G., Campennì, M. (2010) "Internalizing Norms. A cognitive model of (social) norms' internalization", *International Journal of Agent Technologies and Systems (IJATS)*, 2 (1), pp. 63-73, IGI Global.
- Campennì, M., Cecconi, F., Andrighetto, G., Conte, R. (2010) "Norm and Social Compliance. A Computational Study", *International Journal of Agent Technologies and Systems (IJATS)*, 2 (1), 50-62, IGI Global.
- Campennì, M. (2010) "Normative Multi Agent Systems and Normative Architectures: The Emergence of Norms in Artificial Societies", *International Journal of Agent Technologies and Systems (IJATS)*, 2 (1), pp. i-iii, IGI Global.

2009

- Campennì, M., Andrighetto, G., Cecconi, F., Conte, R. (2009) "Normal = Normative? The Role of Intelligent Agents in Norm Innovation", *Mind & Society*, Vol. 8, No. 2, pp. 153-172, Springer Berlin / Heidelberg.
- Conte, R., Andrighetto, G., Campennì, M. (2009) "The Immergence of Norms in Agent Worlds", *Lecture Notes in Artificial Intelligence, LNAI*, Vol. 5881 – H. Aldewereld, V. Dignum, G. Picard (Eds.).
- Andrighetto, G., Campennì, M., Cecconi, F., Conte, R. (in press) "The Complex Loop of Norm Emergence: a Simulation Model", In K. Takadama, C. C. Revilla, G. Deffuant (Eds.) *The Second World Congress on Social Simulation*, Springer-Verlag LNAI.

2008

- Campennì, M.,Andrighetto, G., Cecconi, F., Conte, R. (2008) "Normal = Normative? The Role of Intelligent Agents in Norm Innovation", *The Fifth Conference of the European Social Simulation Association (ESSA 2008)*, Brescia, Italy, September 1-5.
- Andrighetto, G., Campennì, M., Cecconi, F., Conte, R. (2008) "What do Agent-Based and Equation-Based Modelling tell us about Social Conventions", *III Edition of Epistemological Perspectives on Simulation; A Cross-Disciplinary Workshop*, OCTOBER 2 - 3 Lisbon, Portugal.
- Andrighetto, G., Campennì, M., Cecconi, F., Conte, R. (2008) "How Agents Find out Norms: A Simulation Based Model of Norm Innovation", *3rd International Workshop on Normative Multiagent Systems (NorMAS 2008)*, Luxembourg 15-16 July.
- Cecconi, F.,Andrighetto, G., Campennì, M.,Zappacosta, S. (2008) "On the Emergence of Conventions: a Comparison between a Simulative and an Analytical Approach", *International Conference on Economic Science with Heterogeneous Interacting Agents (ESHIA/WEHIA 2008)* Faculty of Physics, Warsaw University of Technology, 19-21 June.
- Andrighetto, G., Campennì, M., Cecconi, F., Conte, R. (2008) "Conformity in Multiple

1. INTRODUZIONE

1.1 Simulazione e Scienze Sociali: storia, problemi, prospettive

Verso la metà degli anni '70 la *Alfred Sloan Foundation* sponsorizzò una serie di conferenze per esplorare una nuova impresa accademica che sotto l'etichetta di *Scienze Cognitive* sintetizzava diverse tendenze scientifiche: la grammatica generativa, la cibernetica, la teoria dell'informazione, le reti neurali, l'intelligenza artificiale e più in generale l'approccio computazionale allo stusio delle menti. La Fondazione stanziò una ingente somma per il finanziamento di grandi programmi di ricerca, contribuendo di fatto alla nascita di un nuovo settore d'indagine scientifica, fortemente interdisciplinare.

Obbiettivo comune a tutte le discipline che confluirono nel programma era quello di indagare le capacità rappresentazionali e computazionali della mente e la loro implementazione strutturale e funzionale nel cervello. L'insieme delle discipline fu rappresentato come un esagono ai cui vertici stavano filosofia, linguistica, neuroscienze, IA, psicologia e antropologia (vedi figura 1).

Contexts: Imitation Vs Norm Recognition", *WCSS-2008 World Congress on Social Simulation 2008 (WCSS-08)*, George Mason University, Fairfax - July 14-17.

2007

- Andrighetto, G., Campennì, M., Conte, R., Paolucci, M. (2007) "On the Immergence of Norms: a Normative Agent Architecture", *Proceedings of AAAI SYMPOSIUM, SOCIAL AND ORGANIZATIONAL ASPECTS OF INTELLIGENCE*, november 8 – 11, WASHINGTON DC, Usa.
- Conte, R., Andrighetto, G., Campennì, M., Paolucci, M. (2007) "Emergent and Immergent Effects in Complex Social Systems", *Proceedings of AAAI SYMPOSIUM, SOCIAL AND ORGANIZATIONAL ASPECTS OF INTELLIGENCE*, november 8 - 11, WASHINGTON DC, Usa.
- Campennì, M., Cecconi, F., Andrighetto, G. (2007) "Uno Studio sull'Emergenza delle Convenzioni", *WORKSHOP AISC (Associazione Italiana Scienze Cognitive) "Cognizione, Complessità, Cittadinanza"* 27-29 Novembre, Roma.
- Andrighetto, G., Campennì, M., Conte, R. (2007) "EMIL-M: MODELS OF NORMS EMERGENCE, NORMS IMMERGENCE AND THE 2-WAY DYNAMIC", *Technical Report*, LABSS-ISTC/CNR, 00507. pp. 1-50.

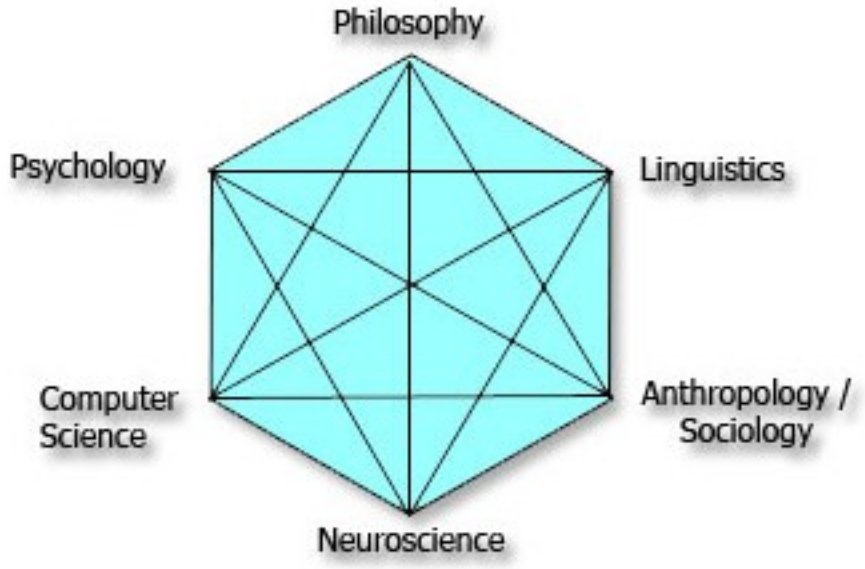


Figura 1. Esagono Cognitivo delle Scienze Cognitive

Nel concreto, nei dipartimenti universitari la nuova disciplina assunse la forma di una particolare coalizione fra due o più discipline che costituivano l'esagono cognitivo; ma in tutte le combinazione scaturite tuttavia restava sempre centrale l'approccio spiccatamente computazionale.

Soltanto in due dei maggiori programmi accademici di scienza cognitiva comparve un approccio "culturale": a Berkeley (in collaborazione con Stanford) e presso l'Università della California a San Diego (UCSD). In questi due casi andò consolidandosi negli anni l'influenza della scienza cognitiva sulle ricerche in ambito culturale ed etnografico. Questo particolare ambito ha utilizzato in particolare la così detta teoria della cognizione (e dell'azione) *situata in contesto* o, come dicono alcuni, la *embodied cognition*.

Dal 1978, presso un laboratorio dell'UCSD in cui l'oggetto prevalente di studio è il funzionamento della cognizione in ambienti socio-culturali organizzati; possiamo dire che in quella università si è realizzato il primo incontro interdisciplinare fra la scienza cognitiva e le scienze sociali (cioè fra psicologi, antropologi e sociologi che vedevano tutti la cultura ed il contesto come elementi costituenti della cognizione).

Fra i primi risultati di questo approccio va ricordato l'articolo (*Twelve issues for Cognitive Science*) di Don Norman, apparso nel 1980 sulla rivista ufficiale della nuova società

di scienza cognitiva ("Cognitive Science" - Norman 1980), nel quale si dava espressamente conto del ruolo della cultura e della società nella cognizione. Fu proprio a seguito di questa impostazione che nel 1993 Don Norman pubblicò *Things That Make Us Smart*, lavoro in cui egli propose la sua teoria della "cognizione distribuita" (Norma 1993). In questa teoria, la nozione di *artefatto* svolge un ruolo fondamentale; si insiste sul fatto che una scienza generale della mente è inevitabilmente anche una scienza della socialità e della cultura. Va da sé che la rappresentazione dei diversi problemi sociali è considerata un evento mentale; le operazioni cognitive di base ed il modo in cui noi le utilizziamo nel giudicare, decidere, agire, ragionare, scegliere, influenzare e persuadere hanno effetti determinanti sulle questioni più importanti di cui si occupano le scienze sociali (Turner 2001).

Tuttavia, come alcuni hanno osservato, la convergenza fra scienza cognitiva e scienze sociali è stata certamente minore di quel che ci si sarebbe potuti aspettare. Seppure il rapporto con la scienza cognitiva è oggi un aspetto della *routine* nelle scienze sociali, lo studio della cognizione non rientra nella formazione curricolare degli studenti di economia, scienze politiche o sociologia.

E' del resto nota la vitalità di settori di ricerca interdisciplinari come la psicologia evoluzionistica, le teorie dell'evoluzione dell'intelligenza basate sulla comunicazione e sull'apprendimento sociale (Dennett 1995, Cosmides e Tooby 1992, Plotkin 1995), oppure l'area della *social cognition*, o quella dei *modelli mentali*; meno noti sono invece altri settori di ricerca in ambiti contigui, come per esempio le *teorie thick* (spesse) della razionalità (Ostrom 1998).

1.2 Artefatti Socio-Culturali

Per una grande famiglia di teorie che prendono le mosse dalla *activity theory* (e quindi – come notato da Cole - risalgono a Luria, Leontiev, Dewey e Mead), la "mediazione dell'azione umana ad opera degli artefatti" è " il momento centrale della costituzione della mente umana" (Cole 1997).

L'artefatto in senso stretto è un oggetto materiale modificato lungo l'arco della sua incorporazione nella azione *goal-directed*. In un senso più ampio, gli artefatti sono simultaneamente ideali, concettuali e materiali. Per esempio, la mente è certamente *artifact-mediated.*

Tutto il filone più ufficiale e rappresentativo della scienza cognitiva, quello che non considera le implicazioni socio-culturali della cognizione, si (auto)confina nel cervello dimenticando gli artefatti, il loro potere rispetto alla cognizione ed il loro carattere simbolico.

Simon (1981) già considerava i *sistemi simbolici* come artefatti alla quintessenza. Secondo questo autore sono quattro le caratteristiche degli artefatti:

1. sono creati dall'uomo (non necessariamente in modo cosciente);
2. possono imitare il reale;
3. possono essere caratterizzati in termini di funzioni, scopi e adattamento;
4. si presentano sia sotto forma di imperativi che di descrizioni.

Tutte queste caratteristiche mostrano la contiguità fra il concetto di artefatto di Simon e quello che caratterizza il così detto *modello standard delle scienze sociali* (quello che risulta appunto dai contributi di Luria, Leontiev, Dewey, Mead ed altri), e quindi giustificano il parlare della cognizione come di un sistema "dentro e fuori" la mente.

Gli artefatti cognitivi sono stati definiti (Norman 1994) come quelli che svolgono un ruolo nella cognizione dell'individuo. Ed Hutchins (1995) vede gli artefatti cognitivi come *costitutivi del pensiero*; sono quelli che interagiscono con strutture interne e permettono, facilitano, modulano il pensiero stesso.

Tuttavia, esiste una seconda classe di artefatti che costituisce un mondo relativamente autonomo; qui, regole e convenzioni non appaiono nel loro diretto valore pratico, ma dirigono comunque l'attenzione e la rappresentazione del mondo da parte dei sistemi cognitivi (vedi le caratteristiche 3 e 4 definite da Simon). Questa differenza si riflette nell'uso linguistico che accompagna le due classi di artefatti. Un artefatto del primo tipo inutilizzato nel contesto previsto *non* viene violato (si pensi ad una cartina stradale non consultata). Una norma operativa ma inapplicata *è* una norma violata. Le norme hanno la proprietà di agire indipendentemente dall'azione e dagli scopi dell'utente (proprietà condivisa con una sottoclasse di artefatti del primo tipo, ossia le tecnologie ad "agente" intelligente ed autonomo). In più hanno la facoltà di dirigere l'azione dell'utente, al di là e al di sopra della sua stessa volontà.

La teoria degli artefatti è ancora lontana da una forma completa e conclusa, come del resto lo è anche quella della cognizione distribuita. Rimangono ancora poco chiari alcuni aspetti cruciali: cosa si intende esattamente quando si afferma che una cartina stradale

costituisce (fa parte della) cognizione? Significa qualcosa di diverso da quello che succede quando leggiamo un libro? In tutte e due i casi si tratta di artefatti inerti, che vengono utilizzati dall'utente come fonte o supporto della propria conoscenza. Attualmente, tuttavia, le tecnologie intelligenti forniscono anche supporti non inerti, bensì proattivi o addirittura autonomi (si pensi ai sistemi ad agente). Quando è che un artefatto costituisce un mero supporto e quando è invece che esso sia *estensione* se non addirittura *costituente* della cognizione?

Gli artefatti possono certamente assolvere funzioni estremamente diverse; tali funzioni si esprimono attraverso un diverso rapporto con la cognizione individuale.

Di solito, siamo abituati a pensare agli artefatti come a strumenti più o meno attivi ed autonomi della cognizione individuale (mezzi di trasporto della conoscenza, calcolatori, software intelligenti). In questo caso, l'artefatto facilita l'acquisizione e la trasmissione di conoscenza, l'attività di ragionamento, di *problem-solving* e di pianificazione dell'utente: è un *artefatto di supporto*.

Esistono però anche artefatti che governano l'azione degli individui, che non sono più utenti, ma destinatari e forse beneficiari; tali artefatti influenzano le credenze e dirigono la volontà, gli scopi, gli affetti, se non addirittura le emozioni degli individui. Questi artefatti sono *artefatti di controllo*.

Di questa seconda categoria fanno parte: le norme, i regolamenti, le istituzioni (ed i relativi ruoli). L'esecutore naturale di un ruolo istituzionale ci offre un interessante esempio di rovesciamento del tipico rapporto fra individui ed artefatti; se l'artefatto di supporto assiste l'individuo, l'esecutore del ruolo assiste l'artefatto di controllo, garantendone la funzione. Gli artefatti di supporto estendono e facilitano la cognizione individuale; quelli di controllo si servono della cognizione, subordinandola alla soluzione di problemi e a fini esterni alla mente individuale. In entrambi i casi (ma soprattutto nel secondo - artefatti di controllo), l'individuo (co)opera in un sistema complesso di *decision-making* e *problem-solving*.

La scienza cognitiva che ignora la componente socio-culturale non è in grado di affrontare la complessa interconnessione fra la cognizione che opera a livello individuale e gli artefatti (di varia natura e funzione: materiali ed immateriali, di supporto e di controllo): la versione neuroscientifica della scienze cognitiva (che confina il pensiero nel cervello) non riesce a dar conto del ruolo degli artefatti; d'altronde, la versione socio-culturale non può

limitarsi a contestualizzare la mente: essa deve considerare le diverse classi di artefatti come co-costituenti della mente (anche quando gli artefatti considerati siano esterni al cervello) e costruire modelli e teorie del sostrato naturale e artefattuale della cognizione.

1.3 Le nuove scienze sociali

C'è una tendenza emergente nelle scienze sociali che consiste nel rivedere la teoria della razionalità ed elaborare nuovi modelli alla luce delle critiche cognitive. Si punta ad una teoria *thick* della razionalità, che si fonda su concetti quali: capitale sociale, norme sociali, reciprocità, fiducia e reputazione.

Le critiche cognitive alla teoria della razionalità non si limitano a sottolineare il ruolo dell'apprendimento o dell'incertezza e incompletezza delle informazioni, ma riguardano anche gli aspetti architetturali e le assunzioni di base della teoria dei giochi, in particolare la natura e la dinamica delle preferenze, la massimizzazione dell'utilità attesa e la supposta centralità dell'individuo. Secondo questo approccio, il nuovo modello della razionalità dovrebbe puntare all'integrazione della componente normativo-affettiva nella razionalità e all'abbandono della centralità dell'individuo.

Molti studiosi della razionalità riflettono oggi sugli aspetti socio-cognitivi; numerosi esperimenti sembrano mettere in luce una diffusa tendenza alla *reciprocità*. Questo fenomeno è stato interpretato da una parte come una *forma di imitazione*, dall'altra come il *risultato di una norma sociale*, la “norma di reciprocazione”. Ad esempio, pagando salari generosi le ditte potrebbero indurre i propri lavoratori a prestazioni migliori rispetto a quelle che ci si potrebbe aspettare sulla base del mero contratto (problema dei contratti imperfetti).

Studi cognitivi sul ragionamento ed il *decision-making*, condotti in psicologia evoluzionistica mostrano il carattere *problem-driven* delle capacità cognitive, la loro natura adattativa, mettendo in luce in questo modo alcune assunzioni problematiche della teoria della razionalità classica. La capacità di ragionamento sembra dipendere dal contesto di applicazione ed è probabilmente evoluta per manipolare norme sociali, piuttosto che regole logiche (Cosmides e Tooby 1992).

Elinor Ostrom (1998) ha analizzato molto attentamente, da una parte, le teorie classiche della razionalità (che ella definisce *thin*, cioè sottili) e dall'altra, gli studi sperimentali di psicologia cognitiva condotti per verificare le assunzioni della razionalità classica. Dal suo

studio emerge un fatto interessante: gli esseri umani sarebbero molto più cooperativi di quello che la teoria *thin* della razionalità si aspettava. Il comportamento cooperativo risulta molto sensibile alle variazioni del contesto sociale, delle condizioni di interazione, del mezzo di interazione e della natura del compito.

Emerge anche un'insufficienza della psicologia evoluzionistica; la maggior attitudine a manipolare norme sociali (piuttosto che regole logiche) spiega la capacità dei soggetti di individuare le trasgressioni ma non riesce a spiegare la volontà di farlo né quella di aderire alle norme.

Boles e Gintis (2001) hanno utilizzato modelli simulativi per indagare la nozione di *strong reciprocity* (reciprocità forte), relativa sia alla restituzione di azioni altruistiche, sia che più in generale a qualunque attitudine cooperativa (inclusa l'osservanza delle norme utili al gruppo, che mette in difficoltà le teorie *thick* della razionalità).

Un paradigma razionale arricchito dall'attitudine alla *strong reciprocity* (e che consideri la natura adattativa del ragionamento e la pressione evolutiva a manipolare regole sociali) potrebbe sembrare una prima soluzione al problema della razionalità; tuttavia, esistono strutture cognitive evolute nel tempo più complesse, rispetto alla semplice attitudine alla reciprocazione; per esempio:

1. la capacità di riconoscere le norme sociali;
2. la volontà di punire i trasgressori (l'aggressione moralistica);
3. l'indignazione;
4. il senso di colpa ed altre emozioni sociali.

Oltretutto, le istituzioni sociali non rappresentano soltanto un'evoluzione della norma di reciprocità. Dati naturali ed artificiali (simulativi) dimostrano l'insufficienza delle sole strutture cognitive individuali nella spiegazione dei fenomeni istituzionali.

Nello specifico, la ricerca antropologica (Dow 1997) ha messo in luce il ruolo dello *story-telling* e del *gossip* nella formazione di gruppi sociali coesi fra gli ominidi del centro Africa, i quali hanno mantenuto la coesione in successive migrazioni, aumentando in questo modo la loro probabilità di sopravvivenza e riproduzione. Addirittura, alcuni studiosi fanno risalire al *gossip* (e alla sua funzione stabilizzatrice rispetto alle relazioni sociali) la nascita del linguaggio nella specie umana (Dunbar 1997). Va detto che il *gossip* non trasmette necessariamente valutazioni accurate su persone o gruppi; la sua efficacia rispetto alla

coesione del gruppo o alla diffusione di norme e regole sociali dipende dalla velocità e stabilità di trasmissione (non dalla corrispondenza delle valutazioni con la realtà).

1.4 Autonomia degli Agenti

Nell'ambito cui abbiamo fatto riferimento, è emerso chiaramente negli ultimi anni che il protagonista indiscusso delle linee di ricerca più promettenti è l'*Agent-Based Modeling* (ABM); in particolare, una versione del modello che tenga conto di alcune proprietà particolari degli agenti:

1. adattatività, flessibilità e capacità di apprendimento;
2. eterogeneità e versatilità;
3. socialità (capacità di rispondere al comportamento altrui ma anche capacità di manipolare la mente altrui);
4. capacità di manipolare artefatti socio-culturali;
5. autonomia.

Il motivo di tanta centralità dell'ABM risiede nella crescente importanza dell'approccio compuazionale e simulativo in numerosi settori delle scienze sociali, dai *social networks* allo studio della cooperazione, dell'altruismo e della reciprocità, della cognizione distribuita, allo studio delle organizzazioni, dall'archeologia e dalla storia evolutiva della cultura e dei sistemi sociali all'etica e allo studio delle istituzioni.

Lo sviluppo di piattaforme simulative basate su agenti per lo studio dei fenomeni sociali ha reso possibile l'elaborazione di nuove teorie sociali, culturali ed etiche, ma ha anche inaugurato nuovi settori di ricerca che si basano sull'intersezione fra scienze sociali, scienze cognitive e scienze dell'artificiale.

Il settore di studio delle società artificiali è relativamente giovane, ma già estremamente produttivo. Nasce ufficialmente nella comunità socio-scientifica europea nei primi anni '90 (Gilbert e Doran 1994, Gilbert e Conte 1995); si istituzionalizza poi nel *Journal of Artificial Societies and Social Simulation*. Negli stessi anni, negli Stati Uniti guadagna immediatamente popolarità grazie al inguaggio SWARM e alla piattaforma *SugarScape* (Epstein e Axtell 1996). Sul finire degli stessi anni '90 il neonato filone di ricerca si specializza in nuovi sotto-settori come l'*artificial economy*, l'*artificial culture*, l'*artificial morality*.

1.4.1 Economia artificiale.

In questo settore di ricerca si è sviluppato in particolare lo studio computazionale di strategie che competono fra loro, nell'acquistare e vendere merci.

L'economia artificiale mira alla costruzione di mercati artificiali con un elevato numero di operatori, allo scopo di indagare cruciali questioni economiche relative alla fluttuazione o equilibrio dei prezzi, come:

1. l'effetto del tipo di setting sui prezzi e sulla stabilità del mercato,
2. l'influenza degli speculatori (cioè di coloro che non producono ma comprano e vendono),
3. la competizione fra diversi tipi di speculatori,
4. la stabilità delle strategie nel tempo
5. i diversi meccanismi di imitazione.

Realizzare sistemi di questo tipo richiede l'utilizzo di agenti eterogenei dotati di capacità di adattamento e apprendimento, comunicazione, previsione degli esiti delle proprie ed altrui azioni, versatilità (ossia capacità di scegliere fra strategie differenti e concorrenti) nonchè capacità di ingannare e bluffare.

1.4.2 Cultura artificiale

Questo nuovo settore pretende di rappresentare una vera e propria "nuova epistemologia" (Gessler 2002).

Nei modelli computazionali dell'evoluzione culturale si tenta di trasferire in linguaggio macchina le teorie più accreditate in questo settore. Proprio come accade negli esperimenti mentali, possono essere generati diversi percorsi o scenari controfattuali, esaminando e controllando (almeno in linea di principio) l'intera gamma di variabili in ingresso ed i risultati che ne seguono. Questi ipotetici mondi permettono allo studioso di indagare le diverse ipotesi di spiegazione di un fenomeno, utilizzando modelli oggettivi, accessibili e rivedibili.

In questi modelli, gli agenti non sono solo eterogenei, adattivi e capaci di apprendimento, ma manipolano e veicolano rappresentazioni sociali e culturali, costruiscono, utilizzano e trasmettono artefatti culturali.

1.4.3 Moralità artificiale

Nel filone chiamato “moralità artificiale” convergono le teorie etiche dell'altruismo, della cooperazione e della reciprocità, implementate su piattaforme simulative ad agente (o utilizzate nella realizzazione di robot virtuosi)(Danielson 1992, 1998). Questo tipo di ricerca punta allo sviluppo di modelli computazionali evolutivi e basati su agenti per la soluzione di problemi di natura etica e per la progettazione di sistemi etici (quali: adattare le convenzioni, le norme e più in generale gli artefatti di controllo al cambiamento tecnologico; per far evolvere comportamenti morali, altruismo e cooperazione).

1.4.4 Etica applicata

Le tecnologie cognitive trovano come si sa larga applicazione in numerosi settori produttivi, dalla comunicazione al commercio, dall'animazione ed intrattenimento fino all'educazione. Un settore molto promettente è la *tecnologia etica*, ossia la produzione di strumenti di supporto all'azione orientata in senso etico.

Questo tipo di tecnologia si basa sempre più spesso sui dati e sulle teorie elaborati dall'etica o moralità artificiale (vedi 1.4.3) ed è orientata allo sviluppo di sistemi computazionali che esibiscano un comportamento etico, come i *robot virtuosi*, gli *agenti etici*, gli *attori sintetici morali*.

L'etica applicata all'impresa, alle tecnologie o al *policy making* è un settore in forte espansione. In generale, potremmo definire questioni di etica applicata tutte quelle in cui si fa riferimento a scelte coerenti con certi standard etici, nei vari settori d'interesse.

Nei centri di etica applicata (per lo più negli Stati Uniti e nei paesi scandinavi) vengono finanziati programmi di ricerca che estendono i fondamenti teorici dell'analisi delle decisioni, accrescendone l'efficacia pratica ed espandendone il raggio di applicazione, per valutare i problemi morali incontrati nell'applicazione e nel consumo di nuovi strumenti e tecnologie.

Diversamente dalle classiche ricerche di Axelrod sull'evoluzione dei comportamenti cooperativi, i sistemi etici puntano ad un crescente livello di adeguatezza sul piano cognitivo, indispensabile per ottenere comportamenti virtuosi flessibili, per la reciprocità indiretta e non basata sull'interazione ripetuta, per risolvere eventuali incongruenze fra norme diverse, per valutare la moralità del comportamento altrui, per evitare i vizi morali di cui sono vittime gli

agenti razionali (come il *moral race*, il punire i partner meno morali di me).

Quelli citati sono solo alcuni dei più importanti filoni della ricerca socio-cognitiva attuale e non rappresentano certamente una sintesi completa ed esaustiva del panorama esistente, costituito da numerose "contaminazioni" trasversali. Bastino, tuttavia, per fornire un quadro di massima delle potenzialità e dell'utilità che la contaminazione reale fra scienze sociali e scienza cognitiva presenta, quando sia realizzata attraverso l'utilizzo dei modelli computazionali.

2. LE NORME

La simulazione sociale permette l'uso di modelli computazionali utili per (ri)creare e studiare aspetti essenziali delle società naturali (umane e non umane) e artificiali. Oramai, coinvolge un numero di discipline sempre maggiore: la sociologia, l'economia, l'antropologia, l'intelligenza artificiale (distribuita e non), la vita artificiale, i sistemi multi-agente, la scienza cognitiva, l'informatica, la psicologia sociale e la biologia, solo per citarne alcune.

La simulazione basata su agenti, come ricordato precedentemente (nel paragrafo 1.4), prevede che gli agenti siano dotati di certe caratteristiche ben precise. La mediazione cognitiva degli agenti e la loro capacità di formarsi rappresentazioni, di ragionare e di stabilire decisioni in modo autonomo, risulta di estrema importanza nella produzione di un comportamento conforme alle norme.

Diversi autori hanno proposto classificazioni del comportamento morale (Piaget 1972 e Kohlberg 1971). Questi studi presentano una lettura morale dello sviluppo del comportamento, ossia del progresso individuale da azioni meno ad azioni più morali.

Il nostro approccio punta invece a fornire una analisi cognitiva del comportamento normativo; non analizza i vari tipi di obbedienza in base al grado di moralità e di autonomia del comportamento in esame, quanto in base alla varietà dei meccanismi cognitivi coinvolti caso per caso, allo scopo di mostrare la complessità cognitiva e il grado di autonomia che ciascuno di essi comporta (anche quelli moralmente meno apprezzabili).

La differenza fra un tipo di approccio ed un altro non risiede dunque nell'autonomia; anche un ragionamento normativo strumentale, calcolato, interessato comporta un notevole grado di autonomia (intendendo questa come la possibilità di vagliare la norma esterna, alla luce degli scopi e degli interessi soggettivi). La differenza sta nel ruolo che la norma gioca nella mente di volta in volta e nelle diverse configurazioni della mente normativa.

2.1 Il problema dell'autonomia normativa.

Nella tradizione, la teoria delle norme (siano esse sociali che legali) segue essenzialmente due filoni di ricerca differenti e per molti versi paralleli (e quindi

apparentemente inconciliabili):

i) la concezione convenzionalistica delle norme, intese fondamentalmente come *convenzioni*;

ii) la concezione logico-giuridica delle norme che si basa sulla nozione di *obbligazione*.

L'idea che sosteniamo in questa sede (Conte 1998) è che in realtà esiste una terza possibile prospettiva che integra le due precedenti.

Tale visione unificata si fonda sul concetto di *prescrizione*, intendendo che tanto le norme sociali che quelle giuridiche altro non sono che comandi e prescrizioni che vigono su agenti (autonomi) sociali che decidono se conformarsi a tali norme o trasgredirle. Non è possibile spiegare il modo in cui le norme operano sul sistema sociale senza spiegare anche come esse operano nella mente degli agenti sociali sui quali vigono.

Gli agenti in questione sono necessariamente agenti intelligenti ed autonomi, cognitivamente complessi.

2.2 L'approccio convenzionalistico.

Nella teoria dei giochi le norme sociali sono definite come convenzioni. Esse sono cioè regolarità comportamentali che non implicano un accordo fra gli agenti, ma che emergono dalla dinamica prodotta dall'interazione dei loro interessi individuali (Schelling 1960 e Lewis 1969). In questa prospettiva una norma è essenzialmente la *soluzione ad un problema di coordinazione* (fra agenti). Le norme sono quindi da intendersi come frequenze comportamentali che producono uno strategico conformismo il quale favorisce l'emergenza di norme di coordinamento.

Le norme sociali secondo questa visione non sono altro che *stati di equilibrio* raggiunti in un sistema sociale, dati certi comportamenti individuali. L'agente non sceglie una norma e non esiste alcuna rappresentazione mentale della norma nella mente dell'agente.

Questa visione delle norme non tiene conto del carattere prescrittivo delle norme né distingue fra *comportamento* normativo e *ragionamento* normativo (un agente può decidere se osservare una norma o trasgredirla). Gli approcci dominanti in questo tipo di studio nella teoria delle norme sono per certi versi “esterni” alle menti degli agenti; non tengono conto del modo in cui le norme operano al livello della mente di ogni agente normativo; studiano

l'utilità, l'applicabilità e la funzione di una norma ma *nessuno si è preoccupato fino ad ora del modo in cui le norme sono stabilite da agenti cognitivi che ragionano sulle loro credenze e che agiscono in base ad esse*.

Nell'architettura mentale di tali agenti normativi non compare alcun meccanismo o elemento che garantisca un corretto ragionamento normativo; il comportamento è descritto dall'esterno come normativo perchè corrisponde ad una norma o convenzione; ma tale conformità non deriva da alcuna elaborazione di oggetti mentali simili a norme da parte degli agenti. Gli agenti non si uniformano, ma sono uniformi; il loro comportamento può essere apparentemente conforme ad una norma senza che vi sia alcuna rappresentazione mentale della norma in questione, nella mente degli agenti.

Nella teoria dei giochi la norma di solito viene definita come una condotta prescritta e seguita dai membri di una società (Ullman-Margalit 1977). In Axelrod (1984) il contesto sociale è considerato come una versione reiterata del *Dilemma del Prigioniero con n giocatori*. Dato un certo numero di agenti razionali, la teoria dei giochi può prevedere il punto di equilibrio, in cui nessun agente desidererebbe cambiare più la propria scelta, essendo la scelta di ciascuno per lui *ottimale* (*equilibrio di Nash* – Nash 1950).

Nel celebre Dilemma del Prigioniero due compagni complici di un reato vengono arrestati e ciascuno, separatamente, viene invitato a confessare; entrambi sanno che se manteranno omertà assoluta se la potranno cavare con una pena ridotta; se entrambi confessano, la pena sarà più dura per tutti e due; se infine confesserà solo uno dei due, questi verrà liberato e all'altro verrà inflitta la pena più dura.

Questo tipo di approccio sembra essere però inadeguato a trattare il tema dell'emergenza di una norma. Mentre una strategia cooperativa può essere vista naturalmente come una caratteristica interna dell'agente, tuttavia esiste una bella differenza fra l'essere inclini a cooperare e *sapere* di *dover* cooperare.

L'emergenza delle norme è cosa ben diversa dall'emergenza della cooperazione, in quanto se è vero che comportamenti e motivazioni endogene sono validi elementi per studiare l'emergenza di entrambi i fenomeni sociali, le norme includono anche le prescrizioni sociali.

Nella vita sociale le credenze relative ai costi di una eventuale trasgressione incidono notevolmente sulla scelta di obbedire o meno; nella teoria dei giochi invece, tali costi sono superficiali rispetto alle decisioni dei giocatori. Anche la reazione dell'altro giocatore, nel

Dilemma sopra citato, non è una vera e propria punizione, quanto piuttosto il tentativo di ridurre i propri costi.

La teoria dei giochi è riuscita a dare conto della diffusione di certi comportamenti in popolazioni di agenti che seguono determinate strategie; riesce a spiegare il conformismo sociale di carattere convenzionale; tuttavia non è riuscita a giustificare un aspetto molto importante del meccanismo normativo che invece incide notevolmente nella diffusione dei comportamenti normativi, ovvero la *richiesta normativa*.

L'intuizione di Hart (1961) che le norme sono obbligazioni imposte non ha ricevuto un'adeguata attenzione e teorizzazione nella teoria dei giochi. Solitamente, la gente vuole che le norme vengano osservate; tale scopo viene espresso in diversi modi: dalle aspettative, ai comandi, alle richieste esplicite, fino alla disapprovazione implicita della trasgressione e al rimprovero esplicito. La tradizionale visione delle norme (in teoria dei giochi) come effetti emergenti dei sistemi sociali complessi non riesce a dar conto di questo particolare aspetto.

Alcuni autori si sono posto il problema relativo al rapporto fra norme e *motivazioni endogene*: Axelrod ha parlato di "sanzioni interne" (1984); sulla scorta di questa visione e del lavoro di Elster (1987) (che tratta le norme come motivazioni esogene), la Bicchieri (1990) ha proposto una teoria delle norme come "preferenze condizionate". La Bicchieri applica alle norme un principio generale di rappresentazione secondo cui un'obbligazione è sempre condizionata a qualcos'altro, in particolare ad un'altra motivazione. L'obbligazione quindi diventa per definizione una motivazione condizionata a, o derivata da, un'altra motivazione (possibilmente endogena).

Questo tentativo di conciliare il carattere esogeno della norma con il paradigma della teoria dei giochi sembra tuttavia insoddisfacente, giacché l'autrice definisce una norma sociale (o preferenza condizionata) come un *equilibrio*, ossia come una combinazione di strategie tali per cui ciascun agente massimizza la propria utilità attesa adattandola, a patto che anche tutti (o quasi) gli altri facciano lo stesso (proprio come nel Dilemma del Prigioniero). Ma le preferenze sono spesso intrinsecamente condizionate ed endogene.

Per modellare il comportamento di un agente (autonomo) normativo sembra allora opportuno far riferimento a una teoria che consideri le norme come:

i) *prescrizioni*, cioè richieste, direttive, comandi posti su un gruppo di agenti da una autorità normativa riconosciuta o dalla comunità sociale (Conte e Castelfranchi

1995);

ii) *oggetti mentali*: cioè oggetti che la mente di un agente possa trattare (Conte e Castelfranchi 1995a).

La prescrizione è un aspetto centrale rispetto al meccanismo normativo; una norma è percepita come tale solo se è in qualche modo associata alla credenza che vi sia una volontà generale che essa sia osservata.

Secondo un certo numero di filosofi morali e del diritto la visione convenzionalistica delle norme incappa nell'erronea identificazione, già segnalata da Hume, fra ciò che è *normativo* e ciò che è *normale* (si veda Hart 1961). Ognuno di noi sa che l'accadere di qualcosa con (grande) frequenza non ne stabilisce il carattere di necessità; inoltre la stabilità di una strategia non rappresenta un buon indicatore di atti conformi a una norma. La soluzione trovata dalla teoria dei giochi si basa sul fatto che il calcolo utilitaristico necessario alla scelta se uniformarsi o meno ad una certa condotta non ha niente a che fare con ciò che di solito si intende per conformismo. Il calcolo della convenienza ha sempre luogo nelle decisioni di un agente autonomo. La questione interessante rispetto alle norme riguarda invece le opzioni rispetto alle quali l'agente si trova a scegliere, quando si trova ad effettuare una decisione normativa; cioè: cosa c'è dietro la singola conformità, la singola decisione di attenersi alla norma? Un qualche meccanismo conformistico o il semplice calcolo utilitaristico, indipendente dalla rappresentazione della norma corrispondente all'azione stessa?

2.3 L'approccio logico-giuridico.

La visione convenzionalistica sembra essere dunque un pò debole.

La visione logico-giuridica presenta invece il problema opposto, riconducendo il concetto di norma a quello esplicito definito dal diritto positivo, quindi alle leggi, alle norme legali.

Kelsen (1991) ha proposto, ad esempio, una concezione troppo forte di norma: la norma non sarebbe riconducibile a un fatto esterno osservabile (comportamento), ma essa indicherebbe, nel suo significato più astratto e generale e per gli effetti che ha, un livello di realtà superiore, la *volontà*. Le norme, secondo questo autore, esprimono una volontà. Tuttavia, la norma sopravvive all'atto della sua emanazione, alla volizione del suo emanatore. Ma che tipo di volontà è in grado di sopravvivere al suo emanatore? La volontà espressa da

una norma è, in realtà, una volontà molto particolare, è la volontà dell'autorità normativa. Nei termini di Kelsen, la norma presuppone un atto di produzione della norma, nello specifico un atto istituzionale (legittimo) di emanazione delle norma. In una visione così restrittiva delle norme, una norma può essere intesa solo come norma positiva, escludendo la maggior parte delle norme sociali che invece sono consuetudinarie, spontanee, implicite.

La concezione logico-giuridica delle norme nasce dunque da due esigenze fondamentali:

i) distinguere ciò che è *normale* da ciò che è *normativo*;

ii) dar conto del *comando*, della *prescrizione*, della *volontà* espressa dalla norma.

Per risolvere questa duplice esigenza, la concezione logico-giuridica postula che la norma esiste se ed in quanto è emanata attraverso un atto di produzione legittimo.

Questa importante ed influente concezione delle norme presenta due fondamentali problemi, non riuscendo a chiarire il nesso :

i) fra norme giuridiche e norme sociali

ii) fra norme ed obbligazioni.

Questa teoria presenta l'indiscusso vantaggio di cogliere nelle norme qualcosa di più di una mera frequenza statistica, di una regolarità comportamentale o del diffondersi di un comportamento meramente conformistico. *La norma è un comportamento prescritto, obbligatorio*; ma non vengono fornite ulteriori spiegazioni su cosa sia una obbligazione.

La concezione logico-giuridica nasce dall'esigenza di contrastare la tesi secondo cui la nascita di una norma coincida con la nascita e diffusione di una regolarità comportamentale. Tuttavia, seguendo l'impostazione suggerita da questo approccio, dovremmo pensare che una norma presupponga sempre l'esistenza di altre norme, le quali legittimino l'autorità emanatrice della norma in questione.

2.4 Rapporto fra convenzioni e norme.

L'emergenza delle norme non è un processo necessario. Esse possono venire emanate in modo deliberato, oppure possono derivare da *protonorme* o *metanorme* (come le consuetudini). Inoltre, le convenzioni non danno luogo necessariamente a norme. Non è detto che l'emergenza di una convenzione produca necessariamente la formazione di una nuova norma; come non è detto che la diffusione di una regolarità comportamentale dia luogo alla

obbligatorietà o prescrizione di un certo comportamento.

Le convenzioni non sono né necessarie né sufficienti per l'emergenza di una norma.

Inoltre, il processo che da una convenzione porta ad una norma non è necessariamente continuo né omogeneo. Gli studi di dinamica sociale hanno messo in luce l'emergenza delle convenzioni in termini di propagazione di una determinata strategia sociale (Schelling 1960), spiegandone la diffusione e non la generazione. Il salto concettuale che si ha nel passaggio da una convenzione ad una norma risiede proprio nell'emergenza di un fenomeno nuovo: la formazione di un oggetto mentale specifico, la norma. Essa definisce l'obbligatorietà di un comportamento, nella mente degli agenti.

Una norma può emergere in una società se prima si dà nella mente degli agenti (immergendosi nelle loro menti). La cognizione normativa, in base a principi evolutivo-adattativi simili a quelli che hanno selezionato le motivazioni morali nei primati (Wright 1994), non sarebbe possibile se la mente umana non presentasse la capacità di ragionare in base a rapporti strumentali, finalistici, mezzo-a-scopo. L'obbedienza normativa non sarebbe possibile se gli organismi non fossero dotati di meccanismi cognitivi per il ragionamento sociale e l'adozione normativa.

3. Ontologia normativa[2]

Dal momento che innovazione normativa è il risultato di un insieme complesso di intricate definizioni teoriche, è necessario fornire una ontologia condivisa, o in altre parole, forgiare un vocabolario di nozioni correlate fra loro. Con il termine ontologia si intende qui uno strumento convenzionale e operativo, una serie di nozioni teoriche che vengono definite una in rapporto all'altra. L'obiettivo è quello di creare collegamenti concettuali tra i concetti normativi espliciti.

Lo scopo della ontologia cui qui ci riferiamo deriva ovviamente dall'esigenza di trattare un argomento particolarmente complesso come quello relativo all'emergenza ed innovazione delle norme sociali avendo a disposizione un vocabolario il più possibile chiaro ed inequivocabile.

Per *innovazione* / *emergenza* normativa si intende un processo complesso, in cui effetti emergenti determinano nuove proprietà, per il livello sottostante di generazione.

Una interazione ricorsiva tra i due livelli è stabilita da un complesso processo di *feedback*. Ciò include due sotto-processi:

- **Emergenza**, cioè il processo mediante il quale gli effetti macro sono generati da e implementati attraverso micro enti (sociali) che (inter-)agiscono;
- **Immergenza**, vale a dire il processo graduale e complesso grazie al quale il macro-effetto sociale, nel nostro caso una norma (sociale) specifica, impatta sulle menti degli agenti, generando un certo numero di *loop* intermedi.

Prima che un qualsiasi effetto globale possa emergere, specifici eventi locali interessano il sistema di generazione, le sue credenze ed i suoi scopi, in modo tale che gli agenti si influenzino l'un l'altro a convergere su effetto macroscopico globale.

L'emergenza di norme sociali è un importante circuito costituito da reti locali, in cui:

- si verificano parziali o iniziali effetti macroscopici osservabili dei comportamenti locali; questi effetti retroagiscono su (un sottoinsieme di) menti degli osservatori,

[2] Per gli argomenti qui di seguito trattata, si veda: Andrighetto, G., Conte, R., Turrini, P. (2007). "EMIL Ontology". *Technical Report*, 00307, LABSS-ISTC/CNR.

modificandole (producendo nuovi stati interni, emozioni, credenze normative, scopi normativi, ecc...);

- gli agenti si comunicano l'uno all'altro i propri stati interni, attivando un processo di influenzamento normativo;
- le credenze normative si diffondono attraverso le menti degli agenti;
- i comportamenti si conformano progressivamente agli stati trasmessi;
- effetti macroscopici iniziali vengono rinforzati / indeboliti, a seconda del tipo di stati mentali che si diffondono.

Abbiamo dunque bisogno di nozioni relative alle norme, delle quali forniremo qui di seguito una lista.

Dinamica: l'indagine scientifica si basa su un approccio simulativo, basato quindi sulla realizzazione di un programma che giri su un calcolatore e che produca una serie di dati (una sorta di laboratorio virtuale). Si presterà particolare attenzione, nel nostro caso, alle *modifiche* piuttosto che alle tipologie di norme e alle loro funzioni.

Orientato verso l'innovazione: si tratta di un caso speciale di dinamica. Per innovazione, si intende un processo progettato o voluto dagli organi istituzionali o sociali (può trattarsi anche di un semplice movimento di opinione). Una visione meramente convenzionalista delle norme e una dinamica spontanea ed emergente non possono dare conto di questo processo: invece di aspettare che emergano spontaneamente nuove regolarità, le agenzie mirano a imporre nuovi obblighi o diritti, nuove autorizzazioni o divieti. In una parola, nuove norme.

Ibrido: integrato sia negli oggetti sociali che mentali. In questa prospettiva, le seguenti nozioni sono ritenute inadeguate:

- *Epifenomeno*: le norme sono *pattern* sociali osservabili, interpretati "come se fossero" il risultato di una qualsiasi forza o processo di carattere normativo. Al contrario, in questa sede noi siamo interessati a realizzare modelli sociali effettivamente derivanti dall'azione delle norme nella e sulla società;

- *Comportamentista*: caratterizza le norme come regolarità osservabili, derivanti da costrizioni reciproche che gli agenti si impongono nella prassi comune. Al contrario, noi partiamo dal presupposto che è importante guardare a cosa accade nella mente degli agenti, al fine di comprendere come le norme operano;
- *Convenzionalista*: le norme sono viste come convenzioni. Anche se necessaria, questa è ancora una visione insufficiente delle norme, soprattutto quando si vuole fare i conti con l'innovazione normativa.

Comando: è una richiesta coercitiva di azione, basata sul potere del comandante sul destinatario.

Norma: qui si considera una norma - sia essa giuridica, sociale, o morale - come "una guida prescritta per un comportamento, che è generalmente rispettata dai membri di una società" (Ullmann-Margalit 1977).

Un vivace dibattito sul concetto di norma si è sviluppato in diversi rami della filosofia, della logica, delle scienze cognitive, della teoria degli agenti, della teoria sociale e della teoria dei giochi.

Una norma si diffonde attraverso la popolazione, grazie alla diffusione di una credenza condivisa, la *credenza normativa*. Una credenza normativa, a sua volta, è la credenza che un dato comportamento, in un dato contesto, per un determinato insieme di agenti, è proibito, obbligatorio, permesso, (Wright 1963; Kelsen 1979; Conte e Castelfranchi 1999; 2006). Detto diversamente, una credenza normativa è la credenza che vi sia un comando basato su un sistema deontico. Anche se necessario per la diffusione del comportamento prescritto, il comando normativo non è sufficiente: fattori aggiuntivi sono costituiti dalla forza della mandatorietà (obbligatorietà e rinforzo) del comando, dalla persuasività e credibilità della fonte, dalla compatibilità con le norme esistenti (i conflitti fra norme spesso portano a violare una norma o l'altra). Naturalmente, una credenza normativa non implica che una data norma sia stata deliberatamente rilasciata da una qualche autorità istituzionale. Le norme sociali sono spesso istituite in virtù di effetti indesiderati. Tuttavia, una volta emersa una data norma sociale, si crede che essa sia basata su una qualche autorità normativa, anche se anonima e impersonale.

Va sottolineato che una norma è una richiesta che si chiede di adottare in quanto si tratta di una prescrizione che si deve rispettare ed è pienamente applicata solo quando è rispettata di per se stessa (anche se questa condizione di "felicità" si applica raramente *de facto*). Anche i comandi normativi sono spesso adottati sotto un qualche effetto di rinforzo. Tuttavia, questo tipo di adozione non è soddisfacente, per così dire, dal *punto di vista della norma*, qualora una simile prospettiva possa essere ipotizzata. La condizione di felicità è che la norma sia stata *accettata*, per dirla in termini di Hart (1961), o *interiorizzat*a, per dirla in termini di Durkheim (1951), perché è riconosciuta come tale (cioè come una norma). In altre parole, affinché la norma sia soddisfatta, non è sufficiente che l'azione prescritta venga eseguita, ma è necessario che la norma sia rispettata, grazie allo scopo normativo, che è lo scopo che deriva dal riconoscimento e dalla successiva adozione della norma.

Così, affinché un comportamento basato su una norma possa avere luogo, una credenza normativa deve essere generata nella mente dei destinatari della norma, ed il corrispondente scopo normativo deve essere formato e perseguito. In questo senso, la nascita e la stabilizzazione di una norma implica la sua *immergenza* nelle menti degli agenti.

Passiamo ora a definire alcune componenti della elaborazione mentale delle norme.

Credenza normativa: credenza che un determinato comportamento, in un dato contesto per un determinato insieme di agenti, sia vietato, obbligatorio, permesso. Più precisamente, la credenza dovrebbe essere che "vi è una norma che vieta, prescrive, permette che ... ". Infatti, le norme sono volte a generare e rilasciare credenze (normative) corrispondenti. In altre parole, le norme devono essere riconosciute come tali, per poter funzionare correttamente.

Scopo normativo: scopo interno, relativizzato ad una credenza normativa. Da un punto di vista cognitivo, gli scopi sono rappresentazioni interne che allo stesso tempo innescano e guidano l'azione: essi rappresentano lo stato del mondo che gli agenti vogliono raggiungere attraverso l'azione e che essi tengono di mira durante l'esecuzione della stessa (Conte 2009). Uno scopo è *relativizzato* quando esso è perseguito in quanto e nella misura in cui un determinato evento o stato del mondo è atteso o previsto essere vero (Cohen e Levesque, 1990a). Un esempio è il seguente: domani, voglio andare a raccogliere dei funghi (scopo relativizzato), in quanto e nella misura in cui credo che domani pioverà (evento atteso). Nel

preciso istante in cui io smetta di credere che domani pioverà, farò cadere ogni speranza di trovare i funghi.

Uno scopo normativo è diverso, da un lato, da un semplice vincolo (che riduce l'insieme di azioni a disposizione del sistema) e, dall'altro, dalla definizione di scopi comuni. Per quanto riguarda i vincoli comportamentali, uno scopo normativo è meno impellente: un agente dotato di scopi normativi ha la possibilità di confrontarli con gli altri scopi (non normativi) e, in certa misura, può scegliere quale scopo verrà perseguito. Solo se un agente è dotato di scopi normativi si può dire che esso rispetta o viola una norma. Per quanto riguarda gli scopi comuni, uno scopo normativo è ovviamente più impellente: quando un agente decide di rinunciarvi, sa che al tempo stesso contrasta uno dei suoi scopi e viola una norma.

Adozione normativa: formazione di uno scopo normativo, a partire da una credenza normativa e grazie ad alcune regole di intervento. Ad esempio, lo scopo normativo di un determinato agente ***x*** circa l'azione ***a*** è lo scopo che l'agente ***x*** mostra di avere per tutto il tempo in cui ha una credenza normativa circa ***a***. Più specificamente, ***x*** ha uno scopo normativa solo se crede di essere soggetto ad una norma.

Ragionamento normativo: operazioni mentali sulla rappresentazione interna di una data norma, che può portare all'adozione di tale norma, formando così uno scopo normativo (nuovo).

Convenzione: regolarità comportamentale (Gilbert 1981, 1989, Lewis 1969, Sugden, 1986/2004, 1998; Young 1993, 2006), vale a dire una pratica o una procedura ampiamente osservata dai membri di una determinata rete sociale, sulla base:

- dello scopo dell'agente di conformarsi a tale comportamento, al fine di agire come gli altri,
- sull'aspettativa reciproca che gli altri si conformino a tale comportamento.

Più in particolare, le convenzioni sono una classe di problemi (arbitrariamente selezionati da un potenziale insieme di candidati alternativi) classificati come (puri) *giochi di coordinamento* (per esempio, la convenzione di tenere la destra o la sinistra durante la guida, sottolineata da David Lewis (1969)) , basati sulla interdipendenza e le aspettative reciproche.

Il confine tra convenzioni e norme non è del tutto chiaro. Le convenzioni possono acquisire forza mandatoria nel corso del tempo; a volte esse possono arrivare ad essere prescritte, e questo è un fattore importante per l'emergenza normativa. Un buon esempio è l'*etichetta* (l'educazione), che è a metà strada tra una norma sociale (con obblighi e, eventualmente, sanzioni) e una convenzione. Il saluto è un comportamento educato, ed il modo in cui si saluta qualcuno - se agitando le mani o dicendo "ciao" - è regolato da convenzioni; d'altra parte, quando si ricevono dei saluti, è *buona norma* rispondere, probabilmente a causa della norma sociale di reciprocità.

Frame normativo: insieme di funzionalità che caratterizzano la norma e che definiscono i suoi aspetti cruciali, i quali sono ulteriormente scomponibili. Ciò significa che quando si specifica una norma, abbiamo bisogno di dire qualcosa su ognuno di questi aspetti. Una possibile ricerca riguarda quindi l'eventualità che queste caratteristiche siano le condizioni necessarie e sufficienti per costruire una norma. Noi individuiamo cinque aspetti essenziali:

1) Deontico: un deontico è sostanzialmente un modo per distinguere le situazioni giuste/ accettabili da quelle ingiuste/inaccettabili. Quello che è importante circa le questioni del riconoscimento in generale è questo: l'autorità che emana l'obbligo ha bisogno di essere riconosciuta e accettata come tale dagli altri agenti, affinché essi si rapportino alla norma e decidano se rispettarla o meno. Per quanto riguarda la validità basata sul deontico, possiamo individuare:

- obbligo,
- divieto,
- permesso.

2) Sorgente: una sorgente è il luogo da cui si emana la norma. Distinguiamo la fonte in:

- personale: "per luogo da cui emana la norma" si intende in questo caso "l'insieme non
- vuoto di agenti che hanno eseguito l'azione la quale consente l'emanazione della norma (che dopo ciò esiste)";

- impersonale: “il luogo da cui emana la norma” significa in questo caso “la comunità che ha permesso la norma”. E' chiaro che considerare la “comunità” come l'equivalente di “l'insieme di tutti gli agenti” farebbe collassare le due nozioni l'una sull'altra.

3) Ruolo normativo: con ruolo normativo si intende la partizione degli agenti coinvolti in una norma. Distinguiamo:

- Legislatori, gli agenti che la emettono;
- Destinatari, gli agenti a cui la norma prescrive una determinata azione come consentita o non consentita;
- Difensori, cioè quegli agenti che condividono lo norma e la fanno rispettare;
- Osservatori, quelli che acquisiscono credenze circa una norma, sia essa applicata, violata, emanata.

4) Meccanismi di rinforzo: operazioni che tentano di modificare le azioni degli agenti, al fine di renderli conformi ad una norma. Molto schematicamente, possiamo distinguere:

- Sanzioni: meccanismi di attuazione che inibiscono l'azione degli agenti;
- Incentivi: meccanismi di attuazione che favoriscono le azioni degli agenti.

Il modo in cui le azioni possono essere favorite o inibite segue percorsi (mentali) precisi negli agenti cognitivi. In questo senso, i meccanismi di applicazione seguono un percorso nella mente di un agente, sfruttando i processi intra-agente. Inoltre, possono essere utilizzati artefatti sociali per sanzionare o per favorire le azioni degli agenti. La reputazione, per esempio, può funzionare come una sanzione normativa; ma può essere utilizzata anche come un incentivo normativo.

5) Controllo: è il modo in cui i meccanismi di rinforzo sono applicati (Conte e Dignum 2001; Conte e Paolucci 2004). Esso implica sia il *monitoraggio* - che controlla la violazione - sia l'influenza - che spinge attivamente gli agenti cognitivi verso il rispetto di una norma. Questi meccanismi possono essere:

- centralizzati: un solo agente (singolo o sovra-individuale) ha il diritto di sanzione;
- distribuiti: ognuno è in grado di difendere la norma.

Pertanto, va osservato che il controllo centralizzato fa uso di regole istituzionali per la regolamentazione, mentre il controllo distribuito non presuppone alcuna delega.

4. Un approccio cognitivo allo studio delle norme: emergenza ed immergenza[3]

In questo capitolo il concetto di *emergenza* nei sistemi sociale complessi è ridiscusso come strumento necessario per una teoria del *link* macro-micro (società - agente). Facendo riferimento al modello di segregazione di Schelling (1960), gli effetti emergenti sono definiti come *effetti generati da micro-organismi sociali (inter)agenti, e implementati dalle loro regole (non incorporati in esse)*. Esamineremo il percorso di ritorno dal macro al micro, vale a dire la *downward causation*. Distingueremo *loop* semplici e complessi, facendo riferimento ad esempi concreti tratti dalla letteratura sociologica e da quella computazionale. Discuteremo poi:

i. come un dato macro-effetto è implementato dai livelli inferiori;
ii. due meccanismi specifici di attuazione, *emergenza di secondo ordine* ed *immergenza*.

4.1 Generazione ed Emergenza

All'inizio del secolo scorso, alcuni scienziati sociali e antropologi (Alexander, 1920; Broad, 1925) si riferirono alle proprietà emergenti a livello macro-sociale come a quelle proprietà che non possono essere desunte dalle proprietà al livello sociale sottostante. Tale affermazione è stata fortemente criticata (Hempel e Oppenheim, 1948) e si è sostenuto che si basi su una confusione logica tra *proposizioni* e *proprietà*. Come gli epistemologi hanno osservato, solo le proposizioni, e non le proprietà, possono essere desunte. Di conseguenza, l'affermazione emergentista deve essere riferita ad una determinata teoria, in una determinata fase del suo sviluppo. L'affermazione in questo modo è indebolita e trasformata in un una relativistica, in cui si afferma che proposizioni su proprietà macro-sociali non possono essere dedotte da proposizioni su quelle micro-sociali, rispetto agli attuali confini teorici.

[3] Per gli argomenti qui di seguito trattati, si veda: Conte, R., Andrighetto, G., Campennì, M., Paolucci, M. (2007). "Emergent and Immergent Effects in Complex Social Systems", *Proceedings of AAAI SYMPOSIUM, SOCIAL AND ORGANIZATIONAL ASPECTS OF INTELLIGENCE*, november 8 - 11, WASHINGTON DC, Usa.

Tuttavia, una tale formulazione dispensa dalla nozione di emergenza: nella nuova affermazione relativistica, le proprietà emergenti non sono (ancora) dedotte. Quindi, ciò che emerge è quello che è (ancora) non-spiegato. Una volta spiegato (Epstein, 2007), ogni fenomeno cessa di essere emergente (in senso forte). Di conseguenza, la nozione di emergenza è poco pregnante, sotto il profilo scientifico.

In questo capitolo, analizziamo il problema da una prospettiva diversa. Partiamo da un aspetto cruciale dei sistemi sociali complessi, cioè la differenza tra *l'implementazione* e *l'incorporazione*: una entità macro-sociale è sempre implementata da entità micro-sociali, dal momento che può agire e ha effetto solo attraverso le azioni di organismi micro-sociali, vale a dire gli individui. Qualche volta, una entità macro-sociale può essere incorporata in un livello inferiore, quando è rappresentata all'interno di esso, ad esempio all'interno delle sue regole.

Sulla base di questa distinzione fondamentale, si può definire *emergente* ogni *effetto che è implementato da organismi micro-sociali che (inter)agiscono, ma che non è incorporato in essi*.

Cercheremo di presentare un processo multilivello, vale a dire la generazione da micro a macro, e discuteremo dell'emergenza sociale in sistemi complessi come una parte del processo di generazione. Faremo questo, illustrando brevemente il modello di segregazione di Schelling.

Infine, discuteremo il percorso di ritorno dal macro al micro, cioè la *downward causation*, un processo che non è certo una novità per la comunità scientifica (Campbell, 1974). Infatti, si può dimostrare che le dinamiche micro-macro consistono in molteplici, semplici e complessi, *loop* di retroazione (Andersen et al., 2000).

Nell'ultima parte di questa sezione, illustreremo come un dato macro-effetto può essere implementato dalla realtà sottostante, introducendo la distinzione tra:

- *immergenza*, dove proprietà macro-sociali causano nuove proprietà micro che riproducono o supportano l'effetto. Di conseguenza, gli effetti emergenti possono anche arrivare ad essere parzialmente incorporati in sistemi micro-sociali, ma non è necessariamente così. In ogni caso, non è la rappresentazione degli effetti di per sé ad implementare l'effetto macrosociale, ma una nuova proprietà, meccanismo o regola, derivante dalle proprietà di livello superiore;
- *incorporazione*, in cui gli effetti emergenti vengono rappresentati nel sistema che

li produce, e questa rappresentazione contribuisce a riprodurre l'effetto.

Da questi diversi tipi di implementazione, si può immaginare che emerge una caratteristica fondamentale della complessità sociale. Un esempio tipico è l'emergenza delle norme, più specificamente, *l'innovazione normativa*. A nostro avviso, l'innovazione normativa è caratterizzata dalla presenza di due complementari processi di emergenza ed immergenza: le norme non possono emergere a meno che non si immergano contemporaneamente nelle menti degli agenti.

4.2 *Downward Causation*

Può una proprietà emergente, macro-sociale, generare effetti al livello più basso? Sì. Esistono due modi principali in cui la *downward causation* occorre:

- *loop* semplice: si tratta della chiusura del circuito macro-micro. L'effetto emergente retroagisce sul livello più basso, determinando una nuova proprietà del sistema generante.
- *loop* complesso o implementazione: l'effetto emergente determina nuove proprietà per mezzo delle quali l'effetto è riprodotto nuovamente.

Questo tipo di *loop* comprende due sottoprocessi:

- *Immergenza*, vale a dire il processo per mezzo del quale l'effetto emergente modifica il modo in cui funziona il sistema generante, alterandone le regole di generazione o i meccanismi.
- *Emergenza di secondo ordine* (o *incorporazione*): vale a dire il processo per mezzo del quale un effetto emergente è riconosciuto dal sistema che lo produce (Dennett 1995, Gilbert, 2001).

Gli effetti emergenti possono retroagire sul sistema che li genera sia chiudendo il circuito, sia aprendo un nuovo *loop*.

4.2.1 *Loop* semplice.

L'effetto emergente retroagisce sui sistemi generanti determinando nuove proprietà che possono interferire negativamente o positivamente con il resto dell'attività dei micro-sistemi. Questo è il caso di una serie di proprietà, come ad esempio i *diritti*, lo *status sociale* e il

potere sociale, nonché le valutazioni che gli agenti si formano l'uno dell'altro (ad esempio, la *reputazione*).

Vediamo un esempio di semplice *downward causation*: il potere di negoziazione.

Reti di dipendenza.

In un ambiente comune, azioni eseguite da un agente producono effetti sugli scopi di altri agenti. Questi sono limitatamente autosufficienti, nel senso che non sempre sono in possesso di tutte le risorse necessarie per raggiungere i loro scopi. In queste condizioni, le reti sociali di dipendenza (Sichman et al., 1994; Sichman e Conte, 2002) emergono come le interconnessioni tra gli agenti dotati di un numero finito di goal e risorse per la loro realizzazione.

Supponiamo, per esempio, che in una serie di agenti ***<a, b, c>***, **a** sia dotato di goal *p* e azione *a(q)*, mentre **b** e **c** siano entrambi dotati di obiettivo *q* e azione *a(p)*. Le loro interconnessioni risultato in una rete di dipendenza, dove gli agenti **b** e **c** sono socialmente dipendenti da **a**, mentre **a** OR-dipende da entrambi **b** o **c**. A sua volta, questa distribuzione non uniforme del potere di scambio determina un nuovo effetto al livello più basso: gli agenti derivano un uguale potere di scelta, o, come lo abbiamo chiamato, il potere di negoziazione (Conte et al. 1998).

In particolare, **a** ottiene un maggiore potere contrattuale rispetto a **b** o **c**: **a** sarà in grado di fare una scelta, vale a dire di scegliere i suoi partner di scambio, mentre **b** e **c** non hanno scelta. Presumibilmente, a causa di questa distribuzione eterogenea di potere, lo scambio fornirà risultati (*payoffs*) ineguali per i partecipanti, in cui l'agente **a** risulterà avvantaggiato rispetto sia a **b** che a **c**. Questo esempio mostra chiaramente che un effetto emergente (per esempio, una rete di dipendenza) può influire sul livello sottostante. Questo tipo di *downward causation* - o, nei nostri termini, *downward generation* - genera nuove proprietà (potere di negoziazione) dei sistemi di livello inferiore, interferendo positivamente o negativamente con i loro futuri successi.

Un problema riguardo alla *downward causation* è fino a che punto contribuisce alle altre dinamiche nel sistema globale. Indubbiamente, le proprietà come il potere sociale, compresi il potere di negoziazione, e la reputazione, hanno un indubbio ma indiretto impatto sui futuri

successi degli agenti: gli agenti possono soffrire o godere dei loro effetti in base alle azioni che gli altri, che interagiscono con loro, intraprendono sulla base delle loro rappresentazioni di tali proprietà.

Talvolta, queste nuove proprietà non solo interferiscono con il grado di adattamento dei singoli agenti, ma fanno anche esplodere nuovi effetti emergenti al livello sociale più alto. Questo è ciò che noi chiamiamo un loop complesso.

4.2.2 *Loop* complesso (implementazione)

A volte, questa retroazione sui livelli più bassi può iniziare nuove dinamiche complesse, per mezzo delle quali le nuove proprietà si rafforzano o riproducono l'effetto emergente. L'effetto al livello superiore più o meno gradualmente si implementa, seleziona una specifica routine per mezzo della quale è (più volte) eseguito dai singoli agenti. Quando questo accadrà, e come?

Immergenza.

Qui, la proprietà macro-sociale influisce sui sistemi generanti attraverso i meccanismi di questi ultimi, aumentando la probabilità di essere riprodotta da essi.

Si consideri la conformità sociale, per esempio le *regolarità comportamentali*, come un fattore fondamentale di regolazione sociale.

Secondo Mary Douglas (1986), gli agenti "si spremono l'un l'altro" in quanto risultato di prassi comuni, o istituzioni sociali. In altre parole, come mostra il comportamento di sciamare di specie inferiori (*swarm behaviour*), effetti collettivi possono essere implementati da regole semplici, senza la necessità di una percezione o comprensione dell'effetto risultante.

Va sottolineato che *l'immergenza* non è un'esclusiva di organismi semplici. Anche tra gli esseri umani, la regolarità comportamentale può essere implementata da una serie di meccanismi diversi, che comprendono, ma non sono ridotti a una vera e propria *regola di maggioranza*.

Infatti, quest'ultima non è sempre applicata, né è sempre efficiente (si veda sotto *Stalemata*).

Regolarità comportamentali si ottengono da più o meno intelligenti forme di adattamento inter-agente (Conte e Paolucci, 2001; Conte, 2002), che comprendono, oltre alle

classiche tipologie di apprendimento sociale (facilitazione sociale, valorizzazione locale, imitazione), meccanismi meno noti:

- *effect arena*, per mezzo del quale il comportamento *i-esimo* degli altri in un ambiente comune spinge ciascuno ad un comportamento successivo analogo che si rafforza in modo crescente al fine di mantenere l'efficienza (si pensi ad un pub rumoroso, in cui ognuno deve spingere la propria voce oltre quella degli altri per essere a mala pena udibile da uno dei vicini). Ovviamente, nessuno vuole che il rumore cresca, ne ha bisogno di percepire che è in crescita. Un caso particolare è la *vulnerable position*, in cui gli agenti sono invitati a comportarsi come gli altri per evitare una posizione di rischio (per esempio, le automobili sono costrette ad accelerare in autostrada);
  - # effetto emergente: asintotico aumento del *i-esima* regolarità comportamentale su una delle sue dimensioni;
  - # regola immergente: “aggiorna l'*i-esimo* comportamento all'*i-esimo* + *n* per mantenere l'effetto”;
- *garden-party shower*, in cui ciascuno può dire cosa sta succedendo osservando gli altri, e conformarsi a ciò, al fine di ottenere un risultato condiviso, ma non ancora comune: la gente cerca riparo, ad esempio sotto una tettoia, osservando gli altri fare lo stesso, in quanto deduce che sta piovendo (Searle, 1999);
  - # effetto emergente: la regolarità del comportamento;
  - # regola immergente: “se il comportamenti degli altri è più efficiente rispetto ai propri obiettivi, adattati ad esso”;
- *social monitoring*, in cui gli agenti cercano le convenzioni e le regole sociali in vigore nel contesto in cui si trovano, confrontandosi con gli altri in quanto fonti di informazione (Conte e Dignum, 2001);
  - # effetto emergente: la conformità;
  - # regola immergente: "comportarsi adeguatamente, e controllare gli altri per sapere che cosa è appropriato nelle attuali circostanze".

Ora, il circuito da proprietà micro-sociale a effetti macroscopici emergenti diventa ricorsivo. Le società sono caratterizzate da fenomeni fuori equilibrio e da processi di questo tipo. Proprio per questo motivo hanno bisogno di uno approccio generativo che coniughi processo

bottom-up e processo *top-down*.

Immergenza delle norme

L'esempio più lampante di immergenza sono le norme sociali. *Le norme sociali sono prescrizioni sociali implicitamente trasmesse da un agente ad un altro, fondate su deontici del tipo "si deve fare ...", "le persone sono obbligate a ...", a volte trasmesse attraverso valutazioni in forma "è bene / male fare ...".*

L'emergenza normativa può essere vista come un meccanismo di regolazione sociale o di soluzione di problemi di coordinamento. Gli agenti non hanno bisogno di rappresentarsi gli effetti delle norme, al fine di conformarsi ad esse. Tutto ciò che devono fare è *accettare* la norma. Come è possibile?

Nella visione delle norme a due facce, una esterna (sociale) ed una interna (mentale) (Conte e Castelfranchi, 1995; Conte, 1998; Conte e Castelfranchi, 1999), *una norma emerge come una norma* solo quando *emerge nelle menti degli agenti coinvolti*, non solo attraverso le loro menti. *La mente è un sistema integrato per la conservazione e la manipolazione di rappresentazioni per raggiungere certi scopi* (Miller, Galanter e Pribram, 1960). In altre parole, funziona come una norma solo quando gli agenti la riconoscono come una norma. L'emergenza di una norma implica la sua immergenza nella mente dell'agente. Una norma sociale è una norma solo dopo la sua immergenza. Quando il suo carattere normativo, vale a dire prescrittivo, è riconosciuto dall'agente, la norma dà luogo ad un comportamento normativo di tale agente.

Così, l'immergenza è un necessario correlato dell'emergenza, almeno in un sottoinsieme di macrofenomeni sociali, come le norme.

Un aspetto ancora poco esplorato delle norme è il *meccanismo che consente loro di influenzare i comportamenti di agenti autonomi e intelligenti* che le implementano. Le norme non solo regolano il comportamento, ma agiscono anche su diversi aspetti della mente.

In Andrighetto et al. (2007) viene presentata l'analisi dei processi inter-agenti e intra-agente necessari per far fronte all'emergenza di una norma. Da un lato, i processi inter-agenti contribuiscono a caratterizzare la trasmissione della norma; dall'altro, i processi e le proprietà intra-agente contribuiscono a definire la sua immergenza. Per quanto riguarda i processi inter-agenti, in questo lavoro particolare attenzione viene dedicata ai meccanismi di insorgenza e

diffusione delle entità o proprietà a livello aggregato, a partire dall'interazione tra gli tra agenti. Per quanto riguarda i processi intra-agente, cercheremo di difendere nel prossimo capitolo l'utilità di una architettura normativa per il *riconoscimento*, l'*innovazione*, il *rispetto* e la *difesa* di una norma (sociale).

Incorporazione (Emergenza di 2 ° Ordine).

A volte, gli agenti diventano gradualmente consapevoli degli effetti che essi contribuiscono a generare. In tal caso, sviluppano uno specifico tipo di proprietà, la rappresentazione mentale dell'effetto emerso. Questo è ciò che alcuni autori chiamano emergenza di secondo ordine.

A volte, le rappresentazioni degli agenti rispetto agli effetti emergenti modificano le loro azioni, producendo in tal modo un ulteriore effetto al livello superiore. Le dinamiche sociali diventano così ricorsive. Come può accadere? Come descritto da Dennett (1995), il processo chiamato emergenza di secondo ordine non è sufficiente a spiegare questa complessa dinamica dal momento che le credenze non bastano a scatenare automaticamente le azione. Infatti, a volte divenire consapevoli di un dato effetto emergente può interferire e contrastare l'azione.

Segregazione.

Nella sua replica del modello di Schelling, Gilbert (2001) fornisce un esempio di emergenza di secondo ordine che rafforza l'effetto emergente (in questo caso, il *raggruppamento*). Questo accade perché la credenza fornisce nuove linee guida per l'azione, ad esempio, "spostati solo se ci sono zone in cui sarai più felice ". La nuova credenza rafforza l'effetto macro-sociale (un più forte effetto di "*cluster*"), nella misura in cui consente una più efficiente soddisfazione della regola locale (la regola della felicità). Il collegamento tra la nuova credenza e il conseguente adeguamento della regola influenza le dinamiche dell'intero sistema. L'effetto macro-sociale è rafforzato dal modello mentale che include la nuova credenza e l'esecuzione della regola. Con questo tipo di emergenza di secondo ordine, la creazione del *cluster* è implementata dalla regola generatrice. In questo modo, Gilbert ha dimostrato come e perché l'emergenza di secondo ordine possa influenzare a sua volta la dinamica del sistema globale, e trasformarlo in un *loop* complesso bidirezionale micro-macro.

Situazione di stallo.

Consideriamo il famoso *witness effect* (effetto testimone) scoperto da Latané e Darley (1970) in situazioni di emergenza sociale. Una grande quantità di prove sperimentali e di dati osservati mostra che la probabilità di un intervento in situazioni di questo tipo crolla drammaticamente quando il numero degli astanti è maggiore di *tre*.

Perché? Gli autori hanno presentato una spiegazione piuttosto elegante, secondo la quale una regola della maggioranza (ad esempio, controllare cosa stia facendo la maggioranza in condizioni incerte) porta ad una situazione di stallo quando esiste una maggioranza, vale a dire quando gli astanti sono almeno tre. In tali condizioni, dal momento che ciascuno controlla ciò che la maggioranza sta facendo, nessuno si muove. I partecipanti sono congelati nel ruolo di testimoni. L'effetto testimone fornisce un chiaro esempio di emergenza: anche se nessuno tende alla realizzazione di questo obiettivo, l'effetto è generato dalla regola di maggioranza, proprio come la segregazione è stata generata dalla regola della felicità. Un frammento del processo generativo è emergente. Per vedere questo, si consideri che gli agenti possono essere addestrati ad evitare l'effetto testimone, che è un effetto fortemente indesiderabile dal punto di vista sociale, semplicemente venendo a conoscenza dell'esistenza di esso. Inoltre, questo esempio dimostra che un effetto emergente può modificare il meccanismo che l'ha prodotta al livello inferiore, senza essere riconosciuto dal sistema che lo ha generato. Nel nostro esempio, la situazione di stallo rafforza la regola locale: maggiore è la probabilità della situazione di stallo (macro-effetto), più forte è la regola locale (regola della maggioranza): meno sono gli agenti e minore sarà la probabilità che siano in grado di interrompere l'effetto. L'effetto testimone ha retroagito sul sistema che lo ha prodotto rafforzando temporaneamente la regola.

In questo processo di implementazione, gli agenti non hanno alcuna idea di ciò che è in corso. Tutto ciò che possiamo dire è che l'effetto testimone viene implementato su una regola a livello dell'agente, cioè la regola di maggioranza, che viene rafforzata dall'effetto in questione, mentre allo stesso tempo lo sta producendo. Vi è una temporanea modifica, vale a dire un rafforzamento della regola locale, prodotto dall'effetto emergente macro-sociale.

4.3 Vantaggi del presente approccio

L'attuale modello cerca di contribuire allo studio del *link* micro-macro, e più

specificamente ad una prospettiva generativo di questo processo. Il paradigma generativo svolgerà un ruolo decisivo per i futuri sviluppi delle scienze sociali, come indicato da alcuni evidenze:

- recente formulazione del paradigma generativo per le scienze sociali (Epstein 2006)
- rapido sviluppo di metodologie generative per lo studio dei fenomeni sociali (simulazione sociale *agent-based*)
- continua crescita di strumenti di simulazione e piattaforme (da librerie *swarm* a linguaggi *logo)
- l'accessibilità di tali linguaggi e strumenti a programmatori non esperti.

Tuttavia, la scienza sociale generativa è ancora formulata in un modo un po' insoddisfacente, ossia come un processo "*bottom-up*" (si veda nuovamente Epstein, 2007, ma più in generale, la stragrande maggioranza di simulazione e modelli computazionali dei processi economici e sociali). Con l'eccezione del *tribute model* di Axelrod (1995), il "*bottom*" è di solito dato (dal programmatore) nel resto dei modelli. Pertanto, la nozione di emergenza è di solito intesa come un processo unidirezionale.

In effetti, questo concetto è stato sostituito da quello di generazione. La presente analisi può contribuire a sviluppare:

- una teoria dell'emergenza come distinta dalla generazione;
- la visione del *link* micro-macro come *loop ricorsivo*, in cui gli effetti emergenti a livello macro retroagiscono sui livelli inferiori, modificandoli, così da fornire una visione più dinamica, generativa delle entità del livello micro.

5. Realizzare un modello computazionale di agenti (autonomi) normativi[4]

Nei precedenti capitoli abbiamo chiarito che esistono due principali approcci per lo studio delle norme e la realizzazione di modelli (computazionali) di fenomeni sociali normativi: l'approccio legato alla consolidata tradizione della *game theory* e l'approccio ispirato ai sistemi multi agente (*multi agent systems* - MAS). Abbiamo visto come questi due approcci fino ad oggi sono stati essenzialmente considerati alternativi e incompatibili e come ciascuno di essi si sia concentrato sullo studio di un particolare aspetto del *problema norma*: nello specifico, il primo si è concentrato sullo studio degli aspetti evolutivi legati all'emergenza di una particolare norma sociale in un gruppo di agenti; il secondo ha cercato di gettare luce sui meccanismi mentali individuali che sono coinvolti nelle scelte comportamentali, una volta che le norme sociali siano già state acquisite.

L'approccio cui fa riferimento il modello computazionale presentato in questo capitolo rappresenta il tentativo di unificare queste due tradizioni, proponendo una teoria *cognitiva* dell'emergenza di una norma in un gruppo di agenti. Questa teoria prevede che l'emergenza di una norma sociale in un gruppo di individui sia possibile solo dopo che questi ultimi si siano formati delle credenze normative relative all'esistenza di una norma; queste credenze devono poi trovare una conferma :

i) nell'osservazione del comportamento altrui,

ii) nella ricezione di espliciti messaggi (ad esempio verbali) da parte di altri individui.

Questa concezione della norma come oggetto a due facce, una *privata* (il processamento mentale delle credenze normative) ed una *pubblica* (la diffusione delle credenze normative e la possibile conseguente convergenza da parte degli agenti su una di esse) trova la sua implementazione in una complessa architettura cognitiva. Un aspetto cruciale di questa

[4] Per gli argomenti qui di seguito trattati, si veda: Conte, R., Andrighetto, G., Campennì, M. (2009) "The Immergence of Norms in Agent Worlds", *Lecture Notes in Artificial Intelligence, LNAI*, Vol. 5881 – H. Aldewereld, V. Dignum, G. Picard (Eds.); Andrighetto, G., Campennì, M., Cecconi, F., Conte, R. (2008) "How Agents Find out Norms: A Simulation Based Model of Norm Innovation", *3rd International Workshop on Normative Multiagent Systems (NorMAS 2008)*, Luxembourg 15-16 July; Campennì, M., Andrighetto, G., Cecconi, F., Conte, R. (2009) "Normal = Normative? The Role of Intelligent Agents in Norm Innovation", *Mind & Society*, Vol. 8, No. 2, pp. 153-172, Springer Berlin / Heidelberg.

architettura è costituito da un particolare *modulo adibito al riconoscimento delle norme*: il modulo permette all'agente che lo possiede di interpretare un comportamento osservato come potenzialmente normativo (dettato dal rispetto di una norma); la credenza normativa (candidata) che il modulo permette di generare deve però poi trovare conferma nel comportamento degli altri per poter diventare "vera" credenza normativa e influenzare da quel momento il comportamento dell'agente che la possiede. In questo senso il modulo di riconoscimento rappresenta la "porta" attraverso la quale si può accedere alla *vita mentale normativa* degli agenti.

In questo capitolo introdurremo il concetto di modello computazionale di agente normativo e presenteremo una breve rassegna degli approcci esistenti in questo settore di ricerca; chiariremo il concetto di innovazione normativa; discuteremo dei processo inter-agenti e di quelli intra-agente; presenteremo l'architettura (EMIL-A) sviluppata nel corso del progetto europeo EMIL. Infine, mostreremo i risultati di alcuni esperimenti simulativi tesi a confermare le assunzioni teoriche che sono alla base della nostra ricerca.

Per verificare l'efficacia del modulo di riconoscimento normativo, abbiamo fatto girare diverse simulazioni *agent-based*, in cui gli agenti interagiscono molto semplicemente, in un mondo ipotetico elementare.

Questo modello molto astratto serve allo scopo di verificare:

i. se vi sono differenze cruciali a livello della popolazione, tra i conformisti sociale (SCs), il cui comportamento è determinato solo per imitazione, e i riconoscitori di norme (NRs), il cui comportamento è invece determinato da un modulo di riconoscimento normativo;
ii. se e in che misura la capacità di riconoscere una norma e di generare (nuove) credenze normative è un requisito preliminare indispensabile per l'emergenza di una norma.

5.1 Verso un'architettura di agente normativo

Lo studio delle norme che adotta un approccio multiagente attinge al trattamento delle norme così come sono viste nelle scienze sociali e si concentra sull'impatto delle norme sui comportamenti degli agenti e sulla comparsa di norme di comportamento.

In questo campo, le norme sono artefatti che hanno a che fare con il coordinamento, la

cooperazione e la sicurezza (Shoham and Tennenholtz 1992, Conte and Castelfranchi 1995, Walker and Wooldridge 1995) e sono utilizzate per modellare questioni giuridiche nel settore delle istituzioni elettroniche e del commercio elettronico, per realizzare modelli multi-agente di organizzazioni, e così via (per un'introduzione sui sistemi multiagente normativi, Boella 2006).

Il dominio di ricerca multiagente normativo ha assorbito anche i vari risultati ottenuti nella formalizzazione di concetti normativi, dalla logica deontica (Henrik 1963, Eduardo and Eugenio 1971), alla teoria della posizione normativa (Lindahl 1977), alla dinamica dei sistemi normativi (Eduardo 1985). In particolare, questi studi hanno fornito sistemi multiagente normativi con una analisi formale delle norme, che è essenzialmente basata sulla distinzione tra idealità e realtà (Jones and Porn 1985), dando così conoscenze fondamentali per rappresentare e ragionare sulle norme.

Tuttavia, questo dominio va al di là delle relazioni logiche tra gli obblighi, i permessi e i divieti spiegando il rapporto tra le norme sociali e le norme giuridiche e indagando come le norme si evolvono, si fondono e cambiano, come gli agenti interagiscono con le norme, come le violano e molto altro ancora.

Nonostante l'indubbia rilevanza dei risultati ottenuti, alcune domande fondamentali sono ancora irrisolte, come ad esempio: come e dove le norme sono originate? Come spiegare l'origine di ogni nuova norma sociale? Come fanno gli agenti ad acquisire le norme? E, più specificamente, come e perché gli agenti ritengono che qualcosa sia una norma, in modo da memorizzarla nella loro memoria normativa?

La nostra sensazione è che la questione su come le norme vengono create ed innovate è un aspetto che non ha ancora ricevuto finora una spiegazione soddisfacente. Noi affermiamo che questa circostanza può essere attribuita al modo in cui l'agente normativo è stato modellato fino ad ora.

Da un lato, nell'ambito delle scienze sociali formali (precisamente in teoria dell'utilità e in teoria dei giochi – Bicchieri 2006, Epstein 2006, Sen and Airiau 2007, Ullman-Margalit 1977, Young 1998), la diffusione delle nuove norme e di altri comportamenti cooperativi di solito non sono spiegati per mezzo di modelli di rappresentazioni interne delle norme. Oggetto di indagine sono di solito le condizioni per la convergenza degli agenti su certi comportamenti: tale approccio si è rivelato efficace nel risolvere i problemi di coordinamento

(Lewis 1969) o di cooperazione (Axelrod 1987), indipendentemente dalle credenze degli agenti e dai loro scopi (Binmore 1994); tuttavia, nessuna teoria sull'acquisizione di atteggiamenti normativi fondata sulle rappresentazioni interne degli agenti è ancora stata fornita.

D'altra parte, nei sistemi multiagente (Dignum 1999, Jones and Sergot 1996, Van der Torre and Tan 1999) le norme sono esplicitamente rappresentate, ma sono già in atto come oggetti mentali *built-in* nelle menti degli agenti. Questo approccio si è concentrato sulla questione relativa al perché gli agenti siano conformi alle norme e come sia possibile che le norme operino su agenti autonomi intelligenti.

Recentemente, il processo decisionale nei sistemi normativi e il rapporto tra i desideri e gli obblighi è stato studiato nell'ambito delle architetture BDI (*Beliefs-Desires-Intentions*), sviluppando una variante interessante di esse, vale a dire la cosiddetta *Beliefs-Obligations-Intentions-Desires* o architettura BOID (Broersen et al. 2001). Questa architettura consiste di un meccanismo di anelli di retroazione, considera tutti gli effetti delle azioni prima di commetterle e risolve i conflitti tra le uscite dei suoi quattro componenti (B, O, I, D). In BOID, così come nel classico BDI, non è previsto che un agente possa (o non possa) riconoscere un dato input, come una nuova norma. Al contrario, gli obblighi sono programmati nelle menti degli agenti quando il sistema è *off-line*.

A differenza del modello game teorico degli agenti normativi, l'approccio a sistemi multiagente presenta certamente tutti i vantaggi che derivano da un rappresentazione esplicita delle norme. Tuttavia, possiamo affermare che questo approccio ha alcuni limiti, che non presentano una rilevanza solo teorica, ma anche pratica e implementativa.

Prima di tutto, si mette in ombra uno dei vantaggi degli agenti autonomi, vale a dire la loro capacità di filtrare le richieste esterne. Una tale capacità di filtraggio riguarda le decisioni, non solo normative, ma anche l'acquisizione di nuove norme. Infatti, gli agenti prendono decisioni, anche quando decidono di formarsi credenze normative, e quindi nuovi scopi (normativi), e non solo quando decidono di eseguire o meno la norma (Conte et al. 1998).

Per quanto riguarda la rilevanza pratica, se gli agenti sono in grado di acquisire nuove norme, non vi è alcuna necessità di ampliare eccessivamente la loro base di conoscenze, in quanto essi possono essere ottimizzati quando sono *on-line* e non solo quando il sistema è *off-*

line (Shoham and Tennenholtz 1992).

In un lavoro successivo, Shoham e Tennenholtz (1994) hanno introdotto il concetto di *co-apprendimento*, che si riferisce ad un processo in cui diversi agenti contemporaneamente cercano di adattare il comportamento dell'uno a quello dell'altro in modo da produrre auspicabili proprietà del sistema globale. Di particolare interesse sono due specifici contesti di co-apprendimento, che si riferiscono, rispettivamente, alla nascita di convenzioni e all'evoluzione della cooperazione nelle società. Nonostante l'indubbia importanza di questo lavoro (il trattamento di norme come convenzioni emergenti derivanti da processi di co-apprendimento può essere utile per spiegare come azioni preesistenti sono gradualmente generalizzate o scartate), l'approccio non può spiegare il processo di accettazione di nuove norme sancite da parte di un'autorità e, più in generale, quei casi in cui venga selezionata una norma che prescrive una certa azione, che nessuno ha mai eseguito prima.

Per modellare e rendere operativo il processo di immergenza di una norma, gli agenti autonomi ed intelligenti devono essere dotati di meccanismi interni e di rappresentazioni mentali che permettano alle norme di incidere sui loro comportamenti. Tali rappresentazioni sono comunemente realizzate da architetture ispirate ad un'architettura modulare, propria dell'approccio dell'intelligenza artificiale (AI) classica. Al giorno d'oggi, non esiste un concetto univoco per la progettazione di agenti normativi. Lo sviluppo di architetture normative è un settore di ricerca in forte crescita. Tuttavia, le architetture attraverso cui si implementano gli agenti normativi sono prevalentemente ispirate, in un modo o nell'altro, ad architetture BDI (*Belief-Desire-Intention*), una "famiglia" di architetture introdotta dal lavoro fondamentale di Rao e Georgeff (Rao and Georgeff 1992) e che può essere considerata come il punto di partenza per ulteriori sviluppi.

L'approccio BDI è destinato ad implementare un modello di azione intelligente e di decisione dell'uomo. Come abbiamo ricordato, un esempio particolarmente chiaro di questo approccio è fornito da una semplice estensione di architettura BDI al ragionamento normativo, indicato come architettura BOID (*Beliefs-Obligations-Intentions-Desires*) (Broersen et al. 2001) che, rispetto ad una semplice architettura BDI, comprende anche gli obblighi fra gli oggetti mentali trattati.

L'architettura normativa che presentiamo in questa sezione, Emil-A, è ispirata a BOID, in quanto anch'essa comporta la rappresentazione di credenze normative e di scopi sulla base

di obblighi. Tuttavia, a differenza di BOID, Emil-A include un modulo per il riconoscimento delle norme, permettendo all'agente di elaborare gli input che riceve ed eventualmente di poterli convertire in norme (nuove credenze normative).

L'obiettivo del presente capitolo è quello di proporre un modello computazionale di riconoscimento autonomo di una norma e di testare la sua efficacia nel contesto dell'emergenza di una norma e della innovazione di una norma attraverso una serie di simulazioni basate su agenti. Noi riteniamo che una capacità autonoma di riconoscimento di una norma potrebbe migliorare notevolmente la flessibilità e le potenzialità dinamiche dei sistemi multiagente. La capacità di generare credenze normative in base ad input esterni (siano essi il risultato di una osservazione o di una comunicazione) rende un agente (sociale) autonomo più adattabile alle condizioni sociali che incontra.

5.2 Innovazione normativa

Le norme sono artefatti dotati di grande capacità di adattamento, che emergono, evolvono, decadono.

Se è relativamente chiaro come possano esistere delle norme giuridiche, è molto meno evidente come lo stesso processo possa riguardare le norme sociali. Come fanno nuove norme e convenzioni sociali a nascere?

Sono stati effettuati alcuni studi simulativi sulla selezione delle convenzioni: per esempio, lo studio di Epstein e colleghi sulla comparsa di norme sociali (Epstein 2006), o lo studio di Sen ed Airiau sull'emergere di una regola di precedenza nel traffico (Sen and Airiau 2007). Tuttavia, tali studi indagano quale comportamento sia scelto da una serie di equilibri alternativi (possibili). Una accezione piuttosto diversa della questione riguarda l'innovazione delle norme sociali quando non sono disponibili per la selezione equilibri alternativi. Questo è un argomento ancora poco indagato e i riferimenti sono scarsi (quando presenti). Per esempio, Posner e Rasmusen (Posner and Rasmusen 1999) trattano la creazione e la distruzione delle norme, ma con particolare riferimento alle sanzioni. Il problema non è certo privo di interesse, ma per l'obiettivo del presente capitolo preferiamo concentrarci solo sulla ricerca degli indizi che portano un agente ad interpretare un dato comportamento sociale come normativo, mettendo per il momento da parte le questioni relative alle sanzioni e ai meccanismi di applicazione.

Proponiamo che una possibile risposta alle difficili domande qui poste sopra potrebbe essere trovata esaminando l'interazione fra i comportamenti comunicati e quelli osservati, e il modo in cui questi sono rappresentati nella mente degli osservatori. Se un qualsiasi nuovo comportamento *alpha* viene interpretato come concernente una norma, una nuova credenza normativa viene generata nella mente dell'agente e un processo di influenza normativa sarà attivato (Conte and Dignum 2001)[5]. Un tale comportamento avrà maggiori probabilità di essere replicato rispetto al caso in cui nessuna credenza normativa si sia formata (Andrighetto et al. 2007). Come mostrato altrove (Conte and Castelfranchi 1999, Andrighetto et al. 2007), quando un agente dotato di credenza normativa replica il comportamento *alpha*, influenza gli altri a fare lo stesso, non solo apparentemente mostrando il comportamento in questione, ma anche esplicitamente trasmettendo una norma (la norma relativa al comportamento in questione). Le persone si impongono l'un l'altra nuove norme, per mezzo di deontici ed esplicite valutazioni normative.

Quindi, proponiamo che il riconoscimento normativo rappresenti un requisito fondamentale per la innovazione normativa, come processo risultante sia dalle interpretazioni degli agenti degli altrui comportamenti, che dalla reciproca trasmissione di tali interpretazioni.

5.3 EMIL-A: un'architettura normativa

Consideriamo una norma come un comportamento sociale che si diffonde attraverso la popolazione, grazie alla diffusione di una credenza, cioè la credenza normativa. Una credenza normativa, a sua volta, è la *credenza che un dato comportamento, in un dato contesto, per un determinato insieme di agenti, sia proibito, obbligatorio o permesso*. Così, affinché un comportamento normativo abbia luogo, una credenza normativa deve essere generata nella mente dei destinatari della norma e il corrispondente scopo normativo deve essere formato e perseguito. La nostra tesi è che una norma emerge come una norma solo quando è *incorporata* nelle menti degli agenti coinvolti (Conte and Castelfranchi 1995, Conte and

[5] Altrove (Andrighetto et al. 2007), abbiamo proposto una tassonomia operativa della innovazione normativa. Abbiamo previsto almeno tre possibili tipi di innovazione normativa, da poco a fortemente innovativa: l'estensione o adattamento di una norma (già esistente); l'istanza di una norma, l'integrazione di norme diverse. Noi affermiamo che l'innovazione normativa ha luogo a partire dalle norme vigenti e che non nasce dal nulla. Più in particolare, consideriamo una nuova norma come un'estensione, o un particolare tipo di evoluzione, vale a dire un'istanza o l'integrazione di norme preesistenti.

Castelfranchi 2006); in altre parole, quando gli agenti la riconoscono come tale. In questo senso, la nascita e la stabilizzazione di una norma implica la sua *immergenza* (Castelfranchi 1998) nelle menti degli agenti.

La nostra architettura normativa Emil-A (si veda figura 1; Andrighetto et al. 2007, per una descrizione dettagliata) è stata pensata per mostrare che le norme non solo regolamentano il comportamento, ma agiscono anche su diversi aspetti della vita mentale degli agenti. Emil-A è costituita da meccanismi e da rappresentazioni mentali che permettono agli agenti:

i. di formarsi credenze e scopi normativi, e decidere se realizzare o meno tali scopi;

ii. di essere più o meno reattivi agli input esterni per mezzo di short cuts (associazioni mentali che favoriscono reazioni particolarmente rapide).

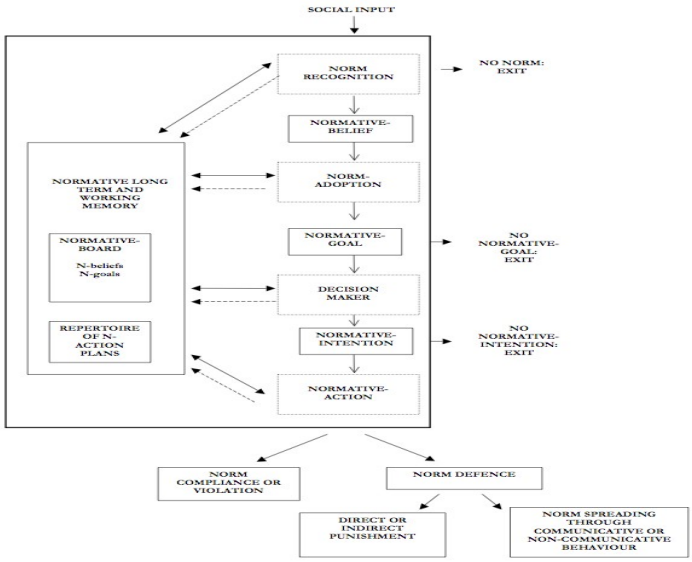


Figura 1. EMIL-A

Ad EMIL-A si accede tramite il modulo di riconoscimento delle norme (vedi figura 2);

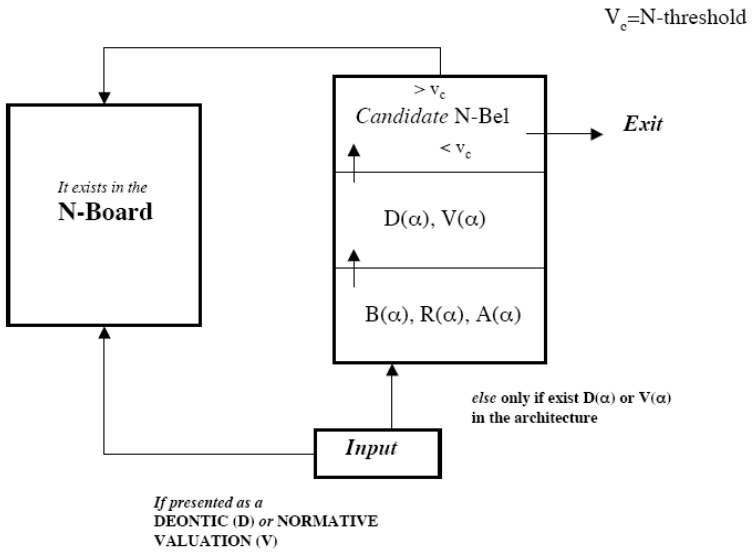


Figura 2. Il modulo di riconoscimento delle norme in EMIL-A

prima che un input sia riconosciuto come normativo, la norma non può immergere nelle menti degli agenti e, di conseguenza, non può incidere sui loro comportamenti ed emergere nella società. Riteniamo che le architetture normative esistenti non siano sufficientemente flessibili ed adattabili per essere realmente plausibili; crediamo che il futuro delle architetture normative sia strettamente legato allo sviluppo di architetture ibride (architetture in cui convivano componenti reattive - potremmo dire senso-motorie, subsimboliche - e componenti razionali simboliche).

RAPPRESENTAZIONI MENTALI NORMATIVE. In questa sezione, cercheremo di chiarire alcuni componenti coinvolte nel processo di elaborazione mentale delle norme.

Credenza Normativa[6]. Prima di tutto, una norma diventa una credenza, cioè la convinzione che un determinato comportamento, in un dato contesto, per un determinato insieme di agenti sia vietato, obbligatorio, permesso, ecc. Più precisamente, la credenza dovrebbe essere che "vi sia una norma che vieta, prescrive, permette che ..." (Henrik 1963, Kelsen 1979, Conte and Castelfranchi 1999, Conte and Castelfranchi 2006). Infatti, le norme

[6] In EMIL-A le credenze normative, insieme con gli scopi normativi, sono organizzati e disposti nella lavagna normativa in base alla loro rispettiva salienza. In base alla salienza si determina il grado di attivazione della norma, che è funzione del numero di volte in cui una determinata norma entra in gioco nel processo di decision-making dell'agente.

sono volte a e rilasciate per generare le corrispondenti credenze normative. In altre parole, le norme devono essere riconosciute come tali al fine di funzionare correttamente. Naturalmente, una credenza normativa non implica che una data norma sia stata in realtà deliberatamente emessa da alcuni sovrani. Le norme sociali sono spesso istituite in virtù di effetti indesiderati (e non previsti, come per esempio una erronea interpretazione di un altrui comportamento). Tuttavia, una volta emersa, una data norma sociale si crede che sia basata su una qualche autorità normativa, se solo si fonda su un anonimo e impersonale destinatario ("Ti è richiesto o ci si aspetta che tu (non) faccia questo..."; " In generale, si prevede che..."; "Così è come vanno le cose...")

Scopo Normativo. Qualcuno che crede non è ancora qualcuno che decide: le credenze sono condizioni necessarie ma non sufficienti affinché una norma sia rispettata. Che cos'è che induce gli agenti ad accettare una norma, che per definizione prevede un comportamento costoso?

Nell'approccio BDI le intenzioni e le azioni provengono solo dai desideri. Al contrario, una grande quantità di nostre azioni non sono suscitate dai nostri desideri, ma dalle pressioni esterne e dalle richieste. I compiti e le norme sono una delle fonti esterne dei nostri scopi. Come è possibile? Come è possibile generare scopi normativi?

Da un punto di vista cognitivo, gli scopi sono rappresentazioni interne che contemporaneamente scatenano e guidano le azioni: essi rappresentano lo stato del mondo che gli agenti vogliono raggiungere attraverso l'azione e di cui effettuano il monitoraggio durante l'esecuzione dell'azione (Conte 2009). Sotto l'effetto di fattori sociali, gli scopi possono essere nuovamente generati attraverso fattori cognitivi, presentandosi come scopi relativizzati ad altri stati mentali (ad esempio, le convinzioni sociali). Uno scopo è relativizzato quando è perseguito in quanto e nella misura in cui un determinato stato o evento del mondo è atteso o previsto essere vero (Castelfranchi 1999).

Quando gli scopi sono positivi o pro-sociali, il processo di generazione è chiamato *adozione-di-scopo* (Conte and Castelfranchi 1995).

Sembra che esista una corrispondenza tra:

1. il processo che porta da una credenza circa una richiesta ordinaria alla decisione di accettare tale richiesta (vale a dire il processo di adozione di uno scopo sociale)

e

2. il processo che porta da una credenza normativa ad uno scopo normativo (adozione di una norma);

uno scopo normativo di un determinato agente ***x*** circa un'azione ***a*** è uno scopo che ***x*** avrà per tutto il tempo in cui avrà una credenza normativa circa ***a***. Più specificamente, ***x*** ha uno scopo normativo solo se crede di essere soggetto ad una norma.

Modulo per riconoscere le norme. Il modulo di riconoscimento delle norme (vedi figura 2) è l'ingresso principale, per così dire, all'architettura Emil-A. Prima che un input sia riconosciuto come normativo, la norma non può immergere nelle menti degli agenti e, di conseguenza, non può emergere nella società. Gli agenti hanno bisogno di essere in grado di discriminare tra le norme e gli altri fenomeni sociali, come la coercizione, le richieste ordinarie, le convenzioni, ecc. La nostra tesi è che le altre architetture normative non rendono giustizia alla procedura di riconoscimento (Andrighetto et al. 2008, Campennì et al. 2009).

Semplificando, una data norma è riconosciuta come tale se

- l'*input* corrente corrisponde a una norma già memorizzata nella nostra memoria (normativa);
- l'agente è in grado di inferire o dedurre l'esistenza della norma in base ai dati ricevuti in *input*.

Nel primo caso, l'agente è facilitata da schemi, script, o altre strutture pragmatiche (Wason and Johnson-Laird 1972, Schank and Abelson 1977, Fiske and Taylor 1991, Barsalou 1999) (Markus and Zajonc 1985, per una panoramica), in cui la norma è inserita (Broersen et al. 2001, per una descrizione). Una volta che questi vengono attivati per qualsiasi motivo, le credenze normative corrispondenti, le aspettative e le regole comportamentali vengono attivate.

La seconda opzione è seguita quando tali *script*, e di conseguenza il corrispondente modello di *pattern matching*, non esistono. L'agente non ha alcuna norma corrispondente. Questo è il motivo per cui il modulo di riconoscimento normativo è necessario. Infatti, il modulo che ci accingiamo a descrivere cerca di rispondere alla domanda su come gli agenti possano comunicare (ad altri o a se stessi) nuove norme, non ancora memorizzate nella memoria (Sripada and Stich 2006). “Raccontare” le norme implica la capacità degli agenti di

considerare input sociali osservati o comunicati (vedi figura 3) come normativi, e quindi di formarsi una nuova credenza normativa.

Emil-A è un modulo di riconoscimento normativo costituito da un architettura normativa grazie alla quale gli input ricevuti vengono elaborati ed interpretati, e una memoria a lungo termine - chiamata lavagna normativa (normative board - N-Board) - in cui le credenze normative e gli scopi normativi, una volta formati vengono memorizzati e classificati in base alla loro salienza.

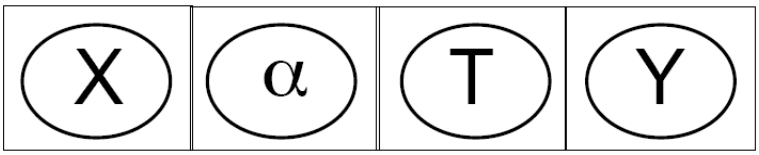


Figura 3. Input sociale

La lavagna normativa (normative board - N-Board). Quando Emil-A ha a che fare con un ingresso esterno, come un cartello di divieto di fumare, il modulo di riconoscimento normativo consulterà la N-Board. Supponiamo che una credenza normativa corrispondente venga individuata (Non fumare quando è PROIBITO); in questo caso una credenza normativa verrà "attivata" e seguirà il percorso descritto in precedenza.

La N-Board è un archivio nella memoria a lungo termine in cui le norme attive sono memorizzate, disposte secondo la salienza acquisita. Differenze di salienza hanno l'effetto che un sottoinsieme di rappresentazioni legate a norme interferiscono più di frequente e fortemente delle altre con i processi cognitivi generali dell'agente. Per decidere l'azione da eseguire, l'agente effettua una ricerca nella N-Board: se trova più di una voce, la norma più saliente verrà scelta.

Se una norma non è mai adottata da parte dell'agente, la sua salienza inizia a diminuire, e prima o poi la credenza normativa decadrà. Al contrario, una norma che è spesso elaborata dal decisore, aumenterà in salienza. La salienza può aumentare fino al punto in cui una norma viene interiorizzata, cioè trasformata in uno scopo ordinario (quindi non necessariamente normativo), o anche in una routine (vedi capitolo 7). In tal caso, la norma sarà uscita dal N-Board.

5.3.1 Il valore aggiunto di EMIL-A

Finora, lo studio relativo all'emergenza di una norma è stato identificato con lo studio delle regolarità comportamentali. Tuttavia, non tutte le regolarità sono obblighi, e non tutte le norme sono rispettate. Quindi, la priorità logica e pragmatica è *come gli agenti scoprono quali sono le regolarità normative*. Solo dopo, ha senso parlare di come realizzare un modello delle ragioni per le quali conformarsi a queste regolarità. Il valore aggiunto di Emil-A è quello di tenere conto di questo aspetto specifico del processo normativo, cioè di come gli agenti scoprono le norme rispetto alle quali decidono di conformarsi (o meno).

Il riconoscimento normativo è un requisito importante per l'emergenza di una norma.

In lavori precedenti (Castelfranchi 1998, Conte et al. 2007), l'emergenza è stata definita come una dinamica progressiva e complessa con la quale il macro-effetto sociale, nel nostro caso una specifica norma, si realizza nella società, mentre immerge nella mente degli agenti, generandosi attraverso una serie di anelli intermedi.

A differenza delle disposizioni morali, il riconoscimento normativo è poco sensibile alla variabilità soggettiva, e piuttosto robusto. Esso ci permette di:

(a) trattare l'aspetto universale delle norme tipiche delle società umane e dei primati;

(b) rendere giustizia dell'intuizione che gli esseri umani violano le norme, ma hanno dei problemi nel trovarne le verifiche;

(c) trattare le evidenze prodotte dalla psicologica evolutiva (Cosmides and Tooby 1992, Cosmides and Tooby 2008) rispetto al fatto che gli agenti applichino facilmente il ragionamento controfattuale alle regole sociali, ma non riescano a fare lo stesso con altrettanta facilità con quelle logiche;

(d) spiegare perché, come rilevato dai dati della psicologia dello sviluppo, l'acquisizione di una norma segue un modello ontogenetico stabile che viene attivato molto presto durante l'infanzia (Bandura 1991, Cummins 1996, Piaget 1965, Kohlberg 1981, Kohlberg and Turiel 1971, Shweder et al. 1987).

In breve, l'intuizione che sta dietro la nostra architettura normativa è duplice: da un lato, l'emergere di norme si basa su una capacità universale di “poter comunicare” norme; dall'altra, questa capacità è supportata da una cornice normativa, un "modello di una norma" interno, che gli agenti utilizzano come strumento di riconoscimento di una norma.

L'accento posto sulle caratteristiche innate e universali di Emil-A non deve trarre in

inganno, portando a pensare che nessuno spazio sia lasciato alla variabilità soggettiva. Se il riconoscimento normativo è un dovere ugualmente condiviso da una stragrande maggioranza degli agenti, gli atteggiamenti morali - cioè i risultati delle esperienze normative e morali accumulati durante la vita che colpiscono diverse procedure normative - non lo sono. Sono sicuramente soggettivi.

Inoltre, gli effetti di rinforzo che si verificano in diverse procedure di EMIL-A variano da agente ad agente. L'esperienza personale, ad esempio, ha delle ripercussioni sulla salienza di una norma. Analogamente, l'architettura normativa, essendo in costante interazione con l'ambiente sociale e le altre procedure (non necessariamente normative), rischia di subire la loro influenza. In questi termini, un'architettura normativa può elegantemente ignorare la polemica cultura/educazione.

5.3.2 Il modulo che riconosce le norme

La nostra architettura normativa (EMIL-A) (Andrighetto et al. 2007 per una descrizione dettagliata) consiste dei meccanismi e delle rappresentazioni mentali che permettono alle norme di incidere sul comportamento di agenti autonomi intelligenti. Emil-A è stata pensata per mostrare che le norme non solo regolano il comportamento, ma agiscono anche su diversi aspetti della mente. In sostanza, Emil-A è una architettura BDI dotata di un modulo di riconoscimento delle norme. Rispetto alle architetture BOID in cui gli obblighi sono già nella mente degli agenti, Emil-A è dotata di una componente per mezzo della quale gli agenti riescono a dedurre che una determinata norma è in vigore anche se non è già memorizzata nella memoria normativa.

In questa situazione la norma non è già inserita in schemi, script, o altre strutture pragmatica (Bicchieri 2006), quindi, gli agenti non sono facilitati da uno di essi. Come avviene nella realtà, la norma deve essere prima scoperta, e solo successivamente, memorizzata.

L'implementazione di tale capacità è condizionata alla modellazione della capacità degli agenti di riconoscere input sociali osservati o comunicati come normativi e di conseguenza di potersi formare una nuova credenza normativa.

Descriviamo ora più in dettaglio il primo componente di Emil-A, vale a dire il modulo di riconoscimento delle norme, in quanto questo è quello più rilevante rispetto alla questione

che abbiamo sollevato, ossia al *come* una nuova norma sia scoperta; il che è ovviamente centrale rispetto al tema dell'emergenza, dell'innovazione e della stabilizzazione delle norme. Inoltre, il modulo di riconoscimento normativo, per il momento rappresenta il solo contributo originale della nostra architettura normativa; per quanto riguarda gli altri componenti, essi sono stati modellati sulla base di una architettura normativa BDI.

Il nostro modulo di riconoscimento normativo (vedi fig. 2, p. 53) è costituito da due strati (o livelli) e da un link alla "lavagna" normativa, che fa parte della memoria a lungo termine dell'agente. La lavagna normativa contiene credenze normative e scopi normativi, organizzati in base alla loro *salienza*. Con il termine "salienza" si fa riferimento al grado di attivazione di una norma: in una particolare situazione, una norma può essere più frequente rispetto ad altre: diremo che la sua salienza è maggiore. La differenza di salienza tra credenze normative e scopi normativi ha l'effetto che alcuni di questi oggetti mentali normativi saranno più attivi di altri e potranno interferire con maggiore frequenza e con più forza con i processi cognitivi generali dell'agente.

Nel livello superiore di questo modulo, sono immagazzinate le azioni presentate come deontici (cioè quelle azioni che hanno in sé la ragion d'essere: obblighi, divieti, permessi); nello strato inferiore, invece, l'azione viene memorizzata solo se è già stata memorizzata anche nel livello superiore; vale a dire, se è stata ricevuta dall'agente come un deontico. Allo scopo di decidere quali azioni produrre, l'agente effettua una ricerca nella lavagna normativa: se è stata scoperta più di una credenza normativa, l'azione eseguita sarà quella associata alla norma più saliente. Una volta ricevuto l'input, l'agente calcola le informazioni al fine di generare / aggiornare le sue credenze normative. Ogni volta che un messaggio contenente un deontico viene ricevuto, l'azione relativa verrà memorizzata come (potenziale) norma. Questo affinerà l'attenzione dell'agente: ulteriori messaggi con lo stesso contenuto, soprattutto se osservato come comportamento, saranno trattati e conservati allo stesso livello. Una volta superata una certa soglia (soglia normativa – NT, *Normative Thrashold*), una nuova credenza normativa verrà generata.

5.4 Simulare l'emergenza normativa

Abbiamo già accennato al fatto che alcuni studi di simulazione per la creazione di norme sociali sono già stati effettuati; ad esempio: da Epstein e colleghi sull'emergenza di

norme sociali (Epstein 2006); da Sen e Airiau sull'emergere di una regola di precedenza nel traffico (Sen and Airiau 2007).

In questi studi, le norme sociali sono essenzialmente viste come convenzioni, cioè, conformità comportamentali che non implicano accordi espliciti tra gli agenti, e che emergono dai loro interessi individuali. All'interno di questa prospettiva, la funzione delle norme si trasforma nella possibilità fornita ai giocatori di un gioco di coordinamento di scegliere uno tra vari possibili equilibri alternativi equivalenti. Gli agenti interagiscono ripetutamente con altri agenti in scenari sociali. Tali interazioni possono essere formulate come *giochi con equilibri multipli* (Myerson 1991), che rendono incerto il coordinamento. Le norme a poco a poco emergono dalla pratica di interazioni; in sostanza, attraverso meccanismi di imitazione e di apprendimento sociale, che stabiliscono chi deve fare cosa.

Finora, gli studi simulativi sono stati utilizzati per indagare quale norma viene scelta, data una serie di equilibri alternativi. In questo quadro, gli agenti non sono dotati di menti normative, bensì di un *ragionamento strategico*. Non si presta attenzione all'immergenza di una norma, e quindi al ruolo dei meccanismi mentali coinvolti nella sua emergenza.

Una specie piuttosto diversa di domanda scientifica è quella che si chiede cosa accade all'emergenza di norme sociali quando equilibri alternativi non sono disponibili per la selezione. Questa è una questione ancora poco indagata e i riferimenti sono scarsi (Posner and Rasmusen 1999).

Proponiamo che una possibile risposta a tutte queste domande dovrebbe essere cercata esaminando l'interazione dei comportamenti comunicati e osservati, e il modo in cui questi vengono interpretati e rappresentati nelle menti degli osservatori. Eventuali nuovi comportamenti possono essere interpretati come obbedienti ad una norma: una nuova credenza normativa verrà generata nella mente dell'agente e un processo di influenza normativa verrà attivato (Conte and Dignum 2001). Suggeriamo che il riconoscimento normativo rappresenti un requisito fondamentale per l'emergenza e l'innovazione di una norma, processi derivanti sia da interpretazioni degli agenti di altrui comportamenti, che dalla trasmissione di tali interpretazioni.

5.4.1 Il modulo al lavoro

Il modulo di riconoscimento delle norme (Andrighetto et al. 2008, Campennì et al.

2009, per una descrizione dettagliata) è costituito da una memoria a lungo termine, da una *normative board* (N-Board), e da una memoria di lavoro, costituita da una architettura a tre strati. Il N-Board contiene credenze normative, classificate in ordine di salienza. La differenza di salienza tra credenze normative e scopi normativi ha questi effetti: alcuni di questi oggetti mentali normativi saranno più attivi di altri e potranno interferire con maggiore frequenza e con più forza con i processi cognitivi generali dell'agente[7].

La memoria di lavoro è una architettura a tre strati, in cui gli input sociali sono elaborati. Questi input sono rappresentati da un vettore ordinato, composto da quattro elementi:

- la fonte (x),
- il tipo di input attraverso cui il messaggio viene presentato (T)[8];
- il destinatario (y);
- l'azione trasmessa (a).

Gli agenti osservano o comunicano input sociali. Una volta ricevuto l'*input* da un altro agente, l'agente calcola, grazie al suo modulo di riconoscimento normativo, le informazioni al fine di generare / aggiornare le sue credenze normative.

Forniamo qui di seguito una breve descrizione di come funziona questo modulo normativo.

Ogni volta che si riceve un messaggio contenente un deontico (D), per esempio, "è *necessario* rispondere alla domanda", o di una valutazione normativa (V), per esempio, "è *scortese* non rispondere alla domanda", si accede direttamente al secondo livello dell'architettura, generando una credenza normativa candidata "*bisogna* rispondere alla domanda", che sarà temporaneamente conservata presso il terzo livello.

Questo sollecita l'attenzione degli agenti: ulteriori messaggi con lo stesso contenuto,

[7] Al momento, la salienza delle credenze normative non può che aumentare, a seconda del numero di istanze della stessa credenza normativa che sono memorizzate nel N-Board. Questa caratteristica ha l'effetto negativo che alcune norme diventano altamente salienti, esercitando una eccessiva interferenza con il processo decisionale dell'agente. Stiamo migliorare il modello, aggiungendo la possibilità che, se la credenza normativa è inattiva per un certo periodo di tempo, la sua salienza diminuirà.

[8] Esso può consistere sia in un comportamento (B), vale a dire un azione o reazione di un agente rispetto ad un altro agente o all'ambiente, o in un messaggio trasmesso attraverso i seguenti possibili modali: asserzioni (A), vale a dire frasi generiche che descrivono gli stati del mondo; richieste (R), cioè le richieste di intervento effettuate da un altro agente; deontico (D), espressioni di giudizio su cosa è bene / male e accettabile / inaccettabile; valutazioni normative (V), ossia affermazioni su ciò che è giusto o sbagliato, corretto o non corretto, appropriato o inappropriato (per esempio è giusto rispettare la coda).

soprattutto se questo dà luogo a comportamenti, o viene trasmesso attraverso affermazioni (A) (per esempio: "quando gli si fa una domanda, Paolo risponde "), o richieste (R) (per esempio: "potrebbe rispondere alla domanda?"), saranno processati e conservati presso il primo livello dell'architettura.

Oltre una certa soglia normativa (che rappresenta la *frequenza* dei corrispondenti comportamenti normativi osservati, ad esempio n% della popolazione), la credenza normativa candidata sarà trasformata in una nuova (vera) credenza normativa, che verrà memorizzata nel N-Board.

La soglia normativa può essere raggiunto in diversi modi: un modo consiste nell'osservare un dato numero di agenti eseguire la stessa azione (*alpha*) prescritta dalla credenza normativa candidata; ad esempio, gli agenti rispondono quando ricevono una domanda. Se l'agente non riceve altre occorrenze dello stesso input (*alpha*), dopo un tempo ***t*** fissato, la credenza normativa candidata lascerà la memoria di lavoro.

Per decidere quali azioni produrre, l'agente effettua una ricerca nella N-Board: se trova più di una voce, verrà scelta la norma più saliente.

5.4.2 Il modello

Nel nostro modello computazionale l'ambiente è costituito da quattro scenari-contesti, in cui gli agenti sono in grado di produrre tre diversi tipi di azioni. Si definiscono due azioni contesto-specifiche per ogni scenario, e un'azione comune a tutti gli scenari. Pertanto, abbiamo nove azioni.

Per capire perché, supponiamo che il contesto primo sia un ufficio postale, il secondo uno sportello informativo, il terzo la nostra casa, e così via.

Nel primo contesto, una azione contesto-specifica potrebbe essere *rispettare la fila*, mentre nel secondo potrebbe essere *occupare il posto giusto* di fronte allo sportello che interessa (magari uno sportello fornisce informazioni di un tipo, uno di un altro). Un'azione comune per tutti i contesti potrebbe essere *rispondere quando interrogati*.

Abbiamo modellato due diversi tipi di agenti: *Social Conformers* (SCs) e *Norm Recognizers* (NRs), ognuno dei quali dispone di una agenda personale (cioè una lista ordinata di contesti), un tempo individuale e costante di permanenza in ogni scenario-contesto (quando il tempo di permanenza è scaduto, l'agente si sposta al prossimo contesto) e una finestra di

osservazione per le azioni prodotte da gli altri agenti.

Inoltre, i NRs sono forniti anche di un'architettura a due strati, necessaria per analizzare le informazioni ricevute, e una N-Board, in cui le credenze normative, una volta generate, sono memorizzate.

Una volta scaduto il tempo di permanenza in uno scenario, ogni agente – sia esso un SC o un NR - si sposta allo scenario successivo seguendo la sua agenda. Tali flussi irregolari (ogni agente ha un tempo diverso di permanenza e una agenda diversa) genera un comportamento complesso del sistema, che *step-by-step* produce una definizione "fuzzy" degli scenari, e una dinamica "fuzzy" comportamentale.

5.4.2.1 I *social conformers*

Ad ogni tick, due SCs sono accoppiati in modo casuale e interagiscono. L'azione che ciascun agente produce è influenzata dalle azioni prodotte dagli ***n*** agenti che hanno agito prima di lui (tasso di conformità sociale): se c'è un'azione che è stata effettuata più spesso delle altre, l'agente imiterà tale azione. In caso contrario, sceglierà in modo casuale una tra le tre possibili azioni per lo scenario in cui si trova. Per l'obiettivo di questo lavoro, abbiamo optato per una euristica indipendente dalla fitness; non siamo infatti interessati alle prestazioni individuali degli agenti, ma a rilevare gli effetti osservabili delle diverse modalità di modellazione rispetto all'emergenza di norme sociali.

Abbiamo eseguito diverse simulazioni, facendo variare il tasso di conformità sociale.

5.4.2.2 I riconoscitori di norme

Ad ogni tick, a differenza di SCs, i NRs (accoppiati a caso) interagiscono scambiandosi messaggi. Questi input sono rappresentati da un vettore ordinato, composto da quattro elementi:

- la fonte (x);
- il modale attraverso il quale il messaggio viene presentato (M); esistono sono sei modali possibili:
    1. asserzioni (A), frasi generiche o che descrivono gli stati del mondo;
    2. comportamenti (B), le azioni o le reazioni di un agente, rispetto ad un altro

agente o all'ambiente;

3. richieste (R), le richieste di intervento effettuato da un altro agente;
4. deontico (D), espressioni di giudizio su cosa è bene / male, accettabile / inaccettabile (si distinguono ulteriormente deontici di tre tipi: obblighi, proibizioni, permessi);
5. valutazioni normative (Vn), affermazioni su ciò che è giusto o sbagliato, corretto o non corretto, appropriato o inappropriato;

– il destinatario (y);
– l'azione trasmessa (a).

Codificare l'input in questo modo ci permette di:

(a) avere accesso alle informazioni anche in seguito, se necessario,
(b) riconoscere la fonte, un pezzo di informazione che potrebbe essere utile per memorizzare gli input da parte di autorità riconosciute;
(c) rappresentare una varietà di informazioni, grazie alla sintassi dei modali,
(d) calcolare le informazioni ricevute al fine di generare una nuova credenza normativa.

I NRs producono comportamenti diversi: se il N-Board di un agente è vuoto (cioè non contiene alcuna credenza normativa), l'agente produce un'azione scelta casualmente dal set di azioni possibili (per il contesto in questione); in questo caso, anche il modale per mezzo del quale l'azione viene presentata è scelto a caso.

Viceversa, se vi sono credenze normative nel N-Board, l'agente sceglie di produrre l'azione corrispondente alla credenza normativa più saliente tra tutte quelle presenti nel N-Board[9]. In questo caso l'azione prodotta è presentata con uno di questi verbi modali: deontico (D), valutazione normativa (VN) o comportamento (B); ciò è in accordo con il fatto che se un agente ha una credenza normativa, riteniamo che (con un'alta probabilità, ad esempio, il 90%) la voglia trasmettere ad altri agenti, attraverso un modale forte (D o Vn) o un comportamento chiaro (B).

Abbiamo eseguito diverse simulazioni, variando il numero di agenti e il valore della soglia che provoca la generazione di nuove credenze normative.

[9] Per l'implementazione usiamo una distribuzione uniforme $\boldsymbol{p}$ della probabilità per scegliere la credenza normativa $\boldsymbol{x}$.

5.4.3 Il Simulatore

Abbiamo realizzato un simulatore per utenti non esperti, con una semplice interfaccia, in modo tale che l'utente può manipolare il numero degli agenti coinvolti nella simulazione, il numero di unità di tempo (cicli della simulazione), il numero di contesti e il tipo di popolazione, sia composta da SCs o da NRs. Simulazioni con popolazioni miste non sono state previste in questa fase. Altre variabili che l'utente può modificare sono il *tasso sociale della conformità* e le *routine standard* (ad esempio, salvare i risultati della simulazione su file).

Il simulatore è implementato per mezzo di funzioni *Matlab* e genera alcune finestre per i risultati.

5.4.4 I risultati sperimentali

Riassumiamo brevemente la struttura delle nostre simulazioni.

Come già detto, noi chiamiamo SCs gli agenti che effettuano un processo di imitazione e NRs gli agenti dotati di un modulo di riconoscimento normativo. Abbiamo messo a confronto una popolazione di SCs, con una popolazione di NRs.

Sia nel caso dei SCs che dei NRs, il processo inizia con la produzione di azioni (e modali nel caso dei NRs) in modo casuale, e continua con gli agenti che tentano di conformarsi.

Il processo di assimilazione è sincronico. Durante la simulazione, l'agente in posizione ***i*** fornisce gli input agli agenti in posizione ***i-1***, ***i-2***,..., ***i-k***, ed i risultati vengono assegnati immediatamente.

Se il numero di agenti prima di ***i*** non raggiunge la soglia, non avviene nessuna imitazione.

L'imitazione tra i SCs è attuata mediante un meccanismo di voto, in modo tale che l'agente ***i*** compie l'azione più frequente tra quelle che sono state eseguite dagli agenti che lo hanno preceduto.

Nel caso dei NRs, il processo è più complesso: l'agente ***i*** fornisce l'input per l'agente che lo precede (k = 1), trasmettendogli una azione e un modale.

La scelta dell'azione è condizionata dallo stato della sua N-Board.

Quando tutti gli agenti hanno eseguito un aggiornamento dello step della simulazione, il complesso processo si riavvia e si passa allo step successivo.

5.4.4.1 Risultati ottenuti per la popolazione di *social conformers*

In figura 3 (si veda alle pagine successive), si mostra la distribuzione delle azioni compiute eseguite da 100 SCs durante una simulazione di 100 tick. Sull'asse ***x*** il tempo scorre dal tempo t = 0 al tempo t = 100 corrispondente alla fine della simulazione. Sull'asse ***y*** è indicato il numero delle azioni eseguite per ogni diverso tipo di azione fra le nove a disposizione.

I risultati illustrati nella figura 4 sono molto chiari: i SCs non producono norme sociali (non convergono infatti su un'unica norma, producendo tutti la stessa azione). In effetti, possiamo vedere che non appare nessuna convergenza verso una singola azione attraverso il tempo: il tasso di convergenza non cambia in modo significativo. Il tasso di convergenza indica la convergenza degli agenti sulle singole azioni: maggiore è la convergenza su una specifica azione, maggiore è il valore di questo tasso. In figura 3, per *t = 100*, mostriamo la distribuzione delle azioni da tick = 0 a tick = 100 (cioè dall'inizio al termine della simulazione).

L'azione comune (linea tratteggiata) è la più frequente. Potremmo chiamare tale l'azione (azione 1), una norma d'*imitazione*, la cui unica caratteristica è la frequenza.

5.4.4.2 Risultati ottenuti per la popolazione di riconoscitori di norme

La situazione appare piuttosto diverse tra i NRs. La figura 5 e la figura 7 mostrano le distribuzioni delle azioni e delle credenze normative per un certo valore di soglia normativa: la figura 5 ci mostra la distribuzione delle azioni, la figura 7 ci mostra il numero complessivo di credenze normative generato attraverso il tempo nella N-Board.

È da notare che una credenza normativa non è necessariamente la credenza più frequente nella popolazione. Tuttavia, le norme sono comportamenti che si sviluppano grazie alla diffusione delle credenze normative corrispondenti. La figura 6 mostra che nella popolazione di NRs possiamo apprezzare una forte convergenza verso una singola azione; a

differenza dei SCs (dove il tasso di convergenza è stabile nel tempo), il tasso di convergenza dei NRs aumenta attraverso il tempo (si confronti il tasso di convergenza dei SCs con quello dei NRs).

Possiamo così riassumere alcuni risultati per quanto riguarda i NRs: le credenze normative permettono

i) una chiara convergenza verso una stessa azione,

ii) una varianza maggiore fra le diverse azioni (cfr. ad es., la figura 5 e la figura 3).

Nella figura 5, dopo il tick = 60, possiamo apprezzare una crescita significativa nel numero di istanze di esecuzioni di una stessa azione, nello specifico l'azione comune (linea tratteggiata), per effetto delle credenze normative che agiscono sulla scelta dei comportamenti degli agenti.

In altre parole, dopo tick = 60, l'azione comune a tutti gli scenari si diffonde; i NRs convergono su una singola azione e una norma specifica può emergere.

La figura 7 ci aiuta a comprendere meglio questo fenomeno. Infatti, essa mostra che, a partire da circa tick = 30, una credenza normativa (quella relativa all'azione comune 1, linea tratteggiata) compare nelle N-Board degli agenti e inizia ad aumentare, nonostante la diffusione dell'azione 1 diventi evidente ed uniforme solo dopo tick = 60 (vedi Figura 5). C'è un intervallo di tempo di 30 tick tra la comparsa di una credenza normativa (tick = 30) e la convergenza degli agenti sull'azione corrispondente (l'azione comune 1) al tick = 60. E' interessante osservare che durante questo intervallo di tempo anche altre credenze normative sono generate e memorizzati nelle menti degli agenti. Affinché una credenza normativa possa incidere sul comportamento, deve trascorrere un certo numero di tick; potremmo chiamare tale intervallo di tempo la *latenza di una norma*. Il riconoscimento dell'esistenza di questo intervallo di tempo tra la comparsa di credenze normative e la convergenza sulle azioni corrispondenti, ha un impatto importante sulla teoria delle norme che qui sosteniamo. Il processo di immergenza si verifica prima che sia l'emergenza, ma ci vuole tempo per ottenere che si verifichi un effetto (sociale) di tale processo. L'emergenza di una norma è una dinamica a due vie che consiste in un processo di immergenza seguito da uno di emergenza, generando un ciclo virtuoso. La popolazione dei NRs ha un *pool* di potenziali norme sociali (vedi figura 7), corrispondenti alle credenze normative poco salienti.

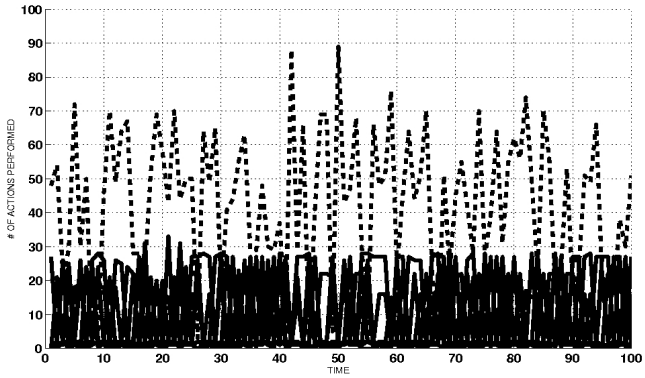


Figura 3. Azioni eseguite da SCs. Sull'asse X è indicato il numero di unità di tempo o tick della simulazione (100) e sull'asse Y il numero di azioni eseguite per ogni diverso tipo di azione. La linea tratteggiata corrisponde all'azione comune a tutti gli scenari.

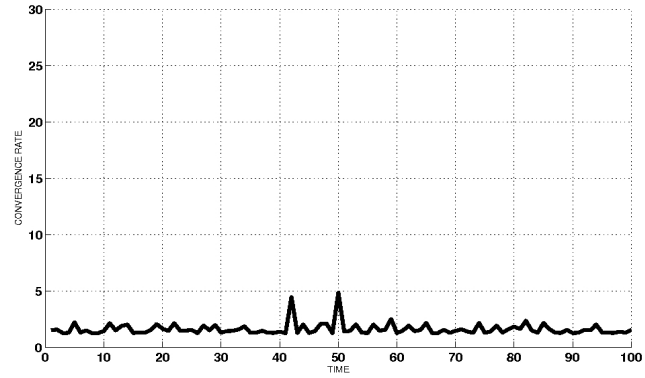


Figura 4. Sull'asse X è indicato il flusso del tempo, in asse Y il valore del tasso di convergenza dei SCs.

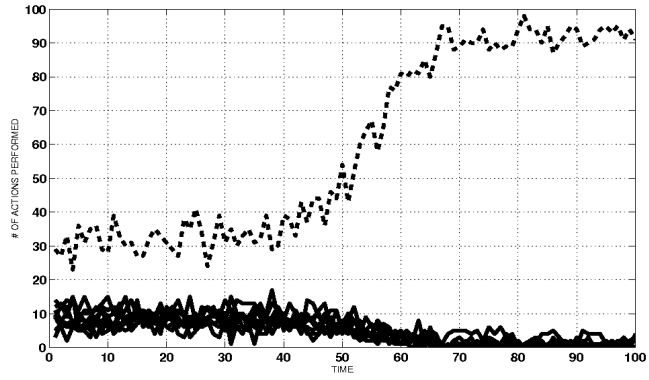


Figura 5. Azioni eseguite dai NRs. Sull'asse X è indicato il numero di unità di tempo o tick della simulazione (100) e sull'asse Y il numero di azioni eseguite per ogni diverso tipo di azione. La linea tratteggiata corrisponde all'azione comune a tutti gli scenari.

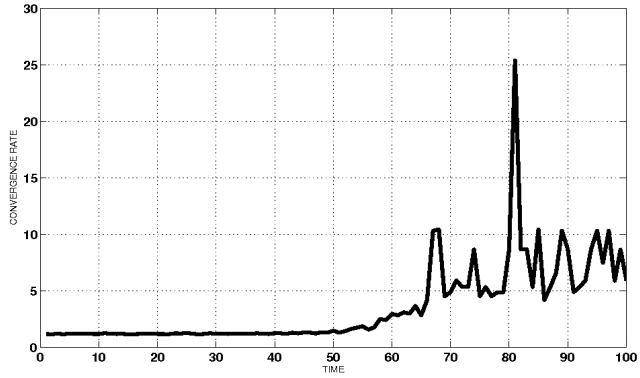


Figura 6. Sull'asse X è indicato il flusso del tempo, in asse Y il valore del tasso di convergenza dei NRs.

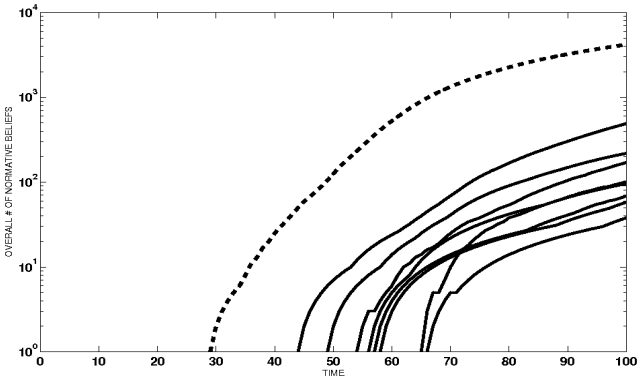


Figura 7. Ogni linea corrisponde all'andamento del numero di nuove credenze normative generate attraverso il tempo per ciascuna diversa azione. Sull'asse X il tempo; sull'asse Y il numero di nuove credenze normative. Questa figura mostra i risultati con 100 agenti e 100 tick. La linea tratteggiata corrisponde all'azione comune a tutti gli scenari.

5.5 Discussione dei risultati

I risultati presentati sembrano suggerire che il modo in cui si decide di modellare la popolazione di agenti determina la convergenza degli stessi su una norma sociale in modi abbastanza diversi: i *social conformers* (SCs) tendono, all'interno della stessa unità di tempo della simulazione (tick), a convergere in modo omogeneo, a causa del fatto che la scelta del comportamento da adottare da parte di ciascun agente sia fortemente influenzata da quella effettuata dai suoi vicini. I SCs convergono in massa su una singola azione, la *norma d'imitazione*, ma in maniera instabile e rapidamente: la conformità varia nel tempo, a seconda delle azioni eseguite da parte degli altri agenti all'interno di ogni scenario, di volta in volta.

Al contrario, i *riconoscitori di norme* (NRs) convergono su una stessa azione pur mantenendo la propria autonomia: essi scelgono come agire considerando le credenze normative che si sono formati osservando e interagendo con gli altri. Così, convergono in modo stabile; dopo un certo periodo di tempo la maggioranza degli agenti inizia a produrre la stessa azione. In questo caso, possiamo dire che una norma *sia emersa dopo essersi immersa* nelle menti degli agenti.

In sintesi, il modulo di riconoscimento normativo sembra rappresentare un requisito fondamentale per l'emergenza di una norma; esso è il processo risultante sia dalle interpretazioni degli agenti di un altrui comportamento, che della loro trasmissione l'un l'altro di tali interpretazioni.

In realtà, la capacità di generare credenze normative con differenti gradi di salienza, in base ad input esterni (osservati o comunicati), dà una maggiore stabilità al processo emerso.

Un possibile sviluppo di questo lavoro sarà introdurre miglioramenti, sia per quanto riguarda le caratteristiche sociali che quelle cognitive. In effetti, al momento, da un lato, è il solo modulo di riconoscimento normativo a guidare il comportamento degli agenti (mentre altri moduli, quali quello relativo all'adozione normativa, quello riguardante il processo decisionale e la pianificazione normativa, non sono ancora stati realizzati); d'altra parte, dinamiche sociali normative, come ad esempio la pena o il meccanismo di rinforzo, non sono stati ancora implementati.

In futuri lavori, sarà interessante

(a) progettare esperimenti con popolazioni miste SCs e NRs,

(b) considerare una popolazione di NRs i cui componenti presentino differenti soglie normative per vedere che cosa questa differenza potrebbe comportare.

Noi consideriamo la nostra piattaforma simulativa uno strumento teorico che ci permette, da una parte, di testare le potenzialità ed i limiti della presente implementazione del modulo di riconoscimento normativo, e che, dall'altra, ci indica la necessità di ulteriori sviluppi in modo da modellare dinamiche interessanti sia a livello intra-agente che a quello inter-agenti.

6. Teoria dei giochi per lo studio delle convenzioni[10]

Abbiamo visto nei precedenti capitoli la distinzione fra *norma* e *convenzione*; abbiamo presentato alcuni risultati simulativi relativi alla possibilità di postulare l'esistenza di una architettura (cognitiva) normativa per realizzare un modello computazionale di emergenza di una norma in un gruppo di agenti. Tuttavia, l'approccio proposto presenta dei limiti: esso non può beneficiare degli importanti strumenti analitici forniti dalla lunga tradizione scientifica che si occupa dello studio dei sistemi complessi in generale e dei fenomeni sociali in particolare.

In questo capitolo presentiamo un modello di emergenza di un altro tipo di artefatto sociale, la convenzione, come risultato di un processo evolutivo-imitativo tra le differenti tipologie di agenti che costituiscono una popolazione. Volendo confrontare i risultati ottenuti dal nostro modello con quelli ottenuti da altri modelli analitici, abbiamo ritenuto opportuno adottare un approccio legato alla tradizione della teoria dei giochi, ed in particolare della *evolutionary game theory*.

Di seguito, presenteremo i risultati simulativi e quelli analitici relativi alla comparsa di stati di *equilibrio* in uno scenario di traffico. L'obiettivo del lavoro è duplice: da una parte, indagare il ruolo delle convenzioni sociali in casi che presentano problemi di coordinamento, e più precisamente nei *congestion game*; dall'altra, confrontare i risultati simulativi con quelli analitici per capire cosa queste metodologie sono in grado di dirci rispetto al problema in esame.

6.1 Introduzione

Dieci anni fa, durante il primo Workshop sui Sistemi Multi-Agente e sulla Simulazione Basata su Agenti (MABS'98), H. Van Dyke Parunak, Robert Savit e Rick L. Riolo presentarono una discussione circa le somiglianze e le differenze tra i Modelli *Agent-Based*

[10] Per gli argomenti qui di seguito trattati, si veda: Cecconi, F., Campennì, M., Andrighetto, G., Conte, R. (2010) "What Do Agent-Based and Equation-Based Modeling Tell Us About Social Conventions: The Clash Between ABM and EBM in a Congestion Game Framework", *Journal of Artificial Societies and Social Simulation (JASSS)*, 13, (1), 6.

(ABM) ed i Modelli *Equation-Based* (EBM), suggerendo alcuni criteri per la scelta dell'uno o dell'altro approccio (Van Dyke Parunak et al. 1998). Questi autori hanno affermato che, nonostante la condivisione di alcuni problemi comuni, ABM e EBM si differenziano in due modi:

i) i rapporti fondamentali tra gli enti che modello

ii) il livello a cui essi operano.

Gli autori hanno osservato che queste due distinzioni sono tendenze, piuttosto che regole rigide, e indicano che i due approcci possono essere utilmente combinati.

Nel corso degli ultimi dieci anni, si è sviluppato un vivace dibattito sul tema (Epstein 2006). Una panoramica completa va oltre lo scopo di questo lavoro. In questo capitolo, si affronterà una questione teorica, riguardante la nascita di convenzioni sociali; vedremo come e in quale misura un approccio integrato di ABM e metodologie EBM ci può aiutare ad affrontare il problema.

Più in particolare, l'obiettivo di questo capitolo è quello di esplorare l'emergenza di stati stazionari (*steady-states*) nei giochi di congestione (*congestion games*, Rosenthal 1973; Milchtaich 1996; Chmura 2007), e più specificamente, di indagare l'emergenza di *una regola di precedenza* in base a micro interazioni, in situazioni di traffico intenso e congestionato (Sen e Airiau 2007).

Un gioco di congestione è una classe di giochi in teoria dei giochi, proposta per la prima volta da Rosenthal nel 1973. In un gioco di congestione si definiscono i giocatori e le risorse; il *playoff* di ogni giocatore dipende dalle risorse che esso sceglie e dal numero di giocatori che scelgono la stessa risorsa.

Facciamo un esempio: immaginiamo un viaggiatore che debba spostarsi da San Francisco a San Jose; si trova in macchina e può decidere di prendere la US Route 101 o l'interstatale 280. Mentre la 101 è più breve, la 280 è considerata più panoramica; in questo modo gli automobilisti possono avere preferenze diverse rispetto ai due flussi di traffico, possono cioè scegliere di percorrere l'una o l'altra strada. Ma ogni auto supplementare su uno dei percorsi, farà aumentare leggermente il tempo di percorrenza della rispettiva tratta.

Un celebre gioco di congestione è il gioco di minoranza (*minority game*), in cui l'unico obiettivo per tutti i giocatori è quello di essere parte del più piccolo tra due gruppi (Challet 1997; Challet 1998; Chmura 2006). Un esempio ben noto del gioco di minoranza è il problema

del Bar di El Farol, proposto da W. Brian Arthur (Arthur 1994), in cui una popolazione di agenti deve decidere se andare al bar ogni settimana, utilizzando un fattore predittivo per la stima del numero di altri agenti presenti la settimana successiva. A tutti gli agenti piace andare al bar, a meno che non sia troppo affollato (vale a dire quando è presente più del 60% degli agenti). Ma poiché non vi è alcun "predittore" unico che può funzionare per tutti allo stesso tempo (ogni agente usa il proprio fattore predittivo per scegliere quando andare al bar), non esiste una soluzione che possa essere dedotta razionalmente. Come possono raggiungere gli agenti l'obiettivo di andare al bar, evitando la folla?

È stato suggerito che, in situazioni di problemi di coordinamento (come i giochi di congestione), una *convenzione* è una *scelta razionale Pareto-ottimale* (Gilbert 1981; Lewis 1969; Sugden 2004; Young 1993), vale a dire una scelta in cui ogni cambiamento effettuato da un qualsiasi giocatore per migliorare la propria situazione è impossibile senza che ciò peggiore la situazione di un altro; così in questo tipo di circostanze, dovrebbe essere la soluzione convenzionale ad emergere.

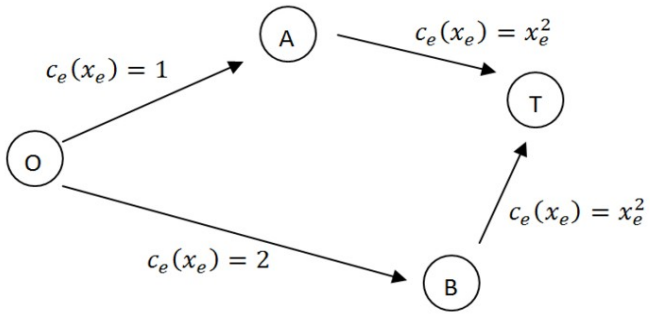
Per capire ancora meglio lo scenario studiato nel nostro modello (sia in quello simulativo che in quello analitico), cerchiamo di definire nel dettaglio un gioco di congestione in una situazione di traffico; si consideri una rete di traffico, in cui due giocatori partono da un punto O, e hanno la necessità di arrivare ad un punto T. Supponiamo che il nodo O sia collegato al nodo T tramite due punti di connessione, A e B, dove A è un po' più vicino ad O di B. Tuttavia, entrambi i punti di connessione hanno facilmente la tendenza a congestionarsi, il che significa che maggiore è il numero di giocatori che passano attraverso il punto (A o B), maggiore diventa il ritardo di ogni giocatore; inoltre se entrambi i giocatori passano attraverso lo stesso punto di connessione, ci sarà un ritardo supplementare. Un buon risultato in questo gioco sarebbe per i due giocatori quello di "coordinarsi" e passare per punti di connessione differenti (uno attraverso il punto A, l'altro attraverso il punto B). Tale risultato può essere raggiunto? E se sì, quale sarà il costo (*payoff*) della scelta per ogni giocatore?

Si consideri il seguente grafo diretto (vedi figura *Grafo* sotto) in cui ogni giocatore ha a disposizione due strategie - passare per A o passare per B - per un totale di quattro possibilità. La matrice seguente esprime i costi dei giocatori in termini di ritardi a seconda delle loro scelte:

Denoti *N* il set dei giocatori ed *E* il set delle risorse; denoti *Si* il set delle strategie del giocatore *i*, dove ciascun $s_i \in S_i$ è subset non vuoto di risorse $s_i \subset 2^E$. Ciascuna risorsa $e \in E$ ha una corrispettiva funzione di costo $c_e(x_e)$ che indica il costo della risorsa *e* quando x_e scelgono di utilizzarla. La funzione di costo per ciascun giocatore è definita da:

$$c_i(S_i) = \sum_{e \in S_i} c_e(x_e)$$

Matrice dei Payoff		
p1/p2	**A**	**B**
A	**5,5**	**2,3**
B	**3,2**	**6,6**



Grafo. Grafo diretto da O a T, passando per A o B.

In questo capitolo, proponiamo una soluzione diversa per l'argomento, sostenendo che in situazioni problematiche specifiche di coordinamento, più precisamente strutture di gioco con quattro strategie, un equilibrio convenzionale non è possibile che emerga.

Pertanto, l'obiettivo del lavoro è duplice, mostrando

a) che in situazioni di problemi di coordinamento, gli agenti razionali possono anche convergere su equilibri non convenzionali;

b) che il confronto e l'integrazione fra risultati ottenuti utilizzando un approccio simulativo e risultati analitici può darci modo di trarre alcune conclusioni interessanti su ciò che queste metodologie sono in grado di dirci rispetto al problema in esame.

Il capitolo è diviso in sette paragrafi: i paragrafi 6.2 e 6.3 descrivono una tecnica generale per scrivere un modello analitico per un gioco di popolazione (come un gioco di congestione), vale a dire la *replicator-projector dynamics*. Nel paragrafo 6.4 si descrive il nostro modello computazionale (simulativo) e un algoritmo per l'attuazione della *replicator dynamics* in esso. Nei paragrafi 6.5 e 6.6 si descrivono i risultati della simulazione e del modello analitico. Infine, nel paragrafo 6.7 si affronta il rapporto tra simulazione e modello analitico.

6.2 *Replicator Dynamics*

I giochi di congestione coinvolgono un gran numero di agenti semplici, ciascuno dei quali può giocare utilizzando una delle regole di cui dispone. In generale, un gioco di congestione si occupa di una società *P* composta da *p* diverse popolazione di agenti; tuttavia, il nostro modello utilizza solo una singola popolazione (caso p = 1) di agenti. Possiamo chiamare questi giochi, giochi di popolazione (population games). Durante il gioco, ogni agente sceglie una strategia della serie S = {1 ,..., n} dove n è il numero totale di strategie (comportamenti che possono essere adottati).

Possiamo descrivere la *replicator dynamics* per un gioco di popolazione.

Sia Σ lo spazio che contiene tutti i vettori delle distribuzioni di frequenza degli agenti che giocano:

$\Sigma = \{\vec{x} \in \Re^n : x_i \in [0,1] \forall i \in S, \sum_{i=1}^{n} x_i = 1\}$	1)

ad esempio potremmo avere una popolazione con due diverse strategie, *n = 2*; la strategia (comportamento) *1* è quella di *mantenere la destra*, la strategia *2* è quella di *mantenere la sinistra*.

x1 è la frequenza di agenti che stanno mantenendo la destra, *x2* è la frequenza degli

agenti che stanno mantenendo la sinistra. Una volta che fissiamo *P* e *S*, i set delle popolazioni e delle strategie, possiamo identificare un gioco con la propria funzione di *payoff*, $P: \Sigma \rightarrow \boldsymbol{R}^n$, vale a dire una mappa che assegna un vettore di *payoff* per ogni distribuzione di frequenza, uno per ogni strategia in ciascuna popolazione. Si definisce $\boldsymbol{\pi}^i$ il *payoff* degli agenti con la strategia (comportamento) i.

Possiamo calcolare tale mappa dei *payoff* in due diversi modi.

a) *Gioco di coordinamento*. La funzione di *payoff* è una matrice costante n x n che contiene i risultati dei *payoff* derivanti dalle interazioni tra due strategie: questo tipo di gioco si chiama anche RPS Game (stando RPS per *Rock-Paper-Scissor*, per celebrare il famoso gioco a 3 strategie in cui il sasso batte la forbice, la forbice taglia la carta e la carta avvolge il sasso).
b) *Gioco di Congestione.* La funzione di *payoff* è funzione sia delle distribuzioni di frequenza che di una sorta di combinatoria tra gli agenti. In un gioco gli agenti interagiscono in maniera complessa (in questo caso sono ammesse interazioni con più di due agenti). Ad esempio, consideriamo un insieme di città collegate da una rete di collegamenti (ad esempio, autostrade). Gli agenti hanno bisogno di spostarsi da una città all'altra, scegliendo un percorso. Il *payoff* di un agente che sceglie un percorso è dato dal contrario del ritardo che accumula prendendo quella strada (dato *r* il ritardo, il *payoff = 1 - r*). Il ritardo su un percorso corrisponde alla somma dei ritardi sui suoi collegamenti (*link*), e il ritardo su un collegamento è una funzione del numero di agenti che utilizzano tale collegamento.

Una *replicator dynamics* in un gioco di popolazione è un processo di cambiamento nel tempo della distribuzione delle frequenze delle strategie: le strategie con *payoff* maggiore si riproducono più velocemente. In termini matematici, la *dinamica è rappresentata dalla traiettoria nello spazio delle frequenze*.

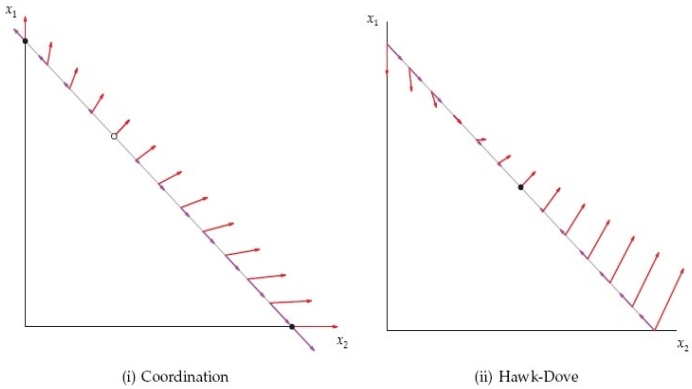


Figura 1. (Sandholm 2008) Mostriamo la dinamica di un gioco popolazione. Le frecce rosse indicano la direzione che seguono le frequenze delle strategie. Le frecce viola indicano la proiezione delle frecce rosse su un sottospazio delle strategie: in questo caso, il sottospazio è una linea, corrispondente ai punti in cui la somma delle frequenze è uguale ad 1. Sulla destra il caso del noto Hawk-Dove Game. In questo secondo caso si mostra che nel Hawk-Dove Game vi è un punto di stabilità (ma non in un semplice gioco di coordinamento, figura 1-i a destra).

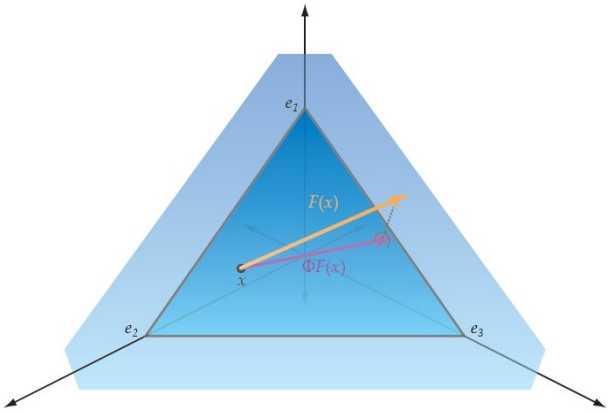


Figura 2. (Sandholm 2008) ci mostra la projector dynamics di un gioco di popolazione con 3 strategie.

Usiamo due figure (figura 1 e figura 2) da (Sandholm 2008) per mostrare che la *replicator dynamics* può essere descritta in termini geometrici, come una traiettoria con alcuni vincoli. Questo punto è alla base della modellazione che utilizza la *projector dynamics*.

Definiamo π il *payoff* medio dell'intera popolazione,

$\overline{\pi} = \sum_{i=1}^{4} \pi^i \frac{x_i}{N}$	2)

La *replicator dynamics* è data da

$\dot{x}_i = x_i[\pi^i - \overline{\pi}]$	3)

vale a dire che in base alla *replicator dynamics*, la frequenza di un comportamento aumenta quando il suo *payoff* è superiore alla media. Per ogni passo *t* di tempo, il sistema di equazioni differenziali (3), definisce un campo vettoriale (vedi figura 1). Chiamiamo campo vettoriale un campo che raggiunge la r*eplicator dynamics F*. Implementiamo *F* attraverso un algoritmo di imitazione globale. Durante il gioco, ogni agente verifica se un altro individuo della popolazione ha un *payoff* superiore al suo. La regola principale di questo modello è che un agente con un *payoff* inferiore imita il comportamento di un altro agente che presenta un *payoff* maggiore. Nel nostro modello, non abbiamo alcuna mutazione nel processo di imitazione. Questo quadro teorico deriva direttamente dalla formalizzazione eseguita da (Ekman 2001).

6.3 Geometria dei giochi di popolazione (population games)

6.3.1 *Projector Dynamics*.

La dinamica descritta da (3) è soggetta ad un vincolo, in quanto il numero di agenti rimane costante, cioè la somma delle x_i deve essere *1*. Si potrebbe, quindi, descrivere la dinamica del gioco, con un approccio diverso, la cosiddetta *Projector Dynamics* (Sandholm 2008). Il punto chiave di questo approccio è che il campo vettoriale *F* che descrive il gioco di popolazione è proiettato su un *nuovo* campo vettoriale

$\phi(F) = F_\Sigma : \Sigma \rightarrow \Sigma_0$	4)

dove

$\Sigma_0 = \{\vec{x} \in \Re^n : x_i \in [0,1]\ \forall i \in S\ ,\ \sum_{i=1}^{n} x_i = 0\}$	5)

Informalmente, noi proiettiamo il vettore da *F* su una linea (se n = 2), su un *simplesso bidimensionale* (se n = 3, vedi figura 2) e così via. La proiezione di un campo vettoriale è una procedura standard. È possibile trovare un algoritmo in (Sandholm 2008).

6.3.2 *Projector Dynamics* con due strategie in un gioco RPS.

Siamo ora in grado di spiegare l'approccio *Projector Dynamics* con due strategie utilizzando un gioco RPS con due strategie, in particolare il gioco Hawk-Dove (Sandholm 2008). Nel gioco Hawk-Dove possiamo riempire la matrice dei *payoff* P_{HD} (matrice dei payoff di un gioco RPS) utilizzando le seguenti regole:

1) quando un Hawk incontra un Dove, Hawk vincere *a*;
2) quando un Hawk incontra un altro Hawk, entrambi perdono *b* con *b* <*a*;
3) quando un Dove in contra un altro Dove, entrambi vincono *b*;
4) quando un Dove incontra un Hawk, Dove ottiene *0*.

Possiamo calcolare le componenti della proiezione di P_{HD} per il gioco Hawk-Dove. Come abbiamo detto prima, P_{HD} denota la matrice dei *payoff* del gioco Hawk-Dove

$P_{HD} = \begin{pmatrix} -b & a \\ 0 & b \end{pmatrix}$	6)

Se *x* è il vettore delle frequenze, abbiamo

$P_{HD}\begin{pmatrix} x_H \\ x_D \end{pmatrix} = \begin{pmatrix} -bx_H + ax_D \\ bx_D \end{pmatrix}$	7)

quindi da (7), possiamo calcolare le componenti della proiezione di P_{HD} per il gioco Hawk-Dove

$\phi(P_{HD})\begin{pmatrix} x_H \\ x_D \end{pmatrix} = \begin{pmatrix} \frac{1}{2}(-bx_H + ax_D) - \frac{b}{2}x_D \\ -\frac{1}{2}(-bx_H + ax_D) + \frac{b}{2}x_D \end{pmatrix} = \begin{pmatrix} \frac{1}{2}(a-b)x_D - \frac{b}{2}x_H \\ \frac{1}{2}(b-a)x_D + \frac{b}{2}x_H \end{pmatrix}$	8)

La relazione ottenuta per la proiezione ci permette di affermare che il gioco ha certamente un punto di equilibrio (si veda la figura 1, a destra).

6.3.3 *Projector Dynamics* in un gioco di congestione.

Nel paragrafo precedente, abbiamo descritto la *projector dynamics* in un gioco RPS, con una matrice costante che rappresenta i *payoff* risultanti dalle interazioni fra 2 agenti. Nel nostro modello (un gioco di congestione), adattiamo l'approccio legato alla *projector dynamics* per calcolare una matrice di *payoff* a partire da una funzione di *payoff* con due parametri, cioè la distribuzione delle strategie al tempo *t* e la combinazione di strategie allo stesso tempo.

Per esempio, abbiamo creato un gioco in cui

1) 2 Hawk *vis-à-vis* ad 1 Dove ottengono una ricompensa diversa rispetto a quello che accade nella situazione opposta (2 Dove e 1 Hawk);
2) agenti che giocando le stesse strategie interagiscono (come nel gioco RPS, Hawk-Hawk e Dove-Dove).

Possiamo riempie la matrice dei *payoff* P_{CG} (matrice dei *payoff* per il gioco di congestione), utilizzando le seguenti regole:

1a) riempiamo la diagonale utilizzando le interazioni della stessa strategia, cioè l'interazione fra le stesse strategie, per esempio Hawk-Hawk e Dove-Dove;

2a) riempiamo le altre celle della matrice con le interazioni di diverse strategie, cioè l'interazione tra le strategie corrispondenti in una matrice costante di *payoff* (nel caso di semplice Hawk-Dove, si considera l'interazione uno-ad-uno Hawk-Dove).

La caratteristica principale del gioco di congestione è: quando si compila la matrice dei *payoff*, si calcola la probabilità di interazioni, ad esempio, la probabilità di avere 2 Hawks v*s* 1 Dove.

6.4 Il modello

6.4.1 Il problema: un incrocio.

Sia *P* una società costituita da una singola popolazione con *N* individui. Nel nostro modello, si prende in considerazione un gioco in cui ogni individuo segue uno fra quattro comportamenti (o strategie) possibili, WatchRight, WatchLeft, Hawk e Dove. Indichiamo il comportamento con l'indice I . All'interno della società, la frequenza degli individui con il comportamento i è x_i .

Definiamo i comportamenti e calcoliamo il *payoff* utilizzando 3 diversi algoritmi: l'algoritmo che modella un comportamento *condizionato*, *compiacente* o *aggressivo* (che per semplicità d'ora in poi chiameremo rispettivamente algoritmo condizionato, algoritmo compiacente ed algoritmo aggressivo). L'algoritmo condizionato descrive i comportamenti di WatchRight e WatchLeft; l'algoritmo compiacente descrive il comportamento di Dove; l'algoritmo aggressivo descrive il comportamento di Hawk.

Usiamo un modello a tempi discreti (seguendo lo *scheduling* descritto in *algoritmo 1*), in cui gli individui si muovono in maniera casuale (compiendo un *random walk*) in un reticolo discreto *L* bi-dimensionale. Usiamo un reticolo regolare, con *C* celle. Consideriamo ogni cella di L un incrocio con quattro direzioni di transito possibili (vedi figura 3). Descriviamo l'incrocio in senso antiorario e partendo da Ovest; se lo guardiamo dall'alto. Se stiamo usando il punto di vista soggettivo degli agenti (cioè l'incrocio come appare al singolo individuo), usiamo per convenzione le possibili direzioni destra, sinistra e avanti (prima si indica quello che trovo sulla mia destra, poi alla mia sinistra, infine, di fronte a me).

Quindi, in Figura 3.a mostriamo (1) un individuo con direzione di spostamento verso est (con un comportamento WatchLeft); (2) la direzione di movimento Nord-Sud è vuota (cioè

nessun individuo si muove in tale direzione); (3) un individuo con un comportamento Hawk è in movimento verso ovest; (4) un individuo con comportamento Dove si muove verso Nord. Chiamiamo *direzioni parallele* le direzioni Nord-Sud e Sud-Nord, e le direzioni Ovest-Est ed Est-Ovest. Definiamo *ortogonali* le altre restanti combinazioni possibili.

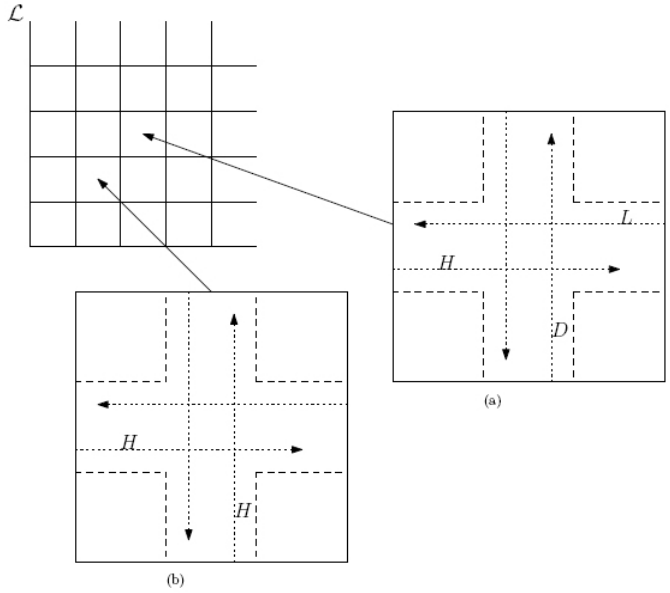


Figura 3. Due istantanee di L. In (a) ci sono 3 agenti, in (b) solo 2; mettiamo gli agenti nei punti d'ingresso all'incrocio, quindi essi cercano di attraversarlo.

Il modello presenta alcuni vincoli:

- gli agenti che sono da soli in una cella non ottengono *payoff*;
- un massimo di quattro agenti può occupare una singola cella;
- è ammesso un unico agente per direzione (non è possibile che, per esempio, due agenti con strategia Hawk abbiamo la stessa direzione di movimento Nord-Sud);
- valgono solo agenti ortogonali (vengono cioè considerati per il computo del *payoff* solo gli agenti provenienti da direzioni ortogonali).

Descriviamo di seguito in pseudo codice l'algoritmo condizionato (*algoritmo 2*), l'algoritmo aggressivo (*algoritmo 3*) e l'algoritmo compiacente (*algoritmo 4*).

Algorithm 1 The scheduling of simulation. We call Crossroads the number of cells with more than 1 individual.

```
for t = 1 to T do
  for a = 1 to N do
    a ⇐ P.
    to set a over L
  end for
  for c = 1 to Crossroads do
    Compute payoffs
  end for
  for a = 1 to N do
    Imitation dynamic
  end for
end for
```

Algorithm 2 Conditionated. *A* is the individual. *X* indicates the direction to monitor. *CRASH* indicates that **A** goes into the crossroad at the same time of an other *orthogonal* individual. *STOP* indicates that **A** does not go into the crossroad. *GOAHEAD* indicate that the individual crosses the crossroad. *H* is an *Hawk* individual. *D* is a *Dove* individual. We define *not blocked* a conditionated individual with the *free* monitored direction.

```
if X is occupied then
  STOP
else
  if there is (orthogonal H) or not blocked C then
    CRASH
  else
    GOAHEAD
  end if
end if
```

Algorithm 3 Aggressive.

```
GOAHEAD
if there is (orthogonal H) or not blocked C then
  CRASH
end if
```

Algorithm 4 Compliant.

```
if there is orthogonal individuals then
  STOP
else
  GOAHEAD
end if
```

6.4.2 Calcolo dei *payoff*

Per realizzare un modello analitico per il nostro gioco di congestione, abbiamo dovuto scrivere un'espressione per $\boldsymbol{\pi}^{i}$, il *payoff* per la strategia *i*.

Più precisamente, dobbiamo scrivere una matrice P_{CG}, 4 x 4. Gli elementi di questa matrice sono funzioni della frequenza delle strategie e della probabilità delle combinazioni delle strategie nel tempo. Il calcolo del $\boldsymbol{\pi}^{i}$ conduce direttamente a riempire la matrice P_{CG}, utilizzando le regole come descritte nel paragrafo 7.3.3, vale a dire ... *riempiamo la diagonale utilizzando la stessa strategia di interazione (ad esempio WatchRight Vs WatchRight), ... riempiamo le altre celle con le interazioni delle diverse strategie (per esempio: Dove Hawk VS WatchRight).*

Noi chiamiamo i comportamenti *i*, *r = 1*, *l = 2*, *d = 3*, *h = 4*.

Le regole generali sono: se un agente *STOP*, il *payoff* che riceve è *0* . Se due agenti CRASH, i loro *payoff* sono fissati a -1. Se un agente GOESAHEAD (quindi riesce ad attraversare l'incrocio senza scontrarsi con un altro agente, il suo *payoff* è 1. Denoti $x_i(t)$ il numero di individui con comportamento *i* nel periodo di tempo *t*. Chiaramente, avremo che *x1 + x2 + x3 + x4 = N*, dove *N* è il numero totale di giocatori. Indichi *C* il numero totale delle celle di cui è composto il mondo. Nel nostro caso, noi studiamo il modello con densità *C = N*.

La probabilità che una cella contenga *k* individui è data da:

$\Pr\left(k; N, p = \frac{1}{N}\right) = \binom{N}{k}\left(\frac{1}{N}\right)^{k}\left(1 - \frac{1}{N}\right)^{(N-k)}$	9)

Chiamiamo a_k il numero di celle con *k* individui. Possiamo calcolare il *payoff* per ogni comportamento come la somma dei prodotti tra a_k e la somma dei *payoff* per ogni combinazione, con *k* fisso.

Ad esempio per *k = 2* avremo 10 combinazioni, *h = {rr, ll, dd, hh, rl, rd, rh, ld, lh, dh}*. Solo 8 di queste combinazioni possibili danno un payoff diverso da zero, *rr, ll, dd, rd, rh, ld, lh, dh.*

hh non da *payoff*, perchè abbiamo *hh parallelo (=)* con *payoff* uguale a 2 e *hh ortogonale (||)* con *payoff* uguale a -2. *rl* non da *payoff* poiché abbiamo 2 per la combinazione parallelo e per le 2 combinazioni ortogonali (con probabilità ½ ed ½) otteniamo 0 e -2. Chiamiamo K_h^i la matrice dei *payoff* differenti da zero per il comportamento *i* e la

combinazione *h*. La probabilità della combinazione h con *h1 + h2 + h3 + h4 = k* condizionato alle frequenze, è una distribuzione multinomiale

$$\Pr(\boldsymbol{h}|x_1,x_2,x_3,x_4) = \frac{\binom{x_1}{h_1}\binom{x_2}{h_2}\binom{x_3}{h_3}\binom{x_4}{h_4}}{\binom{N}{k}}$$	10)

Finalmente, possiamo definire il *payoff* per la strategia (comportamento) *i*

$$\pi^i(\boldsymbol{x}) = \sum_{k=2}^{4} a_k \left[\sum_{\boldsymbol{h}|k} \Pr(\boldsymbol{h}|\boldsymbol{x}) K^i_{\boldsymbol{h}}\right]$$	11)

Questa (figura 4) è una istantanea dell'interfaccia del simulatore che abbiamo implementato utilizzando NetLogo, uno strumento *open source* scaricabile gratuitamente da internet (dal sito: `http://ccl.sesp.northwestern.edu/netlogo/`).

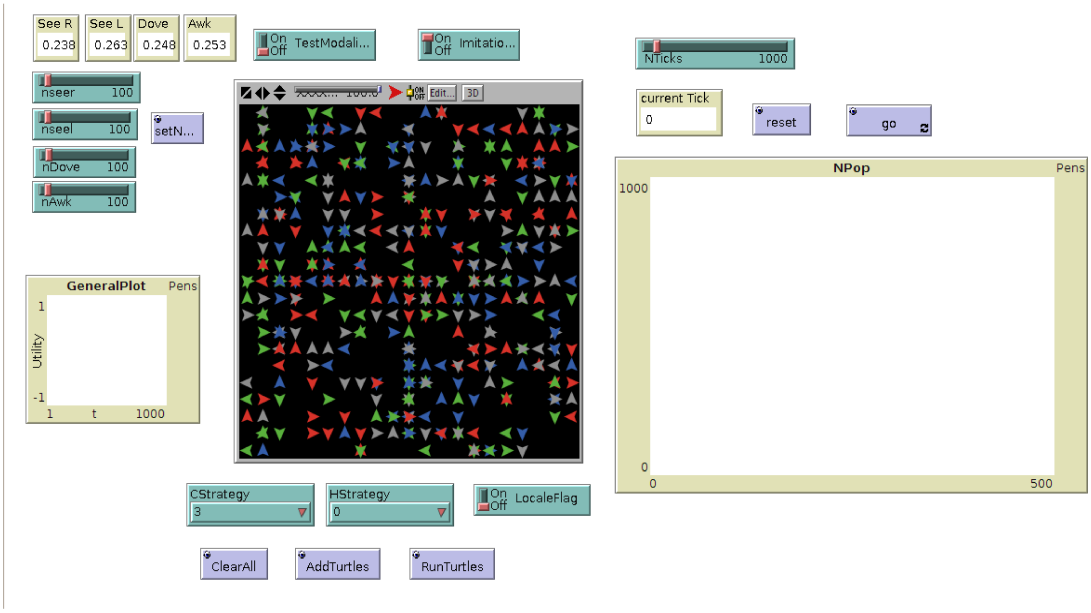


Figura 4. Una istantanea dell'interfaccia del simulatore.

In questa interfaccia (si veda il sito: http://laral.istc.cnr.it/cecconi/AwkDoveModel.html) ciascuna differente strategia (o comportamento) è rappresentata da un differente colore:

- gli agenti blu sono *WatchRight*
- gli agenti grigi sono *WatchLeft*
- gli agenti rossi sono *Hawk*
- gli agenti verdi sono *Dove*

L'oggetto al centro (sfondo nero e piccole frecce colorate) rappresenta il mondo toroidale (in cui cioè non esiste soluzione di continuità fra sopra-sotto e destra-sinistra, il che vuol dire arrivati al margine superiore del mondo se si continua a salire si spunta dal margine inferiore e lo stesso vale per i margini destro e sinistro) in cui gli agenti si muovono casualmente effettuando un classico *random walk*.

Gli *slider* sulla sinistra sono utilizzati per stabilire, all'inizio della simulazione, il numero degli agenti per (cioè che adotta) ciascuna strategia. I due oggetti in cui vengono effettuati i *plot* (sulla destra e sulla sinistra del mondo, rispettivamente) sono utilizzati per monitorare il *payoff* medio ottenuto da ciascuna strategia ed il numero di agenti *sopravvissuti* per ciascuna strategia.

6.5. Risultati delle simulazioni

Abbiamo fatto girare 256 simulazioni facendo variare la dimensione di ciascuna sotto-popolazione di agenti fra 50 e 200 individui, ottenendo così che il *range* in cui varia la popolazione complessiva sia compreso tra 200 ed 800 individui. Ciascuna simulazione prevede 1000 *tick* (unità di tempo o step della simulazione).

Questa è una tabella che riassume i parametri utilizzati

Tabella dei Parametri	
Parameter	***Value***
WatchRight	from 50 to 200
WatchLeft	from 50 to 200
Doves	from 50 to 200
Hawks	from 50 to 200
# of Ticks	1000
Type of imitation	global

Nelle figure 5, 6 e 7 mostriamo i risultati di alcuni singoli *run* di simulazione come esempi delle tre differenti tipologie di risultati che abbiamo ottenuto.

In conformità con la descrizione fatta sopra di tutti i possibili *steady states* raggiungibili, abbiamo ottenuto i seguenti risultati:

1. sopravvive solo una sotto-popolazione

- WatchRight: 3 volte (1.18 %);
- WatchLeft: 5 volte (1.95);

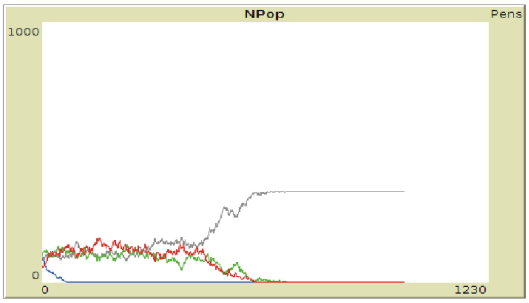


Figura 5. In questo caso la simulazione inizia con 100 WatchRight, 100 WatchLeft, 100 Dove e 50 Hawk e alla fine sopravvivono solo i WatchLeft (la popolazione al termine della simulazione sarà cioè composta da 350 WatchLeft)

2. sopravvivono due sotto-popolazioni:

- (a) Hawk e Dove: 48 volte (18.75%)
- (b) WatchRight e Dove: 15 volte (5.86%)
- (c) WatchLeft e Dove: 19 volte (7.42%)

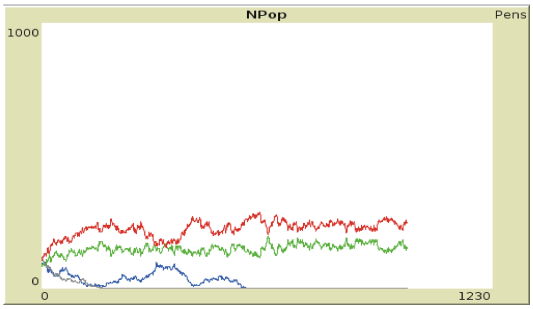


Figura 6. In questo caso la simulazione inizia con 100 WatchRight, 100 WatchLeft, 100 Dove e 100 Hawk e sopravvivono solo Hawk e Dove.

3. sopravvivono tre popolazioni:

- (a) WatchRight, Hawk e Dove: 85 volte (33.2%)
- (b) WatchLeft, Hawk e Dove: 81 volte (31.64%)

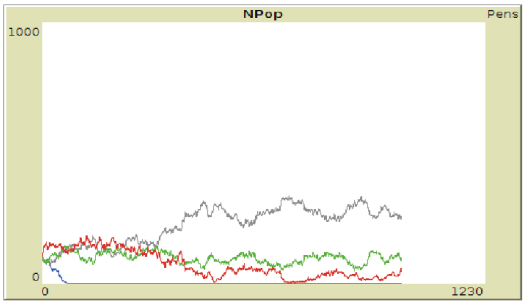


Figura 7. In questo caso la simulazione inizia con 100 WatchRight, 100 WatchLeft, 100 Dove e 100 Hawk

e sopravvivono WatchLeft, Hawk e Dove.

Le simulazioni possono terminare con *steady states* differenti, che possiamo riassumere così (vedi anche figura 8):

i ***sopravvive solo una sotto-popolazione***: in questa situazione, la sotto-popolazione può essere solo di WatchRight o WatchLeft, suggerendo questo che, utilizzando i termini di Lewis, queste sono soluzioni alternative auto-sufficienti a problemi di coordinamento. Essendo le due soluzioni equivalenti, la scelta tra di esse è puramente arbitraria. In tal modo, una volta che una soluzione è stata selezionata, l'altra non può coesistere (*nei paesi in cui si tiene la destra nella guida, non si può tenere la sinistra e viceversa*);

ii ***due sotto-popolazioni sopravvivono***, in questo caso:

- come accade con la strategia evolutivamente stabile (*Evolutionarily Stable Strategy* - ESS) (Smith, 1974; Gilbert 1981), se una sotto-popolazione è Hawk, l'altra sarà necessariamente Dove; Hawk non sopravvive con sotto-popolazioni di WatchRight o WatchLeft;
- se, al contrario, una delle sotto-popolazioni è Dove, l'altra può essere una qualsiasi delle tre rimanenti (Hawk, WatchRight o WatchLeft);
- infine, le strategie non-condizionate (Hawk e Dove) non sono simmetriche: la prima (Hawk) può sopravvivere solo sfruttando la strategia più altruista (Dove).

iii ***tre sotto-popolazioni sopravvivono***, includendo fra le tre o WatchRight o WatchLeft, ma mai entrambe allo stesso tempo;

iv ***quattro sotto-popolazioni*** non possono mai sopravvivere (ciò deriva da *iii*).

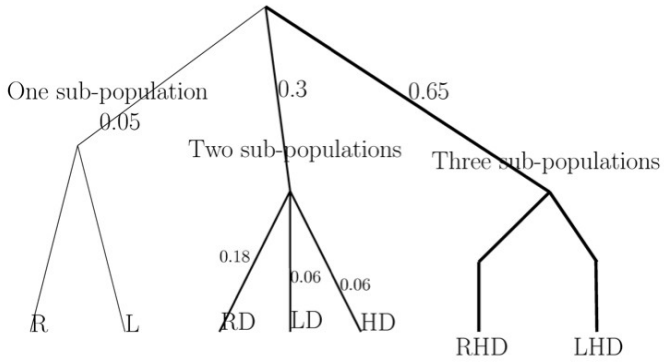


Figura 8. L'albero in figura mostra la distribuzione dei possibili steady states. Lo spessore delle frecce indica la frequenza con cui occorre il particolare steady state (più è spessa, maggiore è la frequenza).

Sommando tutte le percentuali ottenute, possiamo calcolare tutti i casi in cui sopravvive ciascuna sotto-popolazione (da sola o insieme ad altre). In questo modo possiamo suggerire una sorta di *gerarchia di adattabilità*:

1. Dove: 96.87%
2. Hawk: 83.59%
3. WatchRight: 40.24%
4. WatchLeft: 41.01%

Questo risultato ci mostra che nel gioco del nostro modello (così come è stato pensato da noi) la strategia che *paga di più* è quella estremamente prudente di Dove.

6.6 Risultati Analitici.

Vediamo ora di contrastare i risultati delle precedenti simulazioni con quelli che possono essere ottenuti analiticamente.

Siamo in grado di estrarre informazioni dal modello analitico in tre modi diversi:

1. risolvendo numericamente il sistema di equazioni differenziali (vedi figura 9);
2. disegnando il campo vettoriale descritto dal sistema ODE (vedi figura 10);
3. infine, possiamo calcolare gli *steady state* e le pendenze intorno a loro, e stabilire se essi sono stabili o meno (vedi figura 11).

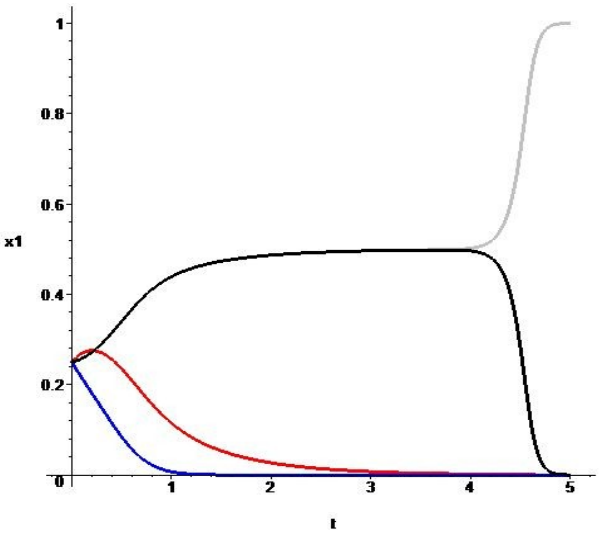


Figura 9. Una soluzione numerica del modello analitico. La curva rossa mostra la frequenza di Hawk, la blu indica la frequenza di Dove; la grigia e la nera mostrano le frequenze delle strategie convenzionali (WatchRight e WatchLeft). La numerosità iniziale per ciascuna sotto-popolazione è uniforme (25% ciascuna). Questa figura ci mostra che, dopo un certo periodo di tempo, una delle due strategie convenzionali vince (in questo caso lo steady state è ad una sotto-popolazione). La dinamica mostra che le strategie convenzionali sono incompatibili.

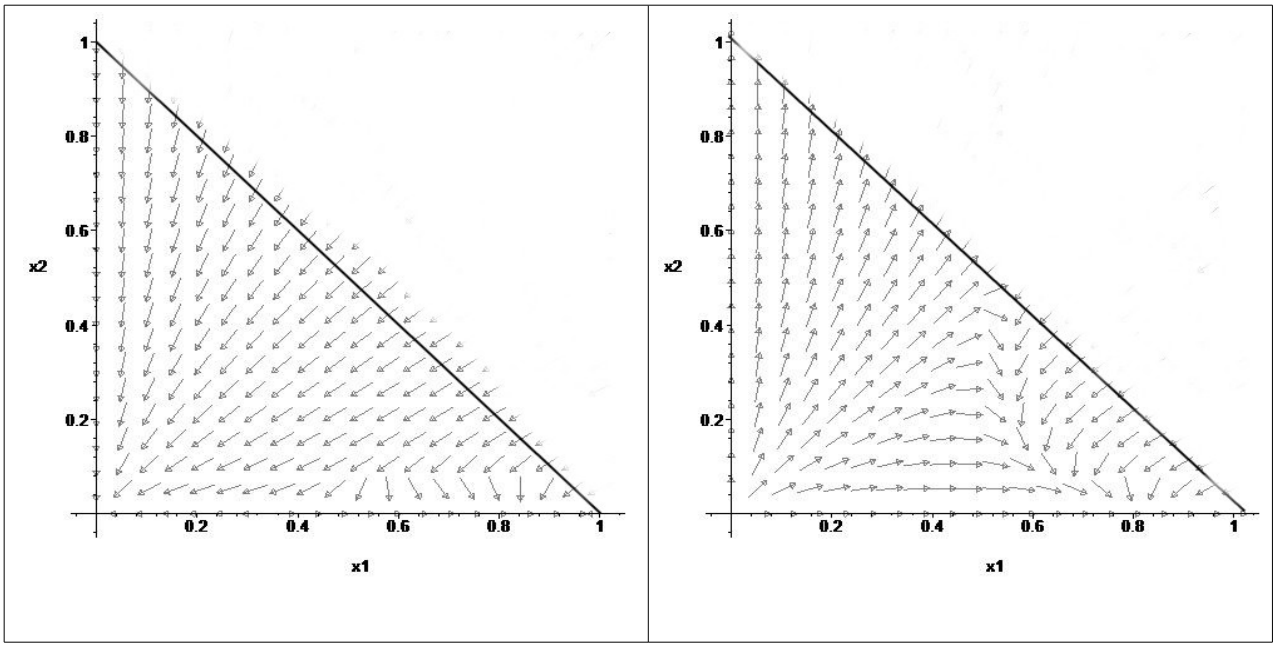


Figura 10. In figura si mostra il campo vettoriale per x1 = WatchRight e x2 = WatchLeft. Nella sotto-figura di sinistra abbiamo fissato il numero di Hawk a zero; in quella di destra il numero di Dove. Le figure mostrano che nel caso in cui il numero di Hawk è zero (sinistra) lo steady state raggiunto è molto meno stabile che nel caso in cui il numero di Dove sia zero (destra).

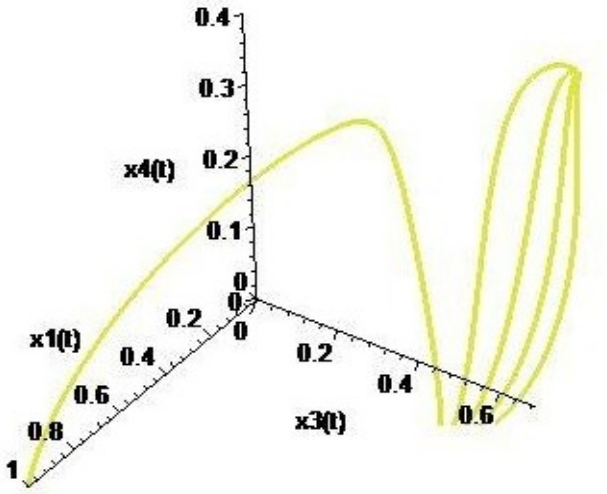


Figura 11. In questa figura abbiamo una mappa delle traiettorie della dinamica a partire da differenti numerosità iniziali delle sotto-popolazioni fino agli steady states finali. In figura si mostra che, partendo con un numero medio di Dove ed un N (popolazione complessiva) medio, troviamo quattro possibili traiettorie su cinque che conducono a steady states finali a due sotto-popolazioni (x1, x3, x4 significano: WatchRight, Dove and Hawk; in questo abbiamo fissato a zero il numero dei WatchLeft).

Utilizzando questi metodi, abbiamo ottenuto alcuni risultati analitici, che *fittano* con i risultati simulativi.

1. Le strategie condizionate WatchRight e WatchLeft sono incompatibili. La relazione ottenuta dalla *projector dynamics* ci permette di affermare che i giochi con una sola strategia hanno un punto stazionario instabile (la derivata delle frequenze ha segno opposto, si veda la figura 9).
2. Le strategie condizionate (WatchRight e WatchLeft), da un lato, e la strategia aggressiva (Hawk), dall'altro, sono incompatibili. Viceversa, troviamo uno *steady state* stabile con le strategie condizionate combinate con la strategia Conforme (Dove) (vedi figura 10).

I modelli analitici suggeriscono alcuni ulteriori spunti:

– La figura 9 mostra che deve trascorrere un certo periodo di tempo prima della separazione tra le strategie condizionate (lo stesso ritardo è osservato anche durante le simulazioni). Così, si potrebbe sostenere che il ritardo non è un effetto stocastico (proprio quindi solo del modello simulativo) proveniente dalle

fluttuazioni, durante la simulazione *agent-based*. Il ritardo è effetto della non-linearità delle interazioni.

- Dalla EBM, abbiamo una mappa generale (vedi figura 11) delle traiettorie a partire dagli stati iniziali, con diverse frequenze, fino ad arrivare agli *steady states* finali stabili. La mappa mostra una certa corrispondenza con la distribuzione degli *steady states* ottenuti dalle simulazioni (0,05 un sotto-popolazione, 0,3 due sotto-popolazioni, 0,65 tre sotto-popolazioni).

6.7 Discussione dei risultati

Cerchiamo di riassumere quello che potrebbe essere il valore aggiunto di questo capitolo

a) dal punto di vista della comprensione del ruolo delle convenzioni in giochi di congestione;
b) nel confronto tra l'approccio ABM e l'approccio EBM, nello studio dei fenomeni sociali.

Per quanto riguarda il punto a), nel nostro modello ABM abbiamo indagato il ruolo delle strategie condizionate - che possono essere viste come convenzioni sociali nel senso di Lewis - nella soluzione di un gioco di congestione, in confronto e in combinazione con le strategie incondizionate.

Che cosa questi risultati ci dicono rispetto alle convenzioni sociali? In linea con la teoria dei giochi evolutiva (*evolutionary game teory*), questo studio aiuta a prevedere quali strategie possono raggiungere uno *steady state*. Inoltre, permette di rintracciare nuove proprietà per le strategie. In particolare, siamo in grado di distinguere non solo tra strategie incondizionate e strategie condizionate, ma anche tra strategie equivalenti incompatibili (ossia le strategie che hanno lo stesso *payoff* per tutti i giocatori) e strategie non-equivalenti complementari (ossia le strategie in cui gli agenti hanno *payoff* differenti; si veda la sezione sui risultati delle simulazioni). Le prime sono auto-sufficienti, le ultime non lo sono.

Quindi, anche se non possiamo prevedere quale sarà l'equilibrio specifico da raggiungere, possiamo calcolare la composizione finale della popolazione, dato l'equilibrio che è garantito da una certa strategia.

Ma è vero che *non possiamo prevedere quale sarà l'equilibrio specifico da raggiungere*? Non completamente. In realtà, utilizzando EBM siamo in grado di *prevedere* se

uno *steady state* implichi una combinazione di strategie, con quale probabilità uno *steady state* segue data una certa distribuzione iniziale, e la stabilità degli *steady states* raggiungibili (si vedano i risultati nella sezione dedicata ai risultati analitici).

Il punto b) (il confronto tra ABM e EBM) è più controverso.

La vera domanda è: sarebbe possibile utilizzare l'ABM senza l'aiuto dell'EBM?

E, d'altra parte, sarebbe possibile realizzare il suddetto modello analitico, senza i dati delle simulazioni? Probabilmente no. Noi sosteniamo che, per quanto riguarda il fenomeno sociale che ci interessa, ossia la soluzione di un problema di coordinamento in una popolazione estremamente numerosa di agenti eterogenei in interazione, i dati della simulazione sono decisivi.

I modelli del primo tipo, cioè quelli ad agenti (ABM), includono già caratteristiche che sono necessarie per la realizzazione del modello matematico. Tuttavia, possiamo affermare che queste due classi di modelli non sono alternative, ma in alcune circostanze complementari. Cerchiamo di difendere questo argomento.

La differenza principale tra ABM ed EBM è la capacità di cogliere i diversi aspetti stocastici dei fenomeni:

- ABM descrive le fluttuazioni stocastiche;
- EBM descrive le statistiche delle fluttuazioni, per esempio la media, e la forma della distribuzione.

I risultati generati dalle simulazioni possono *fittare* con il modello analitico: forniscono dati *in-silico*, per lo sviluppo del modello analitico (utilizzabile per generalizzare i risultati della simulazione) e consentono di fare previsioni (Conte, 2002; Bonabeau 2002).

Sosteniamo che né le simulazioni, né il modello analitico da solo ci possono aiutare a spiegare *perché* sono stati ottenuti certi risultati: in altre parole, nessuna delle due metodologie ci aiuta a comprendere i dati che abbiamo generato.

Tornando al nostro modello, i risultati mostrano che giocando con quattro popolazioni qualitativamente diverse, non solo emergono convenzioni - cioè soluzioni arbitrarie equivalenti a problemi di coordinamento (nel nostro modello, WatchRight e WatchLeft si possono affermare da sole), - ma anche soluzioni di reciproco sfruttamento - rappresentate nel nostro caso dalla coesistenza di Hawk e Dove. In altre parole, eliminando la logica del gioco a due strategie, egemone nella teoria dei giochi, non è affatto garantito che emergerà un equilibrio convenzionale. Possiamo provare ad azzardare qualche spiegazione, utilizzando la

matrice analitica dei *payoff*. Ad esempio, il calcolo del *payoff* mostra che vi è un progressivo deterioramento delle strategie condizionate procedendo da un mondo non affollato verso un mondo affollato. In altre parole, se gli agenti vivono in un ambiente in cui la probabilità di un incrocio a 4 agenti è elevata, il vantaggio di mantenere il *payoff* positivo, utilizzando strategie convenzionali, scompare.

Prendiamo ora in considerazione alcuni dati che emergono dall'approccio ABM.

Supponiamo che sopravviva una sotto-popolazione: in questa situazione, la sotto-popolazione può essere solo WatchRight o WatchLeft, suggerendo che, per dirla in termini di Lewis, queste sono soluzioni alternative auto-sufficienti a problemi di coordinamento. Si dimostra che questi *steady states* si verificano raramente. Perché? Una possibile risposta (suggerita dall'approccio EBM) potrebbe essere che il numero dei cammini (traiettorie) da stati iniziali a *steady states* ad una sotto-popolazione è bassa. Consideriamo ora il caso in cui siano due sotto-popolazioni a sopravvivere: in questo caso, come succede con le strategie evolutivamente stabili (ESS), se una sotto-popolazione è Hawk, l'altra sarà necessariamente Dove; Hawk non sopravvive con sotto-popolazioni di WatchRight o WatchLeft. Questi due ultimi risultati sembrano suggerire che le strategie incondizionate non siano simmetriche: anche se Hawk e Dove non sono auto-sufficienti, la prima (Hawk) può sopravvivere solo sfruttando la strategia più altruista (Dove). Al contrario, la secondo (Dove) è più adattabile, dato che può sopravvivere in interazione con una delle altre due sotto-popolazioni.

Quindi, per un gioco di congestione come quello discusso in questo capitolo, si potrebbe tentare di definire una gerarchia di adattabilità. In cima a tale gerarchia troviamo Dove: Dove può sopravvivere sia con Hawk che con gli agenti condizionati; Hawk può sopravvivere con Dove e, in fondo a tale gerarchia, troviamo gli agenti condizionati. In effetti, una strategia condizionata scaccia via le altre strategie condizionate. Hawk e Dove non sono strategie equivalenti: hanno *payoff* diversi, la cui combinazione può essere stabile a causa della complementarietà delle due strategie. Tre sotto-popolazioni sopravvivono, includendo una fra WatchRight e WatchLeft, ma non entrambe contemporaneamente.

A questo punto, si può sollevare una domanda piuttosto ovvia: *è possibile prevedere, in generale, lo steady state finale a partire dalla dimensione della popolazione*?

Noi abbiamo modellato analiticamente tutte le possibili interazioni nel nostro scenario, e abbiamo quindi calcolato il punto di equilibrio in funzione di N, la dimensione della popolazione (almeno in teoria). Il modello analitico consente di descrivere le relazioni tra le

strategie e i loro *payoff*, attraverso le equazioni a forma chiusa e consente di generalizzare i risultati delle simulazioni, permettendo previsioni accurate. Tuttavia, questo è un punto della discussione aperto.

Infine, una questione cruciale riguarda l'ABM: i risultati della simulazione sono organizzati in qualche struttura gerarchica, in quanto essi sono generati dai nostri algoritmi. Ad esempio, nel nostro modello, i risultati includono simmetrie peculiari: vale a dire strategie equivalenti non possono coesistere, mentre quelle non equivalenti lo possono fare. Un obbiettivo ambizioso potrebbe essere quello di tentare di spiegare questi risultati utilizzando l'EBM. I risultati generati dalle simulazioni ABM forniscono dati strutturati per lo sviluppo del modello analitico, attraverso il quale generalizzare i risultati delle simulazioni e poter fare previsioni. Le simulazioni ABM sono artefatti che generano dati empirici sulla base di variabili, proprietà, regole locali e fattori critici, che chi implementa il modello decide di utilizzare; in questo modo, le simulazioni permettono la generazione di dati controllati (una sorta di laboratorio virtuale), utili per mettere alla prova la teoria e ridurre la complessità; invece, l'EBM consente di *chiudere il cerchio*, rendendo così possibile l'eventuale falsificazione dell'approccio ABM.

In breve, Parunak, Savit e Riolo hanno sostenuto che ABM e EBM differiscono per:

i) i rapporti tra gli oggetti che modellano,

ii) il livello a cui tali oggetti operano.

Un'analisi dei passaggi che abbiamo seguito per realizzare il modello analitico fornisce utili consigli sul confronto tra ABM e EBM. Queste due classi di modelli non sono alternative, ma in alcune circostanze complementari; al di là delle considerazioni di ordine pratico fornite da Parunak, Savit e Riolo, siamo riusciti a rintracciare alcune caratteristiche che distinguono questi due modi di realizzare modelli.

7. Interiorizzare una norma[11]

7.1 Introduzione.

Il problema dell'*interiorizzazione* è un problema sotto esame da lungo tempo nel settore di ricerca delle scienze socio-comportamentali e della filosofia morale.

Negli ultimi tempi, il dibattito si è rinnovato all'interno dell'approccio razionale allo studio della cooperazione e della conformità, in quanto l'interiorizzazione è un sistema meno costoso e più affidabile dell'applicazione del *controllo sociale*. Ma come funziona? Finora, scarsa attenzione è stata dedicata alle *componenti mentali* dell'interiorizzazione. La prospettiva presentata in questo lavoro è evidentemente a favore di un modello cognitivo, ricco di diversi tipi, gradi e fattori di interiorizzazione.

Uno dei problemi intorno a cui si concentra l'attenzione degli scienziati sociali riguarda il modo in cui i sistemi autonomi, come gli esseri viventi, riescono ad esibire comportamenti positivi verso i propri simili e riescono a conformarsi a norme esistenti, specialmente dal momento che gli agenti egoisti sono avvantaggiati rispetto a quelli altruisti, a livello di competizione di gruppo. Fin dai tempi di Durkheim, la chiave di volta per risolvere questo problema è stata individuata nella teoria dell'*interiorizzazione delle norme* (Mead, 1963; Parsons, 1967; Grusec and Kuczynski, 1997; Gintis, 2003).

L'interiorizzazione delle norme è uno dei fili conduttori di tutte le discipline socio-comportamentali. Non soltanto i sociologi, ma anche gli psicologi dello sviluppo, quelli sociali e cognitivi hanno colto la centralità di questo argomento rispetto alla socializzazione. Sulle orme dei primi lavori di Vygotzky (pubblicati negli Stati Uniti alla fine degli anni settanta del secolo scorso) e di Piaget (1978), gli psicologi hanno mostrato che una attitudine parentale orientata a suscitare una interiorizzazione delle norme favorisce il futuro benessere dei figli ed anche le loro inclinazioni verso altri comportamenti altruistici (Ryan and Deci, 2000).

Tuttavia, la nostra definizione scientifica e la nostra comprensione del processo di

[11] Per gli argomenti qui di seguito trattati, si veda: Conte, R., Andrighetto, G., Campennì, M. (2010) "Internalizing Norms. A cognitive model of (social) norms' internalization", *International Journal of Agent Technologies and Systems (IJATS)*, 2 (1), pp. 63-73, IGI Global.

interiorizzazione normativa è ancora frammentaria e insufficiente. Lo scopo principale di questo capitolo è quello di sostenere la necessità di una *modellizzazione cognitivamente ricca*, per poter affrontare l'interiorizzazione normativa al fine di:

(a) fornire una visione unificata del fenomeno, che presenta alcune caratteristiche condivise con fenomeni collegati (ad esempio: robusta conformità, come in un comportamento automatico) e proprietà specifiche che lo mantengono distinto da essi (autonomia);

(b) fornire un modello del processo di interiorizzazione, ossia individuare le sue cause prossimali (rispetto alle distali, quelle evolutive - Gintis, 2003, 2004);

(c) caratterizzare l'interiorizzazione come un processo graduale, che si verifica a vari livelli di profondità e che dà luogo a conformità più o meno robusta;

(d) permettere una conformità flessibile, consentendo agli agenti di recuperare il pieno controllo (Bargh et al., 2001) sulle norme che sono state convertite in risposte comportamentali automatiche (Epstein, 2006).

Grazie a tale modello, sarà possibile adattare le architetture esistenti (come Emil-A - Andrighetto et al., 2007) e le piattaforme simulative (EMIL-S - Troitzsch, 2008), per verificare ipotesi relative

(a) agli effetti individuali e sociali di interiorizzazione,

(b) ai fattori che favoriscono o ostacolano l'interiorizzazione,

(c) all'evoluzione del processo di interiorizzazione nelle società future.

In questo capitolo, il *processo di interiorizzazione normativa è inteso come un processo mentale che prende le norme (sociali) come input e fornisce all'agente che interiorizza (da ora in poi, l'interiorizzatore) nuovi scopi come uscite*. Le emozioni, giocando un ruolo significativo ma non necessario in questo processo, non saranno nostro oggetto di indagine, in questa prima fase.

I contributi per spiegare l'interiorizzazione sono basati a volte sulla teoria dell'*apprendimento per imposizione*. Scott (1971), per esempio, ha teorizzato che l'interiorizzazione normativa conduce al rispetto robusto, a condizione che il sistema esterno sanzionatorio non sia mai completamente abbandonato. Purtroppo, questa spiegazione non è

sufficiente: è incompatibile non solo con l'idea che "*...le norme sociali possono ottenere interiorizzazione nella misura in cui non hanno bisogno dell'imposizione sociale*" (Basu, 1998), ma anche con prove sperimentali. Per esempio, i soggetti che sono coinvolti in esperimenti di *giochi di ultimatum*, si trovano spesso a seguire considerazioni di equità, anche quando inosservati (Bicchieri, 2006).

In questi ultimi anni, nel campo della teoria dei giochi evolutiva (*evolutionary game theory*) è apparsa una forte ripresa di interesse intorno al concetto di interiorizzazione normativ,a, per lo studio della cooperazione e del comportamento pro-sociale.

Gintis (2003) ha sostenuto che l'aumentare della complessità sociale della prime società umane ha prodotto un ambiente in rapida evoluzione, che a sua volta ha posto un problema di adattamento ai meccanismi genetici nella modifica degli scopi. L'interiorizzazione delle norme è adattiva, perché "*facilita la trasformazione di pulsioni, bisogni, desideri e piaceri in forme che siano più strettamente in linea con la massimizzazione della fitness*". Ma come si è evoluta?

Alcuni autori (Bicchieri, 2006; Epstein, 2006) hanno concepito l'interiorizzazione normativa come un processo che porta ad una sorta di conformità automatica, o inconsapevole (*thoughtless*). Le persone, ha osservato Epstein (2006), si conformano ciecamente alla norma: più lo hanno fatto in passato, più saranno disposti a farlo in futuro. Gli agenti non solo imparano a quali norme conformarsi, ma anche a quanto devono pensare ad esse. Dal punto di vista dell'autore, l'interiorizzazione non comporta il dover imparare a *pensare alle norme*.

Bicchieri (2006) ha fornito una spiegazione sofisticata, che conduce ad una conclusione equivalente. Gli agenti imparano cosa sono le norme attraverso le aspettative condivise, alle quali per definizione essi preferiscono corrispondere. Una volta scoperto cosa sono le norme, organizzano le proprie credenze in strutture simili a *script*, compresi i contesti, le relazioni e le condizioni in cui hanno trovato le norme. Quando, in seguito, le strutture di credenze (*script*) saranno attivate da attività in corso o contesti, le norme corrispondenti verranno anch'esse attivate.

Il modello di Bicchieri era destinato ad essere un modello di *conformità normativa*. La domanda è: lo stesso vale anche per quanto riguarda *l'interiorizzazione normativa*, vale a dire il fenomeno rilevato da Durkheim, Hart e definito da Scott come conformità "ad una temporale o spaziale rimozione dalle sanzioni" (1971, p. XIII)?

Il presente capitolo ha lo scopo di proporre modelli basati su agenti, e in particolare agenti dotati di una ricca modellizzazione cognitiva, come un approccio utile per gettare le basi di una teoria della interiorizzazione normativa. In questo quadro, è possibile prendere in considerazione le cause prossimali, cioè le basi cognitive dell'interiorizzazione, e caratterizzare l'interiorizzazione normativa come un progressivo processo *multi-step*, che porta da norme *esternamente o meramente* applicate ai corrispondenti *scopi normativi*, *intenzioni* e *azioni* perseguite per se stesse. Al fine di comprendere questo processo, alcune nozioni preliminari dovono essere chiarite.

7.2 Dinamica degli scopi.

Le persone agiscono in base a predeterminati *scopi* (geneticamente stabiliti e cablati nelle loro menti), che sono modificati, ampliati o ridotti durante il corso della loro vita. Questo processo può essere causato da fattori non cognitivi, come i processi ormonali, sostanze chimiche, ma può anche provenire da meccanismi di apprendimento e ragionamento (Conte e Castelfranchi, 1995). Sotto l'effetto di fattori sociali, gli scopi possono essere generati nuovamente attravreso fattori cognitivi, come scopi relativizzati ad altri stati mentali (ad esempio, le credenze sociali). Uno scopo è *relativizzato* quando è perseguito in quanto e nella misura in cui uno stato del mondo o evento è atteso o previsto essere vero (Cohen e Levesque, 1990a). Domani, voglio andare a raccogliere i funghi (scopo relativizzato), in quanto e nella misura in cui credo che domani pioverà (evento atteso). Il preciso istante in cui io smetto di credere che domani pioverà, farà cadere ogni speranza che io trovi funghi. Nuovi scopi possono essere relativizzati rispetto alle credenze sociali. Questi sono scopi *sociali* relativizzati (Conte e Castelfranchi, 1995). Quando sono positivi o pro-sociali, il processo di generazione è chiamato *adozione di scopi*: un agente, l'adottante, genera un nuovo scopo g_i, perché e fino a quando crede g_i essere uno scopo dell'adottato. *L'interiorizzazione di scopi genera scopi che non sono più relativizzati alle credenze sociali.* Qualunque sia l'atteggiamento iniziale dell'interiorizzatore verso un determinato stato del mondo, finirà col perseguire un nuovo scopo indipendente dallo stato degli scopi dell'agente con cui interagisce. Inutile dire che questo processo non è necessariamente cosciente né razionale.

7.3 Dinamica mentale delle norme.

Il concetto di *norma*, cui si fa riferimento in questa sede, deriva da un quadro teorico ben consolidato (Conte e Castelfranchi, 1995, Conte e Castelfranchi, 2006) e dai risultati ottenuti dal progetto europeo (FP6) EMIL. Nello specifico, le norme sono intese come comportamenti che si diffondono attraverso una data popolazione, grazie alla diffusione delle corrispondenti credenze normative. Ad esempio, tenere allacciate le cinture di sicurezza durante la guida di un'auto è una norma se si diffonde sotto l'ipotesi che questo comportamento è una norma stabilita all'interno di qualcosa, in questo caso della legge sulla circolazione stradale. Nella prospettiva legata agli agenti autonomi, conformarsi alle norme richiede che gli agenti si formino le credenze normative e raggiungano gli scopi normativi, vale a dire, gli scopi relativizzato alle credenze normative. Il processo di adozione degli scopi è trasformato in adozione di una norma quando le credenze normative generano scopi normativi, di solito in riferimento ad un meccanismo di imposizione esterna (sanzioni, approvazione). Se un tale scopo non viene generato, la norma verrà violata. D'altra parte, una norma *è interiorizzata*, quando il destinatario di una norma è conforme ad essa indipendentemente dalle sanzioni e dalle ricompense esterne. In tal caso, lo scopo normativo non è più relativizzato ad un evento sociale percepito (una sanzione), ma solo ad una credenza normativa.

7.4 Tipi e gradi di interiorizzazione.

L'interiorizzazione normativa può avvenire a diversi livelli, riguardanti diversi aspetti della mente. Per capirli, è necessario un breve glossario. In tutto il capitolo, parleremo di scopi, intenzioni e azioni dal punto di vista della scienza informatica e della teoria degli agenti autonomi. In particolare, lo *scopo* è uno stato del mondo desiderato che innesca l'azione e la guida (Conte, 2009); l'*intenzione* è un eseguibile stato del mondo desiderato, scelto per l'esecuzione; un'*azione* è l'esecuzione di una intenzione (al riguardo, si veda la sezione sulle azioni in computer science in Segerberg et al., 2009), vale a dire, l'intenzione incorporata in un dato comportamento. Vediamo allora come questi concetti si possano specificare ulteriormente, nel contesto dell'interiorizzazione.

– *Scopo normativo interiorizzato*: lo scopo normativo non è più relativizzato ad

un'imposizione esterna, ma solo ad una credenza normativa.

- *Scopo interiorizzato*: questa volta lo scopo non è più normativo; vale a dire, non è più relativizzato ad una credenza normativa.

L'interiorizzatore ha perso le tracce dell'origine normativa del suo scopo. La credenza normativa permane ancora, ma l'agente persegue lo scopo corrispondente indipendentemente da esso. Per esempio, io posso adottare la norma di fermarmi ad un incrocio, perché non voglio prendere una multa. Fino a qui, ho generato uno scopo normativo relativizzato ad una sanzione esterna. Se non vedo poliziotti in giro, posso proseguire ignorando la norma. Supponiamo che io poi a poco a poco realizzi (forse grazie all'effetto di una coppia di incidenti stradali non letali, ma gravi) che non arrestarsi ad un incrocio è pericoloso. Se mi fermo sempre, anche se non c'è nessun poliziotto nei dintorni, ho interiorizzato la norma. Se mi fermo anche se so che la norma mi chiede solo di rallentare, ho trasformato la norma in uno scopo interiorizzato.

- *Intenzione interiorizzata*: l'output è uno scopo non più relativizzato ad una credenza normativa, scelto per l'esecuzione e attivato da uno specificato set di eventi percepiti.
- *Azione interiorizzata*: l'output è un'azione condizionata innescata da un *trigger*, da un evento percepito (per esempio, fermarsi quando il semaforo è rosso).

Il processo decisionale è evitato, in quanto il *trigger* attiva una azione condizionata nel repertorio dell'interiorizzatore. È interessante notare, tuttavia, che sotto l'effetto di altri eventi percepiti, le azioni condizionate possono essere bloccate o congelate per l'intervallo di tempo necessario per elaborare un disturbo o interferenza all'evento, per poi ripristinarlo in seguito (Bargh et al., 2001) in una maniera semi-consapevole. Qui, non solo la volontà ma anche la risposta comportamentale è automatica. Nell'esempio del semaforo, si tratta della sequenza dei movimenti necessari per attivare l'arresto della vettura, una risposta comportamentale così profondamente interiorizzata, che difficilmente si può rendere esplicita. Prima di mostrare come una architettura BDI tenga conto di queste diverse forme di interiorizzazione, cerchiamo di analizzare i fattori che potrebbero facilitare la loro comparsa.

7.5 Fattori

I fattori che incidono sulla interiorizzazione devono essere studiati cross-metodologicamente, confrontando le simulazioni (ad agenti) con esperimenti su soggetti reali. Quelle di seguito sono solo ipotesi preliminari.

7.5.1 Scopo normativo interiorizzato.

Perché gli agenti osservano una norma a prescindere da un'imposizione esterna?

Suggeriamo che il fattore principale di questo tipo di interiorizzazione sia la *coerenza* (*consistency* - McAdams, 2008). Questo meccanismo funziona in due fasi: prima selezionando quale norma debba essere interiorizzata, e poi controllando che il comportamento corrisponda ad essa (*self-control*) e imponendoselo (*self-enforcement*). La coerenza delle nuove norme con le proprie credenze, scopi e norme precedentemente interiorizzati, svolge un ruolo cruciale nel processo di selezione. Con successo le strategie educative favoriscono i processi di interiorizzazione (King, 2008), spesso mettendo in relazione nuovi input con le norme precedentemente interiorizzate. Considerazioni analoghe valgono per il *policy-making*.

Si consideri la normativa antifumo: l'efficacia delle campagne antifumo sulla base di annunci-*shock* e di etichette allarmiste (ad esempio, frasi come “Il fumo uccide” sui pacchetti di sigarette - Goodall, 2005) è ancora controversa. Uno dei fattori che riduce l'efficacia è l'effetto noto come *sconto iperbolico* (Bickel e Johnson, 2003; Rachlin, 2000), un meccanismo psicologico che porta a investire nella realizzazione dei propri scopi una quantità di impegno che è una funzione iperbolicamente decrescente della distanza nel tempo dello scopo da raggiungere, e porta la gente a procrastinare gli sforzi necessari (al conseguimento di tale scopo). A causa dello sconto iperbolico, le persone, soprattutto i giovani, non sono in grado di agire in base alla rappresentazione delle conseguenze ritardate delle correnti azioni. Molto più efficace sembra essere un diffuso set di norme sociali precedentemente emerso, i precetti del vivere-in-modo-sano, estremamente coerenti con la normativa antifumo.

La coerenza è fondamentale anche per l'efficacia di auto-imposizione. Questa volta, lo sconto iperbolico opera a favore della interiorizzazione normativa. Gli agenti possono trovare più facile impegnarsi in un determinato corso d'azione che eseguire tale azione nella realtà: posso mostrare più entusiasmo nel promettere a me stesso di smettere di fumare domani, che

non nel mantenere la promessa con tenacia quando arriva il momento. Tuttavia, la mia promessa, che è stata facilitata dalla sconto iperbolico, attiverà uno scopo di mantenimento (Cohen e Levesque, 1990), vale a dire mantenere la promessa fatta a me stesso (o più varianti significative di ciò, come preservare l'autostima, essere coerenti, non perdere credibilità davanti agli occhi degli altri). Grazie allo sconto iperbolico, ho fatto un passo rischioso: ho creato per me un nuovo scopo. Questo è sufficiente, se non per la sua integrazione nel mio comportamento, almeno per l'attivazione dell'auto-imposizione, che è un processo che segue la selezione degli *inputs*. L'auto-imposizione sostiene il processo di esecuzione. L'interiorizzatore si assumerà il compito dell'esecuzione, e inizierà ad applicare l'auto-punizione, quando viola la nuova norma e l'auto-ricompensa, quando rispetta tale norma, sia in termini di auto-valutazione che in termini di emozioni e sentimenti negativi verso se stesso.

7.5.2 Scopo interiorizzato.

L'interiorizzazione della fonte (normativa) dipende da una serie di fattori.

L'*effetto di auto-rafforzamento* del rispetto della norma: il destinatario della norma si rende conto che raggiunge uno dei suoi scopi, osservando una certa norma. Supponiamo che io riesca a astenermi dal fumo e che dopo pochi giorni, mi renda conto di scoprire un vantaggio non previsto: il cibo comincia di nuovo ad essere gustoso. Questa scoperta genera uno scopo (smettere di fumare, per godere del buon cibo), non relativizzato alla norma, ma che lo sostiene: ho trasformato la norma in uno scopo ordinario (se questo scopo sarà abbastanza forte da farmi uscire dalla mia dipendenza, questa è un'altra questione).

La *salienza* della norma (Campennì et al., 2008) è un altro fattore di interiorizzazione. Questa è definito come il numero di volte che una qualsiasi determinata norma viene osservata e difesa per unità di tempo (Troitzsch, 2008). Maggiore è la salienza della norma, più profondamente è interiorizzata.

Un buon esempio è l'osservanza di un particolare regime alimentare. Un animalista decide di adottare una dieta vegetariana per motivi etici. Dopo un po', si sarà disabituato al sapore della carne, e anche al più debole odore di essa.

Questo fenomeno è probabilmente all'origine di una maggiore efficacia dei convertiti nel fare proseliti (Levine e Valle, 1975).

7.5.3 Intenzione interiorizzata.

Quanto più una data norma permette di rispondere a problemi incontrati di frequente in condizioni di urgenza (quando il tempo per il processo decisionale non c'è o è scarso), tanto più è probabile che la norma sarà interiorizzata come un'intenzione, uno scopo scelto per l'esecuzione.

7.5.4 Azione interiorizzata.

Le norme sono convertite in azioni interiorizzate sotto l'effetto di diversi fattori convergenti: la *salienza*, l'*esplicitezza* e l'*operatività*.

Le norme o descrivono gli stati del mondo che (non) devono essere raggiunti senza fare esplicito riferimento a come questo debba accadere ("*tieni la tua camera pulita*"), o le azioni che (non) devono essere realizzate ("*Fai silenzio!*"). In quest'ultimo caso, tanto più la norma è saliente, esplicita e operativa, tanto più probabilmente sarà interiorizzata come azione condizionata, una routine attivata sotto determinate condizioni. Così, gli agenti si coprono la bocca, mentre sbadigliano, sorridono e/o pronunciano una formula di saluto quando entrano in uno spazio privato aperto al pubblico. La norma non solo può essere interiorizzata in routine standard o abitudini, ma anche incorporata in artefatti materiali (argenteria, fazzoletti) che attivano quelle routine. Non è questo tipo di interiorizzazione, dopo tutto, che corrisponde a ciò che Epstein ha chiamato conformità inconsapevole (*thoughtless conformity*)? Probabilmente, si.

Tuttavia, questa non è l'unica forma di interiorizzazione. La questione cruciale, qui, è quella di fornire un terreno comune, un modello di agente che possa esibire tutte queste forme, e, quel che è più difficile, che possa passare dall'una all'altra. Abbiamo bisogno di rendere conto di routine reversibili, o, in altri termini, di una conformità flessibile. Come può l'interiorizzatore riconquistare nuovamente il pieno controllo su un'azione automatizzata, e astenersi dall'applicare una data routine? Come possiamo procedere quando il semaforo è rosso, ma il vigile urbano ci invita a farlo ugualmente? Anche se un'altra routine viene attiva dal nuovo evento (vigile urbano che invita a procedere), come e perché una delle routine (*fermati al rosso*) viene interrotta e la sua complementare (passare) viene attivata? Come è risolto il conflitto? Come possiamo decidere comportamenti automatici?

Infatti, il confine tra comportamenti automatizzati e consapevole volontà è sfumato più di quanto comunemente non si creda. Alcuni psicologi cognitivi (Bargh et al., 2001) hanno mostrato che il comportamento automatico non deve essere rigidamente rimosso dalla coscienza, e che la realizzazione di uno scopo non deve essere (necessariamente) cosciente né intenzionale. In linea di principio, una architettura normativa modulare si adatta bene ad una automaticità flessibile: le azioni interiorizzate non impediscono alle norme di essere trattate a livelli cognitivi superiori. *Input* che interferiscono negativamente con una data routine possono contemporaneamente attivare una credenza normativa e portare al conseguimento di uno scopo normativo. Sarà quindi il *decision-maker* a stabilire se il nuovo scopo debba essere raggiunto o la vecchia routine ristabilita. E 'anche possibile che il nuovo scopo attivi un'altra routine, in conflitto con quella precedente.

Come combinare il conseguimento consapevole e automatico degli scopi in architetture di agenti intelligenti è una questione molto affascinante, ben oltre la portata di questo capitolo. Tuttavia, questa è una delle idee più ispirate alla base del lavoro sulle promettenti architetture cognitive ibride (Sun e Wu, 2006).

7.6 L'interiorizzatore: un'architettura BDI.

L'architettura normativa Emil-A (Andrighetto et al. 2007) – della quale abbiamo già detto nel quinto capitolo – consiste di meccanismi e rappresentazioni mentali che permettono alle norme di incidere sul comportamento di agenti intelligenti autonomi. Emil-A è stata pensata per mostrare che non solo le norme regolano il comportamento, ma anche agiscono su diversi aspetti della mente: il *riconoscimento*, l'*adozione*, la *pianificazione*, e il *processo decisionale*.

Come qualsiasi tipo di architettura BDI, EMIL-A opera attraverso moduli adibiti a diversi sotto-compiti (riconoscimento, adozione) e agisce su rappresentazioni mentali in base a scopi e credenze in una sequenza non rigida. Per mostrare come Emil-A lavori, forniamo la descrizione dell'ideale e completo percorso mentale di una norma. Dopo il riconoscimento (figura 1 – cfr. Campennì et al. 2008), una norma diventa una *credenza* memorizzata nel *normative board* (N-Board) di un agente; la credenza rappresenta un obbligo, una prescrizione o un divieto, rispetto ad un determinato stato del mondo o ad una determinata azione. La N-Board è una porzione della memoria a lungo termine, in cui le credenze normative sono

memorizzate e classificate in base alla loro salienza. Una credenza normativa diventerà l'*input* per il modulo di adozione (la freccia tratteggiata in figura 2). Sotto la pressione di una imposizione esterna, verrà generato uno scopo normativo, a partire dalla credenza normativa relativa alla imposizione prevista.

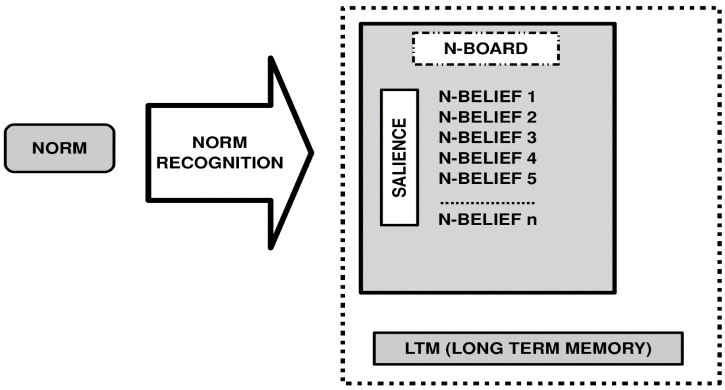


Figura 1. Riconoscimento di una norma.

Sotto l'effetto di fattori come l'alta coerenza, la credenza normativa sarà sufficiente a generare lo scopo normativo (la freccia nera piena figura 2).

Una volta formato, uno scopo normativo viene dato in *input* al *decision-maker* e confrontato con altri scopi eventualmente attivi nel sistema. Il *decision-maker* sceglie quello da eseguire e lo converte in una intenzione normativa (vale a dire uno scopo eseguibile). Una volta eseguito, questo darà luogo a *conformità-ad-una-norma* e/o *difesa-di-una-norma* (punizione diretta o indiretta) e/o *trasmissione di una norma* attraverso la comunicazione.

Altrimenti, la credenza normativafinirà per essere abbandonata, soluzione che porta di nuovo alla violazione della norma. Questa caratterizzazione, ricca di rappresentazioni e processi alla base di un comportamento conforme-a-norma, non deve dare l'idea che la conformità del comportamento è *sempre* basata su un ragionamento complesso e così deliberativo.

Un aspetto cruciale di Emil-A è la possibilità di spiegare la frequenza delle interruzioni, modifiche e deviazioni dai processi descritti finora: la conformità normativa e l'obbedienza possono essere convertite in uno scopo interno, in intenzione e anche diventare un'abitudine,

un (semi)automatismo, un comportamento di routine.

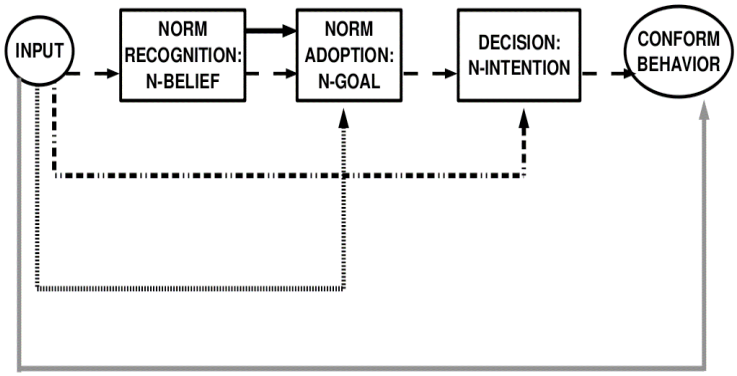


Figura 2. EMIL-A e l'interiorizzazione di una norma.

7.7 Ipotesi e questioni supplementari.

Possiamo ora azzardare alcune ipotesi in base alle caratteristiche del fenomeno dell'interiorizzazione, così come lo abbiamo descritto. In una certa misura, i vantaggi forniti dall'interiorizzazione delle norme sono facilmente identificabili: il rispetto di una norma dovrebbe essere più robusto, se le norme possono essere interiorizzate rispetto a quanto avviene quando le condotte sono governate solo da sanzioni esterne; l'interiorizzazione emancipa i destinatari della norma da sanzioni esterne.

Tuttavia, che cosa dobbiamo aspettarci dal confronto tra gli stati mentali interiorizzati e quelli completamente endogeni? Gli scopi interiorizzati abbiamo ipotizzato che siano più persistenti e portino ad una realizzazione dello scopo più vigorosa (Bargh et al., 2001), di quanto non facciano gli scopi originari. L'argomento si basa sulla teoria della *prospettiva* (Kahneman e Tversky, 1979; Abdellaoui et al., 2007), che presuppone l'avversione alla perdita, ossia la tendenza della gente a preferire con forza evitare le perdite rispetto all'acquisizione di utili, come una caratteristica importante degli esseri umani.

Gli scopi interiorizzati sono già formati nella mente: a differenza degli scopi pienamente endogeni, quelli interiorizzati sono scelti tra gli scopi iniziali acquisiti sotto l'effetto di una influenza esterna. Maggiore sarà lo sforzo investito nella realizzazione di questi scopi, con minore probabilità essi saranno abbandonati in seguito, più vigorosamente saranno raggiunti.

Una specifica ipotesi si basa sul *bias di conferma* (in base al quale le persone sono propense ad accettare *input* che confermano le loro convinzioni, e a respingere quelli che non lo fanno - Nickerson, 1998; Sternberg, 2007). In base ad esso, l'interiorizzatore dovrebbe mostrare maggiore intolleranza - per quanto riguarda la violazione di una norma - rispetto al caso di coloro che seguono la norma spinti da imposizioni esterne, o al caso di coloro che sono spontaneamente mossi a comportarsi in accordo con essa. La violazione è un elemento di disturbo per la l'interiorizzatore, che potrebbe portare a indebolire se non addirittura a rivedere l'impegno assunto. Quindi, gli interiorizzatori di norme dovrebbero essere più coerenti e conformi ad una norma di coloro che la osservano in base ad una imposizione esterna e agli agenti endogenamente motivati.

Un'ulteriore conseguenza di questa teoria è che gli agenti sono molto più disposti a difendere le norme interiorizzate di quanto non lo siano coloro che osservano una norma grazie ad una imposizione esterna. Una conseguenza è che l'interiorizzazione di una norma è decisiva, se non indispensabile, per il *controllo sociale distribuito*. L'interiorizzazione probabilmente non è solo un meccanismo di conformità privata, ma anche un fattore di controllo sociale. In breve, l'interiorizzazione è un buon predittore della conformità e della cooperazione di secondo ordine (cioè invitare gli altri a rispettare le norme - Horne, 2007).

Ma quali sono, eventualmente, gli svantaggi dell'interiorizzazione? Ancora una volta, la teoria porta a formulare alcune ipotesi.

In primo luogo, l'interiorizzazione richiede tempo, e non è necessariamente positiva: l'auto-addestramento può essere troppo difficile, e richiede spesso diverse prove prima di ottenere risultati. La gente quasi mai abbandona cattive abitudini e comportamenti antisociali in poco tempo.

In secondo luogo, gli errori possono avere effetti controproducenti: dopo un certo numero di tentativi infruttuosi, la perdita di autostima ed i sentimenti di impotenza possono rendere troppo deboli e fragili futuri impegni privati, e mettere in pericolo l'interiorizzazione. La questione cruciale diventa: quali sono i fattori che possono favorire il successo al primo tentativo?

In terzo luogo, l'interiorizzazione presuppone la selezione degli *input*, cioè l'autonomia. L'autonomia morale è spesso incoraggiata, almeno nelle società occidentali, ma può avere anche effetti controproducenti. Per esempio, può portare ad una varianza eccessiva nel

rispetto.

In quarto luogo, come può operare l'interiorizzazione in società caratterizzate da un elevato grado di norme, eventualmente in aspro conflitto l'una con l'altra? Ci si può aspettare che l'interiorizzazione diventi incompatibile con la stessa norma percepita e generi conflitti di valore. Potrebbe essere che il futuro della società sia caratterizzato da interiorizzazioni fragili e variabili?

7.8 Discussione dei risultati

Quando Vygotskij per primo formulò la sua teoria sull'interiorizzazione, egli osservò che solo "il contorno spoglio di questo processo è noto" (1978, p. 57). Non sappiamo, tuttavia, come le persone riescano a interiorizzare le credenze ed i precetti, con un successo ragionevolmente adeguato; in parte perché ancora non siamo d'accordo su cosa studiare, cioè su cosa si intenda con il concetto di interiorizzazione. Di conseguenza, nessuna idea utile o modello è disponibile per eventuali applicazioni, nonostante le implicazioni sul piano generale siano profonde. Questioni, ad esempio, su come avvenga l'interiorizzazione normativa, quali siano i fattori che la determinano, quali siano i suoi effetti, gli ostacoli e le contro-indicazioni, sono questioni di interesse per tutte le scienze comportamentali. L'interiorizzazione di *input* sociali è indispensabile per lo studio e la gestione di un ampio spettro di fenomeni, dallo sviluppo di una forte autonomia morale alle indagini e all'esecuzione del controllo sociale distribuito, dalla soluzione del problema della cooperazione, alla promozione della sicurezza e della lotta alla criminalità. L'approccio simulativo può essere molto utile: ci spinge a descrivere il processo di interiorizzazione in maniera chiara ed analitica come richiesto dalla realizzazione di un modello computazionale espresso sotto forma di codice.

Dopo qualche puntualizzazione sull'approccio cognitivo, in questo capitolo abbiamo presentato e discusso gli elementi costitutivi di un modello cognitivamente ricco di interiorizzazione, un progressivo processo a più stadi, inclusi diversi tipi e gradi di interiorizzazione. Abbiamo discusso i fattori che favoriscono i diversi tipi di interiorizzazione. Il carattere modulare delle architetture BDI, come Emil-A, è particolarmente indicato per implementare l'approccio sostenuto. In studi futuri, l'architettura Emil-A potrà essere consolidata da un modello di interiorizzazione lungo le linee qui presentate, per eseguire simulazioni multi-agente, confrontando l'interiorizzazione con altre forme di rispetto

normativo, attraverso una serie di misure (tra cui la convergenza, la robustezza e l'adattamento ad una rapida evoluzione o ad un cambiamento dell'ambiente).

Qual è il valore aggiunto di un modello cognitivamente ricco dell'interiorizzazione, rispetto a quelli semplici (ad esempio, l'apprendimento per rinforzo)? Ci sono diversi vantaggi.

In primo luogo, l'apprendimento per rinforzo tiene conto dell'intuizione principale condivisa da diversi autori, cioè l'idea che l'interiorizzazione renda *indipendente la conformità dall'imposizione esterna*.

In secondo luogo, un modello cognitivamente ricco (come un'architettura BDI con i suoi moduli specifici) *tiene conto della possibilità di diversi tipi e gradi di interiorizzazione*, colmando il divario tra auto-imposizione e risposta automatica.

In terzo luogo, un'architettura BDI, tenendo conto dei diversi livelli di interiorizzazione, *permette di combinare una certa flessibilità con gli automatismi*, nonché la conformità inconsapevole (*thoughtless*) con l'autonomia.

8. Evoluzione della Mente Normativa

Nei precedenti capitoli di questo lavoro abbiamo visto come sia possibile realizzare un modello (computazionale) dell'emergenza di artefatti sociali in società artificiali.

Nello specifico, abbiamo mostrato come in alcuni casi sia sufficiente fare ricorso agli strumenti consolidati della teoria dei giochi (per esempio, per implementare un modello di emergenza delle convenzioni – cap. 6); in altri casi, invece, abbiamo visto come sia necessario postulare l'esistenza nell'agente di una complessa architettura cognitiva in modo da capire come possa emergere una nuova norma sociale in un gruppo di individui (cap. 5).

I modelli computazionali rappresentano uno strumento molto potente e versatile per lo studio dei fenomeni sociali. Essi garantiscono a chi costruisce il modello una piena libertà decisionale e permettono di fissare con estrema accuratezza il livello di indagine desiderato; chi realizza il modello può decidere di interessarsi maggiormente alle dinamiche globali del fenomeno in questione, oppure alle micro-interazioni locali dei singoli elementi che contribuiscono al fenomeno, in una scala di complessità che va dal micro al macro, passando per una serie di gradazioni intermedie.

Negli ultimi anni, il dibattito scientifico sull'argomento "norme" ha riscosso notevole attenzione, spingendo numerosi studiosi operanti in diversi settori scientifici a proporre ipotesi.

Il concetto di *norma sociale* è uno dei concetti più importanti e centrali nell'ambito delle scienze del comportamento; interessa un ampio spettro di discipline e cattura l'attenzione di economisti, sociologi, scienziati politici, antropologi, psicologi, giuristi, filosofi e biologi evoluzionisti.

A conclusione della nostra indagine è ragionevole domandarsi quale sia, in natura, l'origine evolutiva della propensione alla normatività, tipica della mente umana (e forse anche di altri animali). Quale può essere la spiegazione evolutiva della capacità di generare gli artefatti sociali che regolano la nostra vita quotidiana?

In questo capitolo passeremo in rassegna alcune delle ipotesi più accreditate, riferendoci in particolare a quelle proposte scientifiche che prevedono l'utilizzo di modelli computazionali per testare la validità delle proprie assunzioni.

8.1 *Evolutionary Game Theory*

Parte della risposta all'interrogativo circa le origini naturali delle norme sociali risiede probabilmente nella possibilità che queste offrono di favorire il successo adattivo degli individui, dato un particolare contesto socio-ambientale. In questo senso, l'emergere di uno dei più importanti comportamenti (pro)sociali (il comportamento normativo) può essere inteso come una risposta adattiva, la quale aiuta l'individuo a raggiungere un maggior benessere individuale.

L'*Evolutionary Game Theory* (EGT) rappresenta uno dei principali quadri teorico-metodologici nei quali si è affermata questa visione delle norme sociali.

Kameda et al. (2005) forniscono un'interessante analisi dell'origine evolutiva delle norme sociali, presentando un modello computazionale in cui la norma sociale emerge in quanto massimizza la *fitness* individuale.

La condivisione di risorse importanti (come il cibo, ad esempio) è una pratica ben consolidata nelle società umane e in alcune specie di primati non umani; tuttavia, solo gli esseri umani presentano un comportamento così fortemente pro-sociale anche nei confronti di non consanguinei. Gli esseri umani condividono le risorse di cibo più di ogni altro organismo. Diversi altri animali, compresi gli insetti sociali, i carnivori sociali ed alcune specie di uccelli e pipistrelli condividono il cibo; tuttavia, il modo in cui questa pratica è attuata dagli esseri umani è davvero unico. A differenza degli altri mammiferi, per i quali la condivisione di cibo fra madre e prole è strettamente limitata alla fase dell'allattamento, i genitori umani si prendono cura dei propri figli fino all'età adulta. Infatti, lo scambio di cibo fra genitori e figli continua in modo bidirezionale fino alla morte nella maggior parte delle società umane. Inoltre, lo stesso matrimonio può essere visto come una istituzionalizzata condivisione di risorse fra moglie e marito. La condivisione di cibo nelle società umane si basa appunto sulla divisione dei compiti, sull'età e sulle differenze di genere.

La maggior parte delle popolazioni umane sono suddivise in gruppi etnici caratterizzati da tratti apparentemente arbitrari, quali i costumi o le lingue. Le spiegazioni esistenti non colgono adeguatamente il meccanismo che sta sotto all'origine ed al mantenimento della marcatura del gruppo. McElreath et al. (2003) hanno sviluppato un modello matematico il quale mostra come gruppi - che si distinguono in base a differenze rispetto alle norme sociali

e ai *marker* arbitrari che adottano - possano emergere e rimanere stabili nonostante significativi mescolamenti, a patto che:

i) gli individui interagiscano preferibilmente con altri individui che presentano gli stessi *marker* sociali e

ii) gli individui acquisiscano i propri *marker* e comportamenti sociali imitando gli individui che hanno maggior successo.

In questo modello di evoluzione, gli individui utilizzano i *marker* per decidere con chi interagire. Precisamente, nel modello non si assume che i *marker* permettano agli individui di indirizzare selettivamente comportamenti altruistici verso individui della stessa etnia, ma si assume che i *marker* permettano agli individui di interagire con altri individui che condividono le stesse norme sociali. L'assunzione di base è che molte norme sociali possono essere intese come convenzioni e, per questo, modellate attraverso giochi di *coordinamento* piuttosto che di *cooperazione*. I giochi di coordinamento (Sugden 1986, Binmore 1994) hanno luogo quando le interazioni fra individui che condividono credenze circa il modo in cui bisognerebbe comportarsi producono *payoff* maggiori rispetto a quelli prodotti dalle interazioni fra individui che hanno credenze discordanti. McElreath e colleghi mostrano che:

i) se gli individui interagiscono preferibilmente attraverso giochi di coordinamento con individui che hanno gli stessi *marker* e

ii) se essi acquisiscono i loro *marker* e comportamenti coordinati imitando gli individui che hanno maggior successo,

iii) allora possono emergere dei gruppi che si distinguono sia per le norme sociali che per i *marker* che adottano e tali gruppi possono rimanere stabili nonostante significativi mescolamenti (in termini di individui).

Questo in sintesi è il modello. I processi di interazione sociale, imitazione, migrazione e ricombinazione avvengono ciclicamente in sequenza. Ovviamente, nella natura tutti questi processi hanno luogo in parallelo e in maniera apparentemente continua; tuttavia, com'è noto, quando si realizza un modello di un fenomeno reale è necessario considerare eventi discreti e processi seriali.

Secondo il modello esistono due caratteri culturali discreti che vengono trasmessi:

1) un tratto comportamentale che influenza direttamente il benessere degli individui;

2) un *marker* caratteristico osservabile, che non ha effetti diretti sul benessere, ad

eccezione del fatto che può essere utilizzato dagli altri agenti per individuare il gruppo di appartenenza.

Ad ogni ciclo gli individui interagiscono e ricevono un *payoff* dall'interazione; quindi, osservano il *payoff* di un altro individuo scelto a caso e con una certa probabilità copiano il comportamento ed il *marker* di costui, se quest'ultimo ha ottenuto un *payoff* maggiore. In questo modo, alcuni individui abbandonano il loro gruppo ed emigrano in un altro. Così, gli individui acquistano i *marker* e i comportamenti da individui diversi; ciò determina un mescolamento dei *marker* e dei comportamenti. Questo processo complesso ha l'effetto di far emergere combinazioni più o meno stabili di *marker* e comportamenti, che caratterizzano gruppi etnici diversi.

Gruppi etnici differenti sono presenti in ogni periodo storico; mentre alcuni si sono spesso divisi o riuniti nel corso del tempo, altri hanno dimostrato una ragguardevole continuità storica. Dal momento che nessun altro primate si comporta in questo modo, formando gruppi basati sui costumi e sulla lingua (e devono quindi esistere forze che si oppongano all'aggregamento di gruppi diversi), ancora non è chiaro come gli esseri umani si siano potuti organizzare socialmente in questo modo.

Si pongono in particolare due problemi.

In che modo i gruppi vengono identificati da *marker* apparentemente neutrali? Come è possibile che le differenze fra i gruppi perdurino, anche quando (come spesso avviene) hanno luogo interazioni fra gruppi diversi, inclusi scambi, matrimoni e guerre?

Una ipotesi è che gruppi diversi si siano potuti evolvere parallelamente perché isolati gli uni dagli altri. Tale spiegazione sembra poco plausibile perché la maggior parte dei gruppi etnici conosciuti non sono stati sufficientemente isolati da poter generare la propria cultura indipendentemente da quella dei vicini (Nettle 1996).

Un'altra possibilità è rappresentata dalla scelta consapevole che gli individui possono fare dei *marker*; possono usarli per esempio per mostrare la loro propensione verso una certa cultura o economia (Barth 1969). Molti individui adottano chiaramente *marker* sociali per poter avere accesso a determinati benefici. Nel momento in cui gli individui hanno poi accesso ad identità etniche multiple, possono trarre benefici su più fronti; tuttavia, questa sorta di uso strategico dei *marker* non può di per se produrre differenziazione tra i *marker*, in quanto questi ultimi devono già avere valore informativo prima che chiunque li usi per

segnalare l'appartenenza ad un certo gruppo etnico.

Boyd e Richerson (1987) hanno ipotizzato che i marker etnici possano evolvere come *adattamento* ad ambienti spazialmente separati, in quanto essi potrebbero permettere di apprendere selettivamente, incrementando la possibilità di acquisire localmente informazioni utili.

Un certo numero di autori ha suggerito che i *marker* etnici possano garantire agli individui di essere identificati come membri di un particolare gruppo, consentendo loro di cooperare selettivamente e di comportarsi etnocentricamente nei confronti di altri gruppi (Nettle and Dunbar 1997). Nettle e Dunbar ipotizzano che la selezione possa favorire gli individui che indirizzano comportamenti altruistici verso gli individui che presentano i loro stessi *marker* sociali, perché questo permette agli altruisti di beneficiare del comportamento di altri altruisti. Il modello che questi autori presentano suggerisce che un processo del genere funziona soltanto se i *marker* etnici vengono trasmessi con un errore elevato (poco plausibile). Tale conclusione è consistente rispetto ai lavori effettuati in biologia evolutiva che indicano che i *marker* ereditabili non possono funzionare per discriminare un comportamento (realmente) altruistico da uno altruistico solo in apparenza, poiché i truffatori (cioè gli individui che mostrano il *marker* dell'altruismo ma non si comportano altruisticamente) ottengono un *payoff* maggiore di quelli che sono marcati come altruisti, essendolo veramente (Grafen 1990); in questo modo, non esisterebbe alcuna pressione evolutiva che favorisca l'affermazione del comportamento (realmente) altruistico.

Kaplan e Hill (1985) hanno mostrato che una stessa popolazione può esibire norme sociali diverse in base al tipo di risorsa condivisa; se le risorse vegetali (come frutta e verdura) sono facilmente condivise anche con non consanguinei, il bottino di una battuta di caccia viene diviso solo con i parenti. Una possibile spiegazione di questa diversificazione del comportamento in base alla risorsa da condividere sembra risiedere nella frequenza con la quale essa possa essere ottenuta: dal momento che la preda è una risorsa molto più incerta dei frutti della terra, diventa necessario ridurre le dimensioni del gruppo sociale con cui condividerla.

Nel modello simulativo presentato, questi autori cercano di mostrare quanto sia evolutivamente stabile il comportamento legato alla condivisione di cibo in condizioni di incertezza ambientale. Va esplicitato che l'EGT differisce dalla teoria dei giochi classica in

quanto assume che gli agenti non siano in possesso di informazioni complete e perfette. Ciascun individuo sarà maggiormente propenso ad adottare nel gioco una delle possibili strategie; la strategia vincente si diffonderà maggiormente nella popolazione dei giocatori. Facendo variare un certo numero di parametri (quali la dimensione del gruppo sociale, il valore della risorsa ed il costo del conflitto, che può derivare dall'incontro fra due individui con strategie non compatibili), Kaplan e Hill hanno testato la stabilità del comportamento legato alla condivisione di cibo, riscontrando che questa strategia risulta evolutivamente stabile; il che implica la sua affermazione (in modo statisticamente significativo) sulle altre strategie.

8.2 Antropologia Evolutiva

Studi antropologici (Henrich and McElreath 2003) hanno suggerito che le capacità adattive all'origine dell'enorme successo della nostra specie possano essere fortemente *culturali* nel senso che:

i) sono trasmesse "orizzontalmente" tra gli individui all'interno di una stessa generazione, grazie all'apprendimento sociale (siamo la specie che ne fa sicuramente maggior uso);

ii) vengono accumulate "verticalmente" da una generazione all'altra.

Per comprendere i meccanismi che sottostanno a questa duplice dinamica è necessario analizzare:

i) l'evoluzione dei meccanismi psicologici che permettono l'apprendimento sociale;

ii) le dinamiche evolutive dei sistemi culturali.

A differenza di quanto avviene per gli altri animali, nell'uomo le abilità culturali generano strategie adattive e bagagli di conoscenza che vengono accumulati di generazione in generazione; inoltre, le conoscenze vengono spesso incorporate in regole sociali apprese, in tecniche o euristiche che vengono applicate in modo relativamente automatico, senza la necessità di una reale comprensione di ciò che si sta facendo e di come lo si stia facendo.

Per riuscire a comprendere l'evoluzione dell'apprendimento sociale, alcuni teorici (Boyd and Richerson 1985) hanno proposto alcuni modelli formali, tesi ad analizzare in che modo certi cambiamenti spaziali o temporali dell'ambiente possano influenzare il *trade-off* fra

apprendimento individuale, apprendimento sociale e risposte comportamentali innate. L'intuizione principale che sta dietro a questo tipo di proposta è quella secondo la quale l'apprendimento sociale permette agli organismi di rispondere più rapidamente ai cambiamenti ambientali.

Questo tipo di considerazioni si collegano inoltre ad una serie di studi morfologici, i quali hanno dimostrato che l'aumento delle dimensioni del cervello (in rapporto a quelle del corpo) è correlato alle abilità di apprendimento, sia individuale che sociale. Nei primati, ad esempio, le dimensioni del cervello sono fortemente correlate con le abilità sociali, ma anche con abilità legate all'apprendimento individuale e all'uso di utensili (Reader and Laland 2002). Questi dati sembrano dunque suggerire che l'incremento nelle dimensioni del cervello sia stato in parte guidato dall'incremento delle abilità legate all'apprendimento sociale. Inoltre, sembra esserci stata una forte correlazione fra l'incremento delle dimensioni del cervello e l'aumento della variabilità climatica; un ambiente sempre più imprevedibile e relazioni sociali sempre più complesse devono aver esercitato forti pressioni evolutive sulle dimensioni cerebrali di alcuni animali sociali (in particolare i primati umani e non umani), in modo da permettere facoltà cognitive (sociali) superiori.

In questa prospettiva, le capacità culturali umane possono essere viste come un sottoinsieme ipertrofizzato di una classe più ampia di abilità, legate all'apprendimento, che si sono evolute in diverse specie (Box and Gibson 1999).

Molti autori (Pulliam and Dunford 1980, Boyd and Richerson 1985, Lumsden and Wilson 1981) hanno argomentato che l'apprendimento sociale accresce l'adattività degli esseri umani, in quanto risparmia agli individui i costi dell'apprendimento individuale (sia in termini di tempo che di potenziali errori).

Immaginiamo il caso in cui un individuo debba imparare a riconoscere quale fra due varietà differenti di funghi sia velenosa e quale commestibile: in questo caso, il costo della scelta è molto alto (se non fatale: mangiando i funghi velenosi l'individuo potrebbe intossicarsi o addirittura morire); invece, imparare osservando il comportamento degli altri (e indirettamente le conseguenze che tale comportamento comporta) sembra essere una strategia molto meno rischiosa e maggiormente adattiva.

Tuttavia, Rogers (Rogers 1988) ha mostrato che una ipotesi di questo genere è insufficiente a spiegare il successo adattivo della nostra specie (così fortemente *culturale*).

Utilizzando un modello molto semplice, Rogers ha dimostrato che risparmiare agli individui il costo dell'apprendimento individuale non comporta di per se un incremento generale dell'adattività della popolazione (la *fitness* media della popolazione non cresce). Mentre coloro che sono capaci di apprendere socialmente riescono ad agire molto bene quando sono pochi, questi stessi individui non riescono a fare altrettanto quando sono troppo numerosi; senza la presenza di una capacità di apprendimento individuale, coloro che apprendono socialmente non potrebbero adattarsi ai cambiamenti ambientali; il primo che adotta una forma di apprendimento individuale (ed ha la possibilità di unirsi ad un gruppo omogeneo di individui che apprendono socialmente) avrà sempre una *fitness* maggiore rispetto agli altri. Ciò comporta che allo stato di equilibrio, la *fitness* media della popolazione nel suo complesso sarebbe la stessa di una popolazione composta da soli individui che apprendono individualmente: l'apprendimento sociale da solo non accresce l'adattività.

Abbiamo accennato al fatto che le capacità culturali umane possono essere considerate nell'ambito di un più ampio insieme di caratteristiche adattive, tese a migliorare l'apprendimento in un ambiente che può cambiare velocemente. Queste considerazioni evolutive circa l'enorme adattività ed unicità di queste caratteristiche ci pongono davanti ad un interrogativo interessante: come mai le stesse capacità di apprendimento sociale non si sono evoluto nel corso del tempo anche in altri animali sociali? Mentre esistono numerose evidenze che suggeriscono che altri animali (in particolare gli scimpanzé) presentino comportamenti risultanti da un apprendimento di tipo sociale (McGrew 1992, Wrangham et al. 1994, Whiten et al. 1999, Boesch and Tomasello 1998), esistono tuttavia pochi argomenti per poter credere che capacità di apprendimento sociale negli animali non umani possano generare *adattamento cumulativo* paragonabile a quello di *homo sapiens* (Boyd and Richerson 1985, Tomasello et al. 1993, Tomasello 2000).

Per contro, abilità culturali e conoscenza accumulata sono caratteristiche di tutte le società umane. Se i meccanismi psicologici che rendono possibile la *cultura cumulativa* sono ancora poco chiari, tuttavia esistono alcune promettenti supposizioni. Tomasello et al. (1993) suggeriscono che la *vera imitazione* (o *apprendimento per osservazione*, il quale implica la capacità di copiare accuratamente ed in modo diretto il comportamento osservato, le strategie o la conoscenza simbolica) sia una condizione necessaria per l'evoluzione cumulativa della cultura. Altre forme di apprendimento sociale possono rendere conto delle *tradizioni* ma non

dell'accumulazione di informazioni adattive.

Immaginiamo una situazione in cui alcuni individui siano in grado di utilizzare una modesta quantità di apprendimento individuale, in modo tale che l'interazione con l'ambiente produca lentamente un comportamento adattivo. Immaginiamo poi che altri individui meno dotati approfittino dell'abilità dei primi nel trovare e sfruttare particolari risorse di cibo (per mera imitazione, senza condividere quell'abilità cognitiva); anche questa è una forma molto semplice di apprendimento sociale. Tuttavia, in questo modo ogni nuova generazione di individui dovrà "reinventare" la capacità di trovare e sfruttare una certa risorsa di cibo e questo non permetterà una evoluzione inter-generazionale del comportamento. Se invece gli individui sono in grado di acquisire il comportamento attraverso la diretta osservazione, copiando anche i piccoli dettagli delle tecniche utilizzate dagli altri, allora l'apprendimento individuale potrà usufruire di generazione in generazione delle migliori innovazioni raggiunte di volta in volta.

Probabilmente, la vera imitazione non rappresenta la soluzione completa del problema in esame. Nelle popolazioni umane moderne un ruolo centrale è giocato dalle abilità cognitive che trasformano l'apprendimento sociale in accumulo di cultura. Tomasello (2000) argomenta che la *vera* imitazione, radicata in una capacità genetica evoluta fino al punto di permettere una Teoria della Mente, può generare forme di evoluzione culturale linguistiche e non linguistiche; i sistemi di simboli linguistici (incluse le strutture grammaticali) sarebbero stati gradualmente accumulati, migliorati ed adattati attraverso un processo evolutivo culturale analogo a quello osservato nel dominio della cultura materiale e tecnologica.

Al di là della specifica natura di questi meccanismi, resta ancora da chiarire perché i comportamenti legati all'apprendimento sociale sono così rari.

Boyd e Richerson (1996) hanno proposto un modello molto interessante dell'evoluzione delle capacità culturali cumulative. In questo modello, una popolazione vive in un ambiente variabile nel quale esiste un unico valore adattivo ottimale, per un certo tratto quantitativo; ad ogni generazione esiste una certa probabilità che l'ambiente cambi in maniera tale che venga ad essere modificato anche il valore ottimale del tratto. Il fenotipo degli individui è costituito da una combinazione di influenze genetiche e trasmissioni culturali; altri geni determinano la propensione individuale ad imitare, comportando un costo in termini di *fitness*. Per tutti gli individui è prevista una forma di apprendimento individuale, che sposta il fenotipo della

popolazione verso l'ottimo corrente. Gli individui che presentano una forma di apprendimento culturale possono acquisire i fenotipi maggiormente vicini all'ottimo, una volta che questi esistono nella popolazione; questi fenotipi vengono poi resi ancora migliori dall'apprendimento individuale. Il processo si ripete per la generazione successiva.

A differenza del modello di apprendimento sociale presentato sopra e più semplice (Rogers 1988), in quest'ultimo si mostra che un apprendimento culturale affidabile non si diffonde fin dall'inizio, ma diventa stabile solo una volta che sia raggiunta una certa soglia critica. La selezione naturale favorisce l'apprendimento culturale solo quando i costi dello sviluppo e del mantenimento dei meccanismi di apprendimento culturale sono inferiori ai benefici ottenuti dall'acquisizione di semplici comportamenti appresi individualmente.

Sebbene sia difficile avviare tale efficacia dell'apprendimento culturale, una volta avviato questo è relativamente semplice da mantenere. I meccanismi legati all'apprendimento culturale garantiscono l'accesso alla conoscenza accumulata attraverso le generazioni (cosa che non permette di fare il semplice apprendimento sociale).

Dal punto di vista cognitivo, quando l'informazione comporta un costo, la selezione naturale opera secondo un principio di ottimizzazione: vengono favoriti i meccanismi cognitivi che permettono all'individuo di estrarre informazione, strategie, pratiche, euristiche e credenze adattive (dai membri del proprio gruppo sociale) ad un costo inferiore rispetto a quello che si dovrebbe sostenere affidandosi a meccanismi alternativi individuali.

La cognizione umana probabilmente incorpora numerose euristiche e basi di apprendimento, che facilitano l'acquisizione di conoscenza, pratiche, credenze e comportamenti utili.

8.3 Antropologia Cognitiva ed Approccio Epidemiologico

Fin qui, nell'affrontare il tema dell'evoluzione della mente normativa ci siamo riferiti soprattutto a *marker* etnici e alla condivisione di *artefatti culturali*, senza affrontare esplicitamente il tema delle *norme*.

Da dove provengono le norme che decidiamo di adottare o violare? Perché abbiamo a nostra disposizione un certo *set* di norme piuttosto che un altro?

Un'importante proposta naturalistica, l'*etica evolutiva*, vede anche le norme (morali)

come adattamenti evolutivi (Ruse and Wilson 1986). Seppure questa suggestione susciti un discreto interesse nel panorama filosofico (in particolare in quello relativo alla filosofia della scienza), esiste un diffuso scetticismo riguardo alle spiegazioni evolutive circa il possesso di certe norme morali piuttosto che di altre. Inoltre, anche coloro che sono attratti da una spiegazione evolutiva delle nostre capacità mentali, spesso nutrono perplessità circa le spiegazioni evolutive di specifiche norme (Ayala 1987, 1995; Kitcher 1990, 1994).

Una valida alternativa agli approcci appena menzionati è rappresentata da quell'approccio naturalistico che tenta di spiegare la *genealogia* delle norme senza assumere che le norme morali siano *di per se stesse* adattamenti.

L'antropologia cognitiva contemporanea ha beneficiato molto del promettente approccio *epidemiologico* avanzato per spiegare l'evoluzione della cultura (Sperber 1996, Boyer 2000).

Seguendo questo approccio, è possibile studiare l'evoluzione culturale senza affrontare di fatto l'origine degli oggetti culturali, ma chiedendosi quali siano le cause che permettono a certi oggetti culturali di affermarsi. Questo approccio epidemiologico riconosce alla psicologia umana un ruolo centrale, nella determinazione degli oggetti culturali che sopravviveranno più a lungo.

Un tassello importante della spiegazione epidemiologica è rappresentato delle *emozioni*; queste sono fattori centrali, quando si decide di acquisire, di tendere ad affermare un particolare oggetto culturale.

L'idea è questa: le norme che proibiscono azioni foriere di emozioni negative (come fare del male a qualcuno) sopravviveranno con maggior facilità di quelle che sono neutre da un punto di vista affettivo.

La maggior parte dei tentativi genealogici sulle norme si sono concentrati sulle norme *morali*; si è tentato di spiegare l'origine della moralità in senso generale a partire dalla genesi delle norme morali nel nostro passato culturale. Il problema di queste spiegazioni sta in questo: non è che non ci siano *buone* spiegazioni, ma semmai che ce ne sono *troppe*, con scarse evidenze storiche per stabilire a quale di esse dare maggior peso e credito.

Una rassegna sintetica dei tentativi di spiegare le origini della norma morale e sociale non può fare a meno di citare:

- l'approccio in senso lato riferibile a Nietzsche (e in particolare alla *norma dello schiavo*, secondo la quale il debole ha inventato le norme come strategia di

sopravvivenza, per proteggersi dai torti del più forte, Nietzsche 1887);

- l'*altruismo reciproco*, secondo il quale gli individui convengono di non farsi reciprocamente male, almeno in condizioni di scarsità di risorse (in quanto questo accordo arreca beneficio ad entrambe le parti - Trivers 1971);
- la *reciprocità indiretta*, per la quale chi decide di adottare la norma di non offendere sarà maggiormente appetibile per istituire alleanze (Alexander 1987, Frank 1988);
- la *kin selection*, che giustifica la norma di non arrecare danno a parenti in quanto questo conferisce un vantaggio selettivo (con la possibilità di allargare tale norma al gruppo sociale di appartenenza - Sober and Wilson 1998);
- la *sensibilità emotiva*, secondo cui le emozioni giocano un ruolo centrale nella nascita della norma di non danneggiare altri (essere testimoni della sofferenza altrui turba emotivamente e potrebbe essere alla base della norma in questione);
- la *mutazione casuale*, per cui la norma potrebbe essere una imposizione di un individuo dominante che accidentalmente ha ottenuto grande influenza (tale individuo potrebbe punire chi trasgredisce la norma e chi non punisce i trasgressori - Axelrod 1986; Boyd and Richerson 1992).

Sicuramente, ciascuna delle ipotesi sopra menzionate presenta un suo interesse e plausibilità. Tuttavia, per nessuna di esse esistono evidenze storiche che ne confermino la validità. Inoltre, non possiamo assumere che una singola spiegazione circa l'origine di una particolare cultura possa essere valida per *tutte* le culture. Per esempio, potrebbe benissimo darsi il caso che la norma sociale di “non fare del male ad altri” abbia avuto origine per una determinata ragione in una cultura e per una determinata altra ragione in un'altra cultura. In questo senso, una spiegazione che trovi anche riscontri storici documentati per una certa cultura in un certo periodo storico, non potrà essere assunta a spiegazione universalmente vera.

A questo punto, sembra abbastanza evidente che invece di ricercare le origini di una particolare norma sociale, potrebbe essere più utile cercare di determinare quali caratteristiche una certa norma sociale debba avere, per essere maggiormente adottata e prevalere sulle altre. Più in generale, possiamo chiederci quali caratteristiche debbano avere gli oggetti culturali per sopravvivere ed affermarsi in un un certo gruppo di individui.

Numerose sono state le proposte scientifiche avanzate a questo proposito (Dawkins 1976, Dennett 1995), tra le quali ci piace citare la spiegazione *epidemiologica* formulata da Dan Sperber (1996) e condivisa da altri studiosi (Atran 1998; Boyer 1994, 1999, 2000).

L'approccio epidemiologico si focalizza su una particolare classe di oggetti culturali, le *rappresentazioni mentali*. Se assumiamo che le norme possano essere considerate come rappresentazioni mentali (o meglio che esistano rappresentazioni mentali / credenze, concernenti le norme: le *credenze normative*), possiamo cercare di applicare tale approccio allo studio dell'origine delle norme.

Seguendo Sperber, (il quale tenta di valutare quali oggetti culturali, nella forma di rappresentazioni mentali si affermeranno con maggiore facilità), dobbiamo necessariamente evitare di considerare unicamente i fattori ecologici, ambientali; dobbiamo anche guardare alle peculiarità della *psicologia* umana.

Questa è un'idea centrale dell'approccio epidemiologico: per capire la trasmissione culturale non è sufficiente prendere in considerazione i soli *oggetti culturali*; bisogna conoscere anche la *psicologia* umana, in quanto è necessario capire quali oggetti culturali siano maggiormente rilevanti per creature che hanno una vita mentale molto complessa ed intensa.

Cosa dobbiamo sapere della psicologia umana, per capire quali oggetti culturali avranno una *fitness* maggiore?

Innanzitutto, bisogna discutere quali caratteristiche della psicologia umana possono essere considerate *universali*. In questo senso, Sperber adotta la visione prevalente in psicologia evolutiva, secondo la quale la mente è composta da un certo numero di *moduli* / funzioni, risultato adattivo in risposta all'ambiente che ci circonda. Questi moduli - nella visione prevalente - sono gli *stessi* per un'intera specie; ciascun individuo normale di una stessa specie avrà gli stessi moduli.

Per Sperber, i moduli sono cruciali per la trasmissione culturale, in quanto essi riescono a fissare una quantità di contenuti culturali *in un certo dominio cognitivo*, per il quale essi si sono evolutivamente specializzati.

Per strade simili, anche lo sviluppo del modello epidemiologico ad opera di Pascal Boyer si affida a meccanismi cognitivi universali per la specie, specifici per il dominio. Boyer si concentra su uno specifico *cluster* di corpi dominio-specifici, che egli considera parte

dell'ontologia intuitiva (Boyer 1994, 1999, 2000).

In sintesi, per capire quali particolari oggetti culturali sopravviveranno più facilmente, dovremmo sapere il più possibile circa la *psicologia umana universale*.

In questo quadro, sorge il problema di individuare e controllare sperimentalmente quale oggetto culturale sopravviverà con maggiore probabilità. Sperber e Boyer sostengono che gli oggetti culturali che hanno maggiori probabilità di sopravvivere sono quelli che si ricordano più facilmente.

Volendo allora riassumere i concetti chiave dell'approccio epidemiologico potremmo affermare che:

i) è necessario considerare con cura le caratteristiche universali della psicologia umana;

ii) gli oggetti culturali più facili da ricordare sopravviveranno con maggiore probabilità.

Consideriamo ora le componenti *affettive*. L'idea che meccanismi affettivi determinino parzialmente quali oggetti culturali hanno successo è sicuramente compatibile con l'approccio epidemiologico.

Nella psicologia contemporanea le emozioni rappresentano il campo di battaglia per uno dei più importanti dibattiti circa gli universali umani. Ekman e colleghi hanno prodotto una notevole ed impressionante mole di dati i quali indicano che esistono un certo numero di *emozioni base universali*, le quali condividono determinate caratteristiche. Indipendentemente dalla cultura di appartenenza, esisterebbero elementi comuni alla base delle emozioni fondamentali. Tra le emozioni che rispondono ai criteri distintivi delle emozioni di base ci sono la *tristezza*, la *collera*, la *paura*, il *disgusto* (Ekman 1994).

In questo quadro, le emozioni di base vengono considerate adattamenti evolutivi universalmente istanziati nelle specie, (malgrado possano esistere importanti variazioni culturali nelle condizioni che suscitano queste emozioni e nei modi in cui esse possono essere manifestate - Mallon and Stich 2000). Conoscere il carattere dei sistemi affettivi universali potrebbe aiutarci a determinare quali oggetti culturali avranno successo.

Un dato interessante è rappresentato dal fatto che le emozioni possano facilitare la memoria a lungo termine molto meglio di quanto non riescano a fare per quella a breve termine (Kleinsmith and Kaplan 1963). Questi stessi benefici in termini di memoria a lungo

termine (forniti dalle emozioni) si riscontrano anche nelle norme in quanto oggetti culturali. Le richieste normative che sono rinforzate da componenti emotive (come ad esempio quelle che proibiscono una azione che possa turbare dal punto di vista emotivo) saranno ricordate con maggiore facilità rispetto a quelle che non lo sono.

Così, se una rappresentazione mentale che sia ricordata meglio ha una *fitness* maggiore, le norme salienti dal punto di vista emotivo avranno plausibilmente un vantaggio evolutivo maggiore, in quanto legate a credenze (normative) emotivamente salienti.

Recentemente, in una serie di esperimenti (Nichols 2002) sono state confrontate violazioni normative neutre dal punto di vista emotivo (un ospite che fa rumore mentre mangia la minestra) con altre violazioni cariche di una connotazione emotiva (un ospite che sputa in un bicchiere d'acqua prima di bere dallo stesso bicchiere). I soggetti hanno valutato che le violazioni cariche emotivamente (rinforzate dal disgusto) siano più gravi di quelle neutre. Inoltre, hanno confermato questa stima anche nel caso in cui il padrone di casa avesse detto che per lui l'azione in questione non rappresentava un problema (sputare nel proprio bicchiere prima di bere).

Questi risultati suggeriscono che la componente emotiva apporta un significativo contributo alla salienza delle norme. Le violazioni rinforzate dal disgusto sono trattate in modo più serio e indipendente dall'autorità (il padrone di casa); questo sembra essere parzialmente funzione del livello individuale della sensibilità al disgusto.

Possiamo quindi dire che le norme rinforzate da una componente emotiva sono considerate come più serie ed importanti di quelle neutre.

Le evidenza rispetto alla componente emotiva e alla memoria indicano che le norme che sono rinforzate da una componente emotiva saranno ricordate più facilmente di quelle che non lo sono.

Sembra ragionevole aspettarsi a questo punto che la componente emotiva faciliti anche la trasmissione delle norme e che, nello specifico, norme che proibiscono azioni per suscitare emozioni negative verosimilmente avranno una *fitness* maggiore.

9. CONCLUSIONI

In questo lavoro abbiamo presentato un approccio computazionale allo studio dell'emergenza di artefatti sociali in società artificiali.

Il quadro teorico-sperimentale cui abbiamo fatto riferimento è quello relativo all'utilizzo delle metodologie simulative nell'ambito delle scienze sociali e cognitive.

Da quando sono state inaugurate più di cinquant'anni fa, le scienze cognitive hanno faticato molto a trovare spazio nel panorama delle discipline scientifiche. Questo, da una parte, senza dubbio per il loro carattere spiccatamente interdisciplinare; dall'altra, perché la validità dell'interazione con altre discipline tradizionali (che è forse la qualità migliore delle scienze cognitive) non è stata immediatamente riconosciuta.

D'altro canto, i numerosi approcci simulativi non strettamente legati alle scienze cognitive vantano una tradizione ben consolidata. Facendo uso di strumenti formali tradizionali quali quelli messi a disposizione dalla matematica o dalla logica, la teoria dei giochi e i sistemi multi-agente sono applicati allo studio delle norme da più di trent'anni. Si presentano come due tradizioni contrapposte ed inconciliabili che concentrano la propria attenzione solo su una delle due caratteristiche fondamentali delle norme sociali: la prospettiva strettamente sociale (nel caso della teoria dei giochi) o quella individuale (nel caso dei sistemi multi-agente).

Se gli studiosi interessati ad utilizzare il primo dei due approcci menzionati guardano solo agli effetti (macro) sociali emergenti a livello della popolazione, quelli che adottano il secondo approccio si concentrano sullo studio dei meccanismi decisionali che derivano dall'adozione individuale di una certa norma (dando per acquisita la sua esistenza).

Riconoscendo l'immensa eppure parziale utilità di questi due approcci, negli ultimi quindici anni si è affermata una terza prospettiva che cerca di integrare le due preesistenti; dialogando a stretto contatto con le scienze cognitive, questa *terza via* postula la *necessità delle rappresentazioni mentali* (in particolare credenze e scopi) per studiare la complessa dinamica bidirezionale delle norme (dal micro al macro e dal macro al micro).

Facendo riferimento alla *teoria cognitiva delle norme* che le considera oggetti bidimensionali che esistono contemporaneamente in una dimensione privata ed in una sociale, abbiamo sviluppato una architettura di agente normativo e l'abbiamo testata in una serie di

simulazioni.

Abbiamo confrontato i risultati ottenuti con altri raggiunti utilizzando un tradizionale algoritmo di imitazione sociale ed abbiamo mostrato in che modo l'architettura proposta sia in grado di rendere conto dell'emergenza di norme sociali come effetto della loro immergenza nelle menti degli agenti.

L'idea alla base del modello è che esistano due fasi ben distinte nel processo di emergenza di una nuova norma sociale.

In un primo momento gli agenti devono essere in grado di formarsi una credenza (normativa) circa l'esistenza di una (potenziale) norma: a questo punto nelle loro menti non esiste già una credenza normativa relativa ad una norma certa, ma solo una credenza relativa alla possibilità che esista una norma.

Una volta creata tale rappresentazione mentale, gli agenti devono poter osservare nell'interazione con gli altri e con l'ambiente, un numero sufficientemente elevato di conferme a tale ipotesi. Solo dopo aver ricevuto tali conferme, gli agenti possono formarsi la credenza normativa relativa ad una particolare norma ed agire di conseguenza; questo significa che da questo momento in poi, in base ad un meccanismo dinamico della salienza delle norme (cioè della loro attuale efficacia ed incisività) che indica *hic et nunc* quale norma è più "viva", l'agente inizierà ad agire in modo conforme alla norma (nel caso decida di adottarla) o difforme (nel caso decida di trasgredirla).

Solo una volta completato questo processo di immergenza delle credenze normative nelle menti degli agenti può avere luogo la fase di emergenza sociale della norma.

Dal momento in cui gli agenti posseggono credenze normative ed agiscono in base ad esse, sono in grado i) di comunicare ad altri il contenuto di tali credenze e ii) di cercare di influenzare il comportamento altrui "promuovendo" le norme che ritengono più salienti.

Ovviamente, questo comporta che non necessariamente ogni agente cercherà di diffondere le stesse credenze normative; la dinamica che ne risulta rispecchia fino in fondo la filosofia dell'approccio adottato, quella relativa all'implementazione di un modello di *agenti autonomi eterogenei*. Gli agenti sono *autonomi* perché decidono come agire in base alle proprie rappresentazioni mentali; sono *eterogenei* in quanto il contenuto delle rappresentazioni mentali di uno può essere estremamente diverso dal contenuto della rappresentazioni mentali di un altro.

Scambiandosi messaggi dal contenuto normativo ed osservando comportamenti più o meno conformi ad una norma, gli agenti aggiornano le proprie credenze e si formano nuove credenze normative. A questo punto, se per alcuni di essi una stessa credenza normativa sarà più saliente di altre, potranno dare l'avvio ad un processo di diffusione normativa che porta all'emergenza di una norma sociale: l'intera popolazione, o gran parte di essa, adotterà la stessa norma sociale e diremmo che tale norma *è emersa*.

In questo modo risulta evidente come sia possibile integrare un approccio strettamente sociale (come quello game-teorico) con uno individuale (come quello dei sistemi multi-agente).

Assumendo l'esistenza di due fasi, una in cui le norme possono immergersi nelle menti degli agenti sotto forma di rappresentazioni mentali ed una in cui possono emergere a livello della popolazione grazie alla condivisione delle stesse rappresentazioni mentali, possiamo rendere conte della duplice natura delle norme; la natura privata relativa al singolo individuo e quella sociale.

Chiaramente la prospettiva scientifica proposta è molto differente da quelle esistenti e sarebbe ingenuo non ammettere che presenta dei limiti oltre che dei vantaggi.

Se da una parte siamo in grado di "osservare" cosa accade nelle menti degli agenti prima che una data norma si affermi socialmente, non siamo più in grado di utilizzare i numerosi strumenti formali della matematica o della logica. Questo comporta delle conseguenze nella fase di analisi dei risultati ottenuti dalle simulazioni e rende decisamente più difficile la fase di disseminazione dei risultati nei termini di comunità scientifica con cui poter dialogare (che necessariamente sarà limitata).

In questo lavoro abbiamo anche mostrato (vedi cap. 6) come per lo studio di alcuni particolari artefatti sociali, le convenzioni, sia utile adottare uno dei metodi "tradizionali"; se si desidera confrontare i risultati delle simulazioni con quelli ottenuti attraverso l'uso di modelli analitici, l'approccio game-teorico è estremamente vantaggioso in quanto permette di fare assunzioni confrontabili con quelle di approcci diversi. Nonostante l'uso di una metodologia consueta, abbiamo mostrato come sia possibile ed interessante inserire elementi di novità (in questo caso l'aggiornamento *runtime* dei payoff di ciascuna strategia in un gioco di congestione).

Abbiamo avanzato alcune ipotesi in merito ai possibili meccanismi di interiorizzazione

delle norme ed indicato alcune interessanti linee di ricerca future. Questo per riconoscere che non sempre avviene il complesso processo cognitivo proposto; anzi, nella maggior parte dei casi, nella vita quotidiana, probabilmente noi agiamo in maniera automatizzata, non ragionata, proprio perché le norme sociali possono essere interiorizzate in diversi modi e misure.

Infine, ci è sembrato utile presentare una panoramica il più possibile acritica delle spiegazioni esistenti sull'origine evolutiva delle norme in quanto oggetti culturali.

Abbiamo passato in rassegna tre possibili *classi* di spiegazioni mostrando come anche in questo caso possa essere utile realizzare dei modelli computazionali per testare le proprie assunzioni teoriche, se non addirittura per proporne di nuove in base ai risultati di esperimenti effettuati *in-silico*.

Abbiamo individuato tre possibili approcci per spiegare la genealogia delle norme.

Il primo è rappresentato da una evoluzione della teoria dei giochi, l'*evolutionary game theory* (EGT) che differisce dalla teoria dei giochi classica in quanto assume che gli agenti non siano in possesso di informazioni complete e perfette e competano per la disponibilità di risorse limitate adottando strategie differenti.

Il secondo, quello legato alla tradizione dell'*antropologia evolutiva*, tenta di rintracciare le origini evolutive delle norme sociali nella capacità dei primati (umani e non-umani) di trasmettere oggetti culturali in modo intra-generazionale ed inter-generazionale, mettendo in luce i vantaggi adattivi di tale spiegazione.

Infine, il terzo approccio è rappresentato dall'*antropologia cognitiva*, con particolare riferimento alla teoria epidemiologica e al ruolo giocato dalle emozioni nella trasmissione di artefatti sociali e culturali.

10. BIBLIOGRAFIA

- Abdellaoui, M. Bleichrodt, H., Paraschiv, C. 2007. "Loss Aversion Under Prospect Theory: A Parameter-Free Measurement", *Management Science*, Vol. 53, Issue 10, pp. 1659–1674.
- Alchourrón, C. 1993. "Philosophical Foundations of Deontic Logic and the Logic of Defeasible Conditionals", in J-J. Ch. Meyer and J. R. Wieringa (Eds.), *Deontic Logic in Computer Science: Normative System Specification*, Chichester: John Wiley & Sons, pp. 43-84.
- Alexander, S. 1920. *Space, Time, and Deity*. 2 vols. London: Macmillan.
- Alexander, R. 1987. The Biology of Moral Systems. New York: Aldine.
- Andersen, P. B., Emmeche, C., Finnemann, N. O, Christiansen, P. V. (Eds.) 2000. *Downward Causation. Minds, Bodies and Matter*, Århus: Aarhus University Press.
- Andrighetto, G., Campennì, M., Conte, R., and Paolucci, M. 2007. "On the Immergence of Norms: a Normative Agent Architecture", in *Proceedings of AAAI Symposium, Social and Organizational Aspects of Intelligence*, Washington DC. Forthcoming.
- Andrighetto, G., Giardini, F., and Conte, R. 2009. "Norms through minds", in *Proceedings of WOW04*, Indiana University, Bloomington, 3-6 Giugno, 2009.
- Andrighetto, G., Tummolini, L., Castelfranchi, C., Conte, R. (forthcoming). "A convention or (tacit) agreement betwixt us". In Johan van Benthem, Vincent F. Hendricks, John Symons and Stig Andur Pedersen (Eds.) *Between Logic and Intuition: David Lewis and the Future of Formal Methods*, in Philosophy Synthese Library book series as Dordrecht: Springer.
- Andrighetto, G., Campennì, M., Cecconi, F., and Conte, R. (forthcoming). "The Complex Loop of Norm Emergence: a Simulation Model", in K. Takadama, C. C. Revilla, G. Deffuant (Eds.) *The Second World Congress on Social Simulation*, Springer- Verlag LNAI. Aspects of Intelligence Washington DC, 2008.
- Arthur, W. B. 1994. "Inductive Reasoning and Bounded Rationality", *American Economic Review*, 84,406–411.
- Atran, S. 1998. "Folk Biology and and the Anthropology of Science: Cognitive

Universals and Cultural Particulars", Behavioral and Brain Sciences, 21:547-609.

- Axelrod, R. 1984. *The Evolution of Cooperation*, New York, Basic Books.
- Axelrod, R. 1986. "An Evolutionary Approach to Norms", American Political Science Review, 80:1095-1111.
- Axelrod, R. 1987. *The Evolution of Strategies in the Iterated Prisoner's Dilemma*. Los Altos, CA: Kaufmann.
- Axelrod, R. 1995. "A model of the Emergence of New Political Actors". In N. Gilbert and R. Conte (Eds.) *Artificial Societies: the Computer Simulation for Social Life*. London: UCL Press.
- Ayala, F. 1987. "The Biological Roots of Morality", *Biology and Philosophy*, 2:235-252.
- Ayala, F. 1995. "The Difference of Being Human: Ethical Behavior as an Evolutionary Byproduct", in H. Rolston (Eds.) Biology, Ethics, and the Origins of Life. Boston: Jones & Bartlett, 113-135.
- Bandura, A. 1991. "Social cognitive theory of self-regulation", *Organizational Behavior and Human Decision Processes*, 50, 248-287.
- Bargh, J. A., Gollwitzer, P. M., Lee-Chai, A., Barndollar, K., and Troetschel, R. 2001. "The automated will: Unconscious activation and pursuit of behavioral goals", *Journal of Personality and Social Psychology*. 81:1004–1027.
- Barsalou, L. W. 1999. "Perceptual symbol systems", *Behavioral and Brain Sciences*, 22: 577-660.
- Barth, F. 1969. "Introduction" in Ethnic groups and boundaries. Boston: Little Brown & Co.
- Basu, K. 1998. "Social Norms and the Law", in Peter Newman (Eds.), *The New Palgrave Dictionary of Economics and Law*. London: Macmillan.
- Bicchieri, C. 1990. "Norms of cooperation", *Ethics*,100: 838-861.
- Bicchieri, C. 2006. *The Grammar of Society: The Nature and Dynamics of Social Norms*. New York: Cambridge University Press.
- Bickel, W. K., and Johnson, M. W. 2003. "Delay discounting: A fundamental behavioral process of drug dependence". In G. Loewenstein, D. Read, and R. F. Baumeister (Eds.), *Time and Decision*. New York: Russell Sage Foundation.
- Binmore, K. 1994. Game theory and the social contract. Cambridge, MA: MIT Press.

- Boella, G., Van der Torre, L., and Verhagen, H. 2006. “Introduction to normative multiagent systems”, *Journal of Computational and Mathematical Organization Theory*, 12(2/3):71–80.
- Boesch, C., and Tomasello, M. 1998. “Chimpanzee and human culture”. Current Anthropology, 39:591-604.
- Bonabeau, E. 2002. “Agent-based modeling: Methods and techniques for simulating human systems”, *PNAS*, vol. 99, pp. 7280-7287.
- Box, H. O., and Gibson, K. R. (Eds.) 1999. *Mammalian social learning: comparative and ecological perspectives*. Cambridge, MA: Cambridge University Press.
- Boyd, R., and Richerson, P. J. 1985. *Culture and the evolutionary process*. Chicago: University of Chicago Press.
- Boyd, R., and Richerson, P. J. 1987. “Evolution of ethnic markers”. *Cultural Anthropology*, 2:65-79.
- Boyd, R., and Richerson, P. J. 1992. “Punishment allows the evolution of cooperation (or anything else) in sizeable groups”, *Ethology and Sociobiology*, 13:171-195.
- Boyd, R., and Richerson, P. J. 1996. “Why culture is common, but cultural evolution is rare”, in Runciman, W. G., Maynard Smith, J., Dunbar, R. I. M. (Eds.). *Evolution of social behaviour patterns in primates and man*. Proceedings of The British Academy, Vol. 88. Oxford: Oxford University Press. Pp 77-93.
- Boyer, P. 1994. “Cognitive Constraints on Cultural Representations: Natural Ontologies and Religious Ideas”, in L. Hirshfeld and S. Gelman (Eds.), *Mapping the Mind*. Cambridge, MA: Cambridge University Press, 391-411.
- Boyer, P. 1999. “Cognitive Tracks of Cultural Inheritance: How Evolved Intuitive Ontology Governs Cultural Transmission”, *American Anthropologist*, 100:876-889.
- Boyer, P. 2000. “Evolution of the Motern Mind and the Origins of Culture: Religious Concepts as a Limiting Case”, in P. Carruthers and A. Chamberlain (Eds.) Evolution and the Human Mind. Cambridge, MA: Cambridge University Press, 93-113.
- Bowles, S. e Gintis, H. 2001. “The evolution of strong reciprocity”, (http://www.umass.edu/preferen/gintis/evolsr.pdf).
- Broad, C. D. 1925. *The Mind and Its Place in Nature*. London: Routledge & Kegan Paul.
- Broersen, J., Dastani, M., Hulstijn, J., Huang, Z., and Van der Torre, L. 2001. “The

boid architecture. Conflicts between beliefs, obligations, intentions and desires", in *Proceedings of the fifth international conference on Autonomous Agents*, Montreal, Quebec, Canada, pages 9 – 16.

- Campbell, D. T. 1974. "Downward causation' in Hierarchically Organized Biological Systems", in *Studies in the Philosophy of Biology*, F.J. Ayala & T. Dobzhansky (Eds.), Macmillan Press, p. 179-186.
- Campennì, M., Andrighetto, G., Cecconi, F., and Conte, R. 2008. "Normal = Normative? The Role of Intelligent Agents in Norm Innovation". In *The Fifth Conference of the European Social Simulation Association (ESSA)*. University of Brescia, September 1-5, 2008.
- Campennì, M., Andrighetto, G., Cecconi, F., and Conte, R. 2009. "Normal = Normative? The role of intelligent agents in norm innovation", *Mind & Society*, Vol. 8, n. 2, 2009.
- Castelfranchi, C. 1998. "Simulating with Cognitive Agents: The Importance of Cognitive Emergence Multi-Agent Systems and Agent-Based Simulation", *Lecture Notes in Computer Science (LNCS)*. Springer Berlin / Heidelberg. ISSN - Volume 1534/1998 , p. 26-44.
- Castelfranchi, C. 1999. "Prescribed mental attitudes in goal-adoption and norm adoption", *Artificial Intelligence and Law*, 7(1):37–50.
- Cecconi, F., Zappacosta S., Marocco, D., and Acerbi, A. 2007. "Social and individual learning in a microeconomic framework", in *Proceedings of the Econophysics Colloquium And Beyond*, September 27-29, Ancona, Italia.
- Cecconi, F., Zappacosta, S. 2008. "Low correlations between dividends and returns: the alitalia's case", in *Proceedings of the IASTED International Conference on Modelling and Simulation (MS 2008)*, February, 13 - 16, Quebec, Canada.
- Challet, D., Zhang, Y.-C. 1997. "Emergence of cooperation and organization in an evolutionary game", *Physica A*, 246, 407-418.
- Challet, D., Zhang, Y.-C. 1998. "On the minority game: Analytical and numerical studies", *Physica A*, 256, 514-532.
- Chmura, T., Pitz, T. 2006. "Successful Strategies in Repeated Minority Games", *Physica A*, 363, 477-480.
- Chmura, T., Pitz, T. 2007. "An Extended Reinforcement Algorithm for Estimation of

Human Behaviour in Experimental Congestion Games", *Journal of Artificial Societies and Social Simulation*, 10 (2).

- Cohen, P. R. and Levesque, H. J. 1990. "Persistence, Intention, and Commitment", in Cohen, P. R., Morgan, J., Pollack, M. A. (Eds), *Intentions in Communication*, pp. 33-71. Cambridge, MA: MIT Press.
- Cohen, P. R. and Levesque, H. J. 1990a. "Intention is choice with commitment". *Artificial Intelligence*, 42(2-3):213-261.
- Cole, M. 1997. "Culture and Cognitive Science", talk presentato il 15 maggio al *Cognitive Science Program*, U.C., Santa Barbara.
- Coleman, J. S. 1990. *Foundations of Social Theory*. Cambridge, MA: Harvard University Press.
- Conte, R. and Castelfranchi, C. 1995. *Cognitive and social action*, London: London University College of London Press.
- Conte, R. and Castelfranchi, C. 1995a. "Norms as Mental Objects. From Normative beliefs to Normative goals", in J.P. Mueller e C. Castelfranchi (Eds.), *From reaction to cognition*, Springer.
- Conte, R. 1998. *L'obbedienza intelligente*. Bari: Laterza.
- Conte, R., Castelfranchi, C., and Dignum, F. 1998. "Autonomous norm-acceptance", in *Proceedings of the 5th International Workshop on Intelligent Agents V, Agent Theories, Architectures, and Languages*, pp. 99–112.
- Conte, R. and Castelfranchi, C. 1999. "From conventions to prescriptions. Towards a unified theory of norms", *Artificial Intelligence and Law*, 7: 323-340.
- Conte, R. 2000. "Memes Through (Social) Minds". In R. Aunger (Eds.), *Darwinizing culture: The status of memetics as science*. London: Oxford University Press.
- Conte R. and Dignum F. 2001. "From Social Monitoring to Normative Influence", *Journal of Artificial Societies and Social Simulation*, 4 (2).
- Conte, R. and Paolucci, M. 2001. "Intelligent Social Learning", *Journal of Artificial Societies and Social Simulation*, 4 (1).
- Conte, R. 2002. "Agent-based modeling for understanding social intelligence", *PNAS*, vol. 99, pp. 7189-7190.
- Conte, R., and Castelfranchi, C. 2006. "The mental path of norms", *Ratio Juris*, 19(4):501–517.

- Conte, R., Andrighetto, G., Campennì, M, Paolucci, M. 2007. "Emergent and Immergent Effects in Complex Social Systems", in *Proceedings of AAAI Symposium, Social and Organizational Aspects of Intelligence*, 8-11 November 2007 Washington DC.
- Conte, R. 2009. "Rational, Goal-Oriented Agents". *Encyclopedia of Complexity and Systems Science*, 7533-7548, Springer.
- Cosmides, L. e Tooby, J. 1992. "Cognitive adaptations for social exchange". In J. Barkow, L. Cosmides, & J. Tooby (Eds.). *The adapted mind*, New York: Oxford University Press.
- Cosmides, L. e Tooby, J. 2008. "Can evolutionary psychology assist logicians? A reply to Mallon". In W. Sinnott-Armstrong (Ed.), *Moral psychology*. (pp. 131-136) Cambridge, MA: MIT Press.
- Cummins, D. D. 1996. "Evidence for deontic reasoning in 3- and 4-year olds", *Memory and Cognition*, 24 (6): pp. 823-829.
- Danielson, P. 1992. *Artificial Morality: Virtuous Robots for Virtual Games*. New York: Routledge.
- Danielson, P. 1998. (Eds.) *Modeling Rationality, Morality and Evolution*. Oxford University Press.
- Dawkins, R. 1976. The Selfish Gene. New York: Oxford University Press.
- Dennett, D. 1995. *Cognition, Computation and Consciousness*, New York: Oxford University Press.
- Dennett, D. 1995. *Darwin's Dangerous Idea: Evolution and the Meanings of Life*. New York: Simon and Schuster.
- Dignum, F. 1999. "Autonomous agents with norms", *Artificial Intelligence and Law*, 7(1):69–79.
- Douglas, M. 1986. *How Institutions Think*. Syracuse, New York: Syracuse University Press.
- Dow, J. W. 1997. "How Cultural Anthropology Contributes to Culture: the Scientific Method in Late Twentieth Century Cultural Anthropology", paper presentato al *74th Annual Meeting of Central States Anthropological Society*, Wisconsin.
- Dunbar, R. 1997. *Grooming, Gossip, and the Evolution of Language*, Harvard: Harvard University Press.

- Durkheim, E. 1951. *Suicide*, New York: The Free Press.
- Eduardo, A. C., and Eugenio, B. 1971. *Normative Systems*. Wien, Springer-Verlag.
- Eduardo, A. C. 1985. "On the logic of theory change: partial meet contraction and revision functions", *Journal of Symbolic Logic*, (50):510–530.
- Ekman, P. 1994. "All Emotions are Basic", in P. Ekman and R. Davidson (Eds.) The Nature of Emotions. New York: Oxford University Press, pp. 15-19.
- Ekman, J. 2001. "Game theory evolving. a problem-centered introduction to modelling strategic interaction", *Ecological Economics*, 39(3):479–480.
- Elster, J. 1987. *Rationality and social norms*, manoscritto, University of Chicago.
- Epstein, J. M. e Axtell, R. L. 1996. *Growing Artificial Societies: Social Science from the bottom up*. Cambridge, MA: MIT Press.
- Epstein, J. M. 2006. *Generative Social Science. Studies in Agent-Based Computational Modeling*. Princeton-New York: Princeton University Press.
- Fiske, S. T. and Taylor, S. E. 1991. *Social cognition* (2nd edn.). New York: McGraw Hill.
- Frank, R. 1988. Passions within Reason. New York: W. H. Norton.
- Gessler, N. 2002. "Computer Models of Cultural Evolution", in *Evolution in the computer age*. Sudbury, Massachusetts: Jones and Bartlett Publishers.
- Gilbert, M. 1981. "Game theory and convention", *Synthese*, 46:41–93.
- Gilbert, M. 1989. *On Social Facts*. London, New York: Routledge.
- Gilbert, N. e Doran, J. 1994. (Eds.) *Simulating Societies: the computer simulation of social processes*. London: UCL Press.
- Gilbert, N. e Conte, R. 1995. (Eds.) *Artificial Societies: the computer simulation of social life*. London: UCL Press.
- Gilbert, N. 2002. "Varieties of emergence". Paper presented at the *Agent 2002. Conference: Social agents: ecology, exchange, and evolution*, Chicago.
- Gintis, H. 2003. "Solving the Puzzle of Prosociality", *Rationality and Society*, 15, 2, 155-187.
- Gintis, H. 2004. "The genetic side of gene-culture coevolution: internalization of norms and prosocial emotions", *Journal of Economic Behavior & Organization*, 53(1):57-67.

- Goodall, C. E. 2005. “Modifying smoking behavior through public service announcements and cigarette package warning labels: A comparison of Canada and the United States”. Senior Honors Thesis, Ohio State University, June 2005.
- Grafen, A. 1990. “Do animals really recognize kin?”, Animal Behaviour, 39:42-54.
- Hart, H. L. A. 1961. *The concept of law*. New York: Oxford University Press.
- Hart, H.L.A. “Prolegomenon to the Principles of Punishment.” In *Punishment and Responsibility: Essays in the Philosophy of Law*. New York: Oxford University Press, 1968. pp. 1-27.
- Hempel, C., and Oppenheim, P. 1948. “Studies in the Logic of Explanation”. *Philosophy of Science*, 15: 135-175.
- Henrik, G. 1963. *Norm and Action. A Logical Inquiry*. London: Routledge and Kegan Paul.
- Henrich, J., and McElreath, R. 2003. “The Evolution of Cultural Evolution”, Evolutionary Anthropology, 12:123-135.
- Horne, C. 2007. “Explaining norm enforcement”, *Rationality and Society*, 19(2):139-170.
- Horty, J., F. 2001. *Agency and Deontic Logic*. Oxford: Oxford University Press.
- Hutchins, E. 1995. *Cognition in the wild*. Cambridge, MA: MIT Press.
- Jones, A. J. and Porn, I. 1985. “Ideality, subideality and deontic logic”, *Synthese*, pp. 275–290.
- Jones, A. J. and Sergot, M. J. 1993. “On the Characterisation of Law and Computer Systems: The Normative Systems Perspective”. In J.-J. Ch. Meyer & J. R. Wieringa (Eds.) *Deontic Logic in Computer Science: Normative Systems Specification*. Chichester: John Wiley & Sons.
- Jones, A. J., and Sergot, M. 1996. “A formal characterization of institutionalized power”, *Logic Journal of the IGPL*, 4(3):429–445.
- Kahneman, D., Tversky, A. 1979. “Prospect Theory: An Analysis of Decision under Risk”, *Econometrica*, Vol. 47, No. 2, pp. 263-291.
- Kameda, T., Takezava, M., and Hastie, R. 2005. “Where Do Social Norms Come From?”, *American Psychological Society*, 14:6, 331-334.
- Kaplan, H., and Hill, K. 1985. “Food sharing among Ache foragers: Tests of

explanatory hypotheses", Current Anthropology, 26, 223-246.

- Kelsen, H. 1991. *General Theory of Norms*. Oxford: Oxford University Press. (1st ed. 1979).
- King, K. 2008. "The politics of partnerships: Peril or promise", Special Issue of *NORRAG NEWS: Network for Policy Research Review and Advice on Education and Training (NORRAG)*, 41, Dec. 2008.
- Kitcher, P. 1990. "Developmental Decomposition and the Future of Human Behavioral Ecology", *Philosophy of Science*, 57:96-117.
- Kitcher, P. 1994. "Four Ways of 'Biologicizing' Ethics", in E. Sober (Eds.) *Conceptual Issues in Evolutionary Biology*. Cambridge, MA: MIT Press.
- Kleinsmith, L., and Kaplan, S. 1963. "Paired-associate Learning as a function of Arousal and Interpolated Interval", Journal of Experimental Psychology, 65:190-193.
- Kohlberg, L. 1971. "From Is to Ought: How to commit to naturalistic fallacy and get away with it in the study of moral development", in T. Mischel (Eds.), *Cognitive development and epistemology*. New York: Academic Press.
- Kohlberg, L. 1981. "Justice and reversibility". In L. Kohlberg (Eds.) *Essays on Moral Development*, vol. 1. San Francisco: Harper and Row.
- Kohlberg, L. and Turiel, E. 1971. "Moral development and moral education". In G. Lesser (Eds.) *Psychology and educational practice*. New York: Scott Foresman.
- Latane, B., and Darley, J. 1970. *The unresponsive bystander: Why doesn't he help?* New York: Appleton- Century-Crofts.
- Levine, J.M., Valle, R. 1975. "The convert as a credible communicator", *Social Behavior and Personality*, 3 (1), 81-90.
- Lewis, D. K. 1969. *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.
- Lindahl, L. 1977. *Position and Change*. Dordrecht: Reidel Publishing Company.
- Lumsden, C., and Wilson, E. O. 1981. Genes, mind and culture. Cambridge, MA: Harvard University Press.
- Macy, M. W., and J. Skvoretz. 1998. "The evolution of trust and cooperation between strangers", *American Sociological Review*, 63, 638-660.
- Macy, M. W., and Sato, Y.. 2002. "Trust, Cooperation, and Market Formation in the U.S. and Japan.", *PNAS*, 99: 7214-20.

- Mallon, R., and Stich, S. 2000. "The Odd Couple: The Compatibility of Social Construction and Evolutionary Psychology", Philosophy of Science, 67.
- Markus, H. and Zajonc, R. B. 1985. "The cognitive perspective in social psychology". In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology*, pp. 137-229, 3rd Edition. New York: Random House.
- McAdams, R. H. 2008. "Norm Internalization: A Comment on Philip Pettit, Norms, Commitment and Censure", Draft of 3 December 2008.
- McElreath, R., Boyd, R., and Richerson, P. J. 2003. "Shared Norms Can Lead to the Evolution of Ethnic Markers", *Current Anthropology,* 44: 122–130.
- McGrew, W. C. 1992. Chimpanzee material culture: implications for human evolution. Cambridge, MA: Cambridge University Press.
- Mead, M. 1963. *Cultural patterns and technical change*. New York: The New American Library.
- Meyer, J.-J. 1988. "A Different Approach to Deontic Logic: Deontic Logic Viewed as a Variant of Dynamic Logic", *Notre Dame Journal of Formal logic*, 29 (1): 109-136.
- Milchtaich, I. 1996. "Congestion Games with Player-Specific Payoff Function", Games and Economic Behavior, 13, 111–124.
- Miller, G.A., Galanter, E., and Pribram, K.H. 1960. *Plans and the Structure of Behavior*. New York: Holt, Rinehart & Winston.
- Myerson, R. B. 1991. *Game Theory: Analysis of Conflict*. Harvard: Harvard University Press.
- Nash, J. F. 1950. "Equilibrium points in N-Person Games", *PNAS*, Vol. 36 No.1 pp. 48-49.
- Nettle, D. 1996. "Language diversity in West Africa: An ecological approach", *Journal of Anthropology Archaeology*, 15:403-408.
- Nettle, D., and Dunbar, R. I. M. 1997. "Social markers and the evolution of reciprocal exchange", Current Anthropology, 38:93-99.
- Nichols, S. 2002. "Norms with Feeling: Towards a Psychological Account of Moral Judgment", *Cognition,* 84, 221-236.
- Nickerson, R. S. 1998. "Confirmation bias: A ubiquitous phenomenon in many guises". *Review of General Psychology*, Vol. 2, pp. 175-220.
- Nietzsche, F. 1887. On the Genealogy of Morals. Translation of W. Kaufman and R.

Hollingdale. New York: Vintage Books.

- Norman, D. 1994. "Twelve issues for Cognitive Science", *Cognitive Science*, 4, 1-32.
- Nucci, L. P. 2001. *Education in the Moral Domain*. Cambridge, UK: Cambridge University Press.
- Ostrom, E. 1998. "A Behavioral Approach to the Rational-Choice Theory of Collective Action", *American Political Science Review*, 92, 1, 1-22.
- Parsons, T. 1967. *Sociological Theory and Modern Society*. New York: Free Press.
- Piaget, J. 1965. *The moral judgment of the child*. New York: The Free Press.
- Piaget, J. 1972. *Il giudizio morale nel fanciullo*. Firenze: Giunti Barbera.
- Piaget, J. 1978. *Success and Understanding*, London: Routledge and Kegan Paul.
- Plotkin, H. 1995. *Darwin Machines and The Nature of Knowledge*, London: Penguin.
- Posner, R., and Rasmusen, E. B. 1999. "Creating and enforcing norms, with special reference to sanctions", *International Review of Law and Economics*, pp. 369–382.
- Pulliam, H. R., and Dunford, C. 1980. Programmed to learn: an essay on the evolution of culture. New York: Columbia University Press.
- Rachlin, H. 2000. *The Science of Self-Control*. Cambridge, UK: Harvard University Press.
- Rao, A. S., and Georgeff, M. P. 1992. "Social plans: Preliminary report". In Werner, E. and Demazeau, Y., (Eds.) Decentralized AI 3 - *Proceedings of the Third European Workshop on Modelling Autonomous Agents and Multi-Agent Worlds (MAAMAW-91)*, pp. 57-76. Amsterdam: Elsevier Science Publishers.
- Raz, J. 1975. *Practical reason and norms*. Oxford: Oxford University.
- Reader S. M., and Laland, K. N. 2002. "Social intelligence, innovation, and enhanced brain size in primates". *PNAS*, 99:4436-4441.
- Rogers, A. R. 1988. "Does biology constrain culture?", American Anthropology, 90:819-831.
- Rosenthal, R.W. 1973. "A Class of Games Possessing Pure-Strategy Nash Equilibria", *International Journal of Game Theory*, 2, 65–67.
- Ruse, M., and Wilson, E. 1986. "Moral Philosophy as Applied Science", Philosophy, 61:173-192.
- Ryan, R. M., Deci, E. L. 2000. "Self-Determination Theory and the Facilitation of

Intrinsic Motivation, Social Development, and Well-Being", *American Psychologist*, 55(1): 68.

- Sandholm, W.H., Dokumaci, E., Lahkar, R. 2008. "The projection dynamic and the replicator dynamic", *Games and Economic Behavior*, 64-2:666–683.
- Schank, R. C. and Abelson, R. P. 1977. *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Schelling, T. C. 1960. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.
- Scott, J. F. 1971. *Internalization of Norms: A Sociological Theory of Moral Commitment*. Englewoods Cliffs, N.J.: Prentice-Hall.
- Searle, J. 1990. "Collective Intentions and Actions". In Cohen, P., Morgan, J., and Pollack, M. (Eds.) *Intentions in Communication*, 401-415, Cambridge, MA: MIT Press.
- Segerberg, K., Meyer, J. J., and Kracht, M. 2009. "The Logic of Action", *Stanford Enciclopedia of Philsophy.*
- Sen, S., and Airiau, S. 2007. "Emergence of norms through social learning". In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence, (IJCAI'07)*, Hyderabad, India.
- Shoham, Y., and Tennenholtz, M. 1992. "On the synthesis of useful social laws in artificial societies". In *Proceedings of the 10th National Conference on Artificial Intelligence*, 276-282. San Mateo, California: Kaufmann.
- Shoham, Y., and Tennenholtz, M. 1994. "Co-learning and the evolution of social activity". *Technical Report CS- TR-94-1511*, Stanford University.
- Shweder, R., Mahapatra, M., and Miller, J. 1987. "Culture and moral development". In J. Kagan & S. Lamb (Eds.), *The Emergence of Morality in Young Children*. Chicago: Chicago University Press.
- Sichman, J. S., Conte, R., Castelfranchi, C., and Demazeau, Y. 1994. "A Social Reasoning Mechanism Based On Dependence Networks". In A.G. Cohn (Eds.) *Proceedings of the 11th European Conference on Artificial Iintelligence, ECAI*, 188-192. Baffin Lane, England: John Wiley and Sons.
- Sichman, J. S., and Conte, R. 2002. "Multi-agent dependence by dependence graphs".

In *Proceedings of Autonomous Agent & MAS, AAMAS 2002*, 483-91. ACM Press, Part I .

- Simon, H. 1981. *Sciences of the artificial*. Cambridge, MA: MIT Press.
- Smith, J.M. 1974. "The theory of games and the evolution of animal conflicts", *Journal of Theoretical Biology*, 47:209–21.
- Sober, E., and Wilson, D. 1998. *Unto Others*. Cambridge, MA: Harvard University Press.
- Sperber, D. 1996. Explaining Culture. Cambridge, MA: Blackwell.
- Sripada, C. and Stich, S. 2006. "A Framework for the Psychology of Norms". In P. Carruthers, S. Laurence and S. Stich (Eds.) *The Innate Mind: Culture and Cognition*, 280-301, Neew York: Oxford University Press.
- Sternberg, R. J. 2007. "Critical Thinking in Psychology: It really is critical", in Robert J. Sternberg, Henry L. Roediger, Diane F. Halpern (Eds.) *Critical Thinking in Psychology*. Cambridge, MA: Cambridge University Press.
- Sugden, R. 1986/2004. *The Economics of Rights, Co-operation, and Welfare*, 2nd ed. New York: Palgrave Macmillan.
- Sugden, R. 1998. "The Role of Inductive Reasoning in the Evolution of Conventions", *Law and Philosophy*, 17: 377–410.
- Sun, Y ., and Wu, B. 2006. "Agent Hybrid Architecture and Its Decision Processes", in *Proceedings of International Conference on Machine Learning and Cybernetics*, 13-16 Aug. 2006, 641 – 644.
- Tomasello, M., Kruger, A. C., and Ratner H. H. 1993. "Cultural learning", *Behavioral and Brain Sciences*, 16:495-552.
- Tomasello, M. 2000. *The cultural origins of human cognition*. Cambridge, MA: Harvard University Press.
- Trivers, R. 1971. "The Evolution of Reciprocal Altruism", Quarterly Review of Biology, 46:35-57.
- Troitzsch, K. G. 2008. "Simulating Collaborative Writing: Software Agents Produce a Wikipedia", in *Proceedings of The Fifth Conference of the European Social Simulation Association (ESSA)*. University of Brescia, September 1-5, 2008.
- Turner, M. 2001. *Cognitive Dimensions of Social Science*, New York: Oxford University Press.

- Ullman-Margalit, E. 1977. *The Emergence of Norms*. Oxford: Clarendon Press.
- Van der Torre, L., and Tan, Y. H. 1999. "Contrary to duty reasoning with preference-based dyadic obligations". *Annals of Mathematics and Artificial Intelligence*, pp. 1239–1246.
- Van Dyke Parunak, H., Savit, R., and Riolo, R. 1998. "Agent-Based Modeling vs Equation-Based Modeling: A Case Study and Users' Guide", in Sichman, J., S., Conte, R., Gilbert, N. (Eds.) *Multi-Agent Systems and Agent-Based Simulation*, Lecture Notes in Artificial Intelligence, pp. 10-25.
- von Wright, G. H. 1963. *Norm and Action. A Logical Inquiry*. London: Routledge and Kegan Paul.
- Vygotsky, L. S. 1978. Mind in society: The development of higher psychological processes. Cambridge, MA: Harvard University Press. Published originally in Russian in 1930.
- Young, H. P. 1993. "The evolution of conventions", *Econometrica*, 61: 57-84.
- Young, H. P. 1998. *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. Princeton: Princeton University Press.
- Young, H. P. 2006. "Social Norms", Prepared for the *New Palgrave Dictionary of Economics* Second Edition, Steven N. Durlauf and Lawrence E. Blume (Eds.), London: Macmillan.
- Walker, A., and Wooldridge, M. 1995. "Understanding the emergence of conventions in multi-agent systems". In M. Press (Eds.) *Proceedings of the First International Conference on Multi-Agent Systems*, number 384–389.
- Wason, P. and Johnson-Laird, P. 1972. *Psychology of Reasoning: Structure and Content*. Cambridge, MA: Harvard University Press.
- Whiten, A., Goodall, J., McGrew, W. C., Nishida, T., Reynolds, V., Sugiyama, Y., Tutin, C. E. G., Wrangham, R. W., and Boesch, C. 1999. "Cultures in chimpanzees", *Nature*, 399:682-685.
- Wrangham, R. W., McGrew, W. C., de Waal, F. B. M., and Heltne, P. 1994. Chimpanzee cultures. Cambridge, MA: Harvard University Press.
- Wright, R. 1994. *The Moral Animal. Why we are the way we are*. London: Abacus.