i-lex Scienze Giuridiche, Scienze Cognitive e Intelligenza Artificiale
Rivista quadrimestrale on-line: www.i-lex.it
Settembre 2004, numero 1

MORALE ARTIFICIALE E DIRITTO ARTIFICIALE

LOTHAR PHILIPPS[*]

Traduzione italiana dall'inglese: ***F.ROMEO, G.GIULIANI, L.FIORI.***

1. Introduzione.

Intelligenza Artificiale e diritto: questo accostamento suggerisce una '*liaison*': Artificial Law [Philipps 1989]. La disciplina della Vita Artificiale esiste già [Levy 1992]: forme di vita artificiale, simulate al computer, si adattano ad un ambiente artificiale, combattono per la vita, si diversificano, combinano le loro risorse naturali ed evolvono. Non potrebbe valere lo stesso per la morale ed il diritto? regole di comportamento guiderebbero persone immaginarie nel lottare, cooperare, diversificarsi ed unirsi.

Il filosofo canadese Peter Danielson ha recentemente pubblicato un libro nel quale si esplorano tali possibilità avvalendosi di programmi per elaboratore: "Artificial Morality - Virtuous Robots for Virtual Games" [Danielson 1992]. Alcuni dei principi studiati da Danielson si estendono già al campo del diritto.

2. Il Dilemma del Prigioniero e le situazioni contrattuali.

Una delle più importanti intuizioni della moderna filosofia morale è che molte relazioni contrattuali condividono la struttura del dilemma del

[*] Istituto di Filosofia del Diritto ed Informatica Giuridica, Ludwig-Maximilian Universität Monaco di Baviera, Germania

prigioniero. Questa intuizione è collegata all'uso del computer che diviene così strumento del filosofare.

Il dilemma del prigioniero descrive una vicenda ambientata negli USA: dopo una rapina, due vagabondi vengono arrestati vicino al luogo del crimine. Lo sceriffo è convinto di avere catturato gli autori del reato ma non è in grado di provarlo. Decide dunque di chiudere gli indiziati in celle separate e spiega loro la situazione:

1 - Se uno di loro si dichiarerà colpevole - ma l'altro no -, quello che avrà confessato sarà rilasciato per essere diventato testimone dell'accusa[1]. L'altro dovrà fronteggiare ad un lungo periodo di carcerazione.

2 - Se tutti e due si dichiareranno colpevoli, non si potrà dar luogo alla rinuncia all'azione penale e saranno entrambi condannati alla prigione, ma solo per periodi non molto lunghi, giacché la confessione sarà considerata una circostanza attenuante per entrambi.

3 - Se nessuno dei due si dichiarerà colpevole, la Corte non avrà altra scelta che condannare entrambi solamente ad un breve periodo di carcerazione - per vagabondaggio.

La situazione del prigioniero può essere trasferita in una matrice di teoria dei giochi (un numero alto non implica un lungo periodo in prigione, ma al contrario più alto è il numero, migliore è la situazione).

	Diniego (cooperazione)	Confessione (defezione)
Diniego (cooperazione)	2,2	0,3
Confessione (defezione)	3,0	1,1

Questa matrice mostra cosa probabilmente accadrà: entrambi confesseranno. Confesseranno perché ognuno di loro dirà a se stesso: se il mio partner confessa, sarà meglio anche per me confessare, altrimenti il mio periodo di carcerazione sarà lungo. Se invece l'altro non confessa, trarrò maggior vantaggio confessando perché sarò rilasciato.

[1] Nel testo "for giving State's evidence". La "State's evidence" statunitense corrisponde alla "Queen's evidence" cioè la testimonianza contro se stessi ed i propri complici in cambio della rinuncia all'azione (immunity from prosecution). (N.d.T.)

In termini di teoria dei giochi questo implica che la strategia della confessione prevale su quella del silenzio.

O forse qui si applica il cauto principio "maximin". Tale principio sostiene che la linea di condotta da scegliere è quella che garantisce nel caso peggiore il risultato comparativamente migliore. Il risultato peggiore, nel caso della confessione, è un medio periodo di carcerazione; in caso di diniego, è un lungo periodo di carcerazione.

È degno di nota che la linea razionale di condotta per ogni prigioniero individualmente, non è invece sensata per entrambi. Sarebbe meglio mantenere il silenzio; ciò comporterebbe per tutti e due solo un breve periodo di carcerazione.

L'importanza filosofica fondamentale del dilemma del prigioniero è stata intuita già da lungo tempo. Secondo le mie conoscenze, il filosofo canadese David Gauthier fu il primo a porre l'attenzione sul fatto che nei contratti ci si trova di fronte ad una situazione sostanzialmente uguale [Gauthier 1969]. Ciò è certamente vero per contratti conclusi nello "stato di natura", dove il potere dello Stato nulla può imporre. Ognuna delle parti potrebbe ragionare in questo modo: mi piacerebbe adempiere la mia parte del patto, ma come posso essere sicuro che l'altro farà lo stesso? Senz'altro, l'altro esaminerà la possibilità che io possa non adempiere; quindi, per minimizzare il potenziale danno, nessuno onorerà l'accordo. Questo sarà vero non solo per il totale inadempimento del contratto, ma anche per il parziale adempimento, che appare anzi il caso più realistico. Tali contratti, conclusi in 'stato di natura', esistono ancora nella nostra società. Ad esempio il patteggiamento, cioè l'accordo sulla pena concluso fra giudice, difesa e pubblico ministero, nel processo penale. In Germania il patteggiamento nel processo penale non è considerato lecito, ciononostante ha luogo frequentemente [Schünemann 1990]. E' possibile per il giudice ignorare l'accordo, o per un imputato dare falsa confessione ed incriminare una terza parte innocente.

Tutti noi conosciamo gli strumenti legali a tutela delle obbligazioni contrattuali. Ma forse molte persone onorano i loro accordi senza riguardo all'ombra minacciosa dello Stato. Come nel caso di relazioni d'affari di lungo periodo, che possono essere simulate iterando il dilemma del prigioniero. Ogni parte adempie in buona fede per timore che un rapporto lucrativo possa interrompersi. A questo proposito il politologo americano Robert Axelrod ha sfidato scienziati di tutto il mondo in un torneo – politologi, psicologi, biologi, e teorici dei giochi [Axelrod 1984]. Ogni concorrente aveva sviluppato un programma per computer per verificare il dilemma del prigioniero iterato. Ogni

programma veniva messo in gara contro tutti gli altri programmi, incluso se stesso, ed anche con un programma che eseguiva mosse casuali. Il risultato straordinario alla fine di due tornei fu che ogni volta vinse il programma più semplice: TIT FOR TAT[2]. Questo programma inizia cooperando e poi continua imitando la mossa dell'avversario. Se l'avversario coopera, continuerà la cooperazione. Se il partner defeziona, anche lui defezionerà – finché il partner ancora una volta non passi a cooperare: quindi seguirà l'esempio immediatamente (adattandosi alla cooperazione).

Il motivo di questi successi è facilmente esplicabile, confrontando TIT FOR TAT al comportamento di un programma che agisce solo con la defezione (abbiamo visto che questo è individualmente razionale in un dilemma del prigioniero non iterato). Se i due programmi giocano l'uno contro l'altro, TIT FOR TAT perderà a causa della sua iniziale manifestazione di fiducia. Questa sconfitta, in ogni caso, avverrà soltanto al primo round e non porterà al concorrente molti punti. TIT FOR TAT d'altra parte accumulerà punti su punti giocando contro se stesso. Il risultato moralmente soddisfacente sarà lo sviluppo di una stabile popolazione di giocatori di TIT FOR TAT da cui il defezionista ostinato sarà escluso[3].

Il principio è plausibile; ma ancor più è degno di nota che nessun altro programma, non importa quanto ingegnoso o sofisticato, sia riuscito a mettere nel sacco TIT FOR TAT. Peter Danielson ammette di aver trascorso la *"better part of a weekend"* per sviluppare una strategia, che sicuramente avrebbe potuto battere TIT FOR TAT. Il suo desiderio si dimostrò un'illusione. Tutto ciò sembra confermare l'asserzione dell'antropologo Levi-Strauss: la legge di reciprocità è tanto

[2] (NdT) Traducibile con "colpo su colpo". Nota Philipps in una e-mail successiva che "occhio per occhio" o "pan per focaccia" appaiono troppo unilateralmente negativi, in tedesco si direbbe "Wie du mir, so ich dir!" (come tu a me così io a te), che comprende anche reazioni positive.

[3] Quelli che sono lieti per i risultati dei tornei di Axelrod, dovrebbero ricordare due cose: 1) Oltre al dilemma del prigioniero c'è il "chicken game" che è anche di fondamentale importanza filosofica ma – come vedremo – porta a risultati discutibili dal punto di vista morale. 2) Nei tornei di Axelrod i giocatori si incontrano in coppie variabili di due giocatori. Nel dilemma del prigioniero di n-persone, i risultati potrebbero essere completamente differenti, specialmente, se i partecipanti restano anonimi. Mentre il comportamento parassitario è destinato a fallire nel lungo periodo in un gioco di due giocatori, sembra plausibile in un gioco di n-giocatori che singoli parassiti si associno ai gruppi cooperanti. Schüßler ha attenuato le rigide regole di Axelrod ed ha eseguito (in tal senso) simulazioni al computer [Schüßler 1990]. Ha certamente raggiunto risultati notevolmente 'positivi'.

fondamentale nelle interazioni sociali quanto la legge di gravità lo è nella fisica.

In effetti, il fenomeno della reciprocità è stato riscontrato anche tra gli animali. I biologi distinguono due varianti dell'altruismo animale [Danielson 1992 pag. 39-51]. L'altruismo di parentela significa che i vantaggi individuali sono sacrificati a favore del moltiplicarsi dei geni. L'altra variante è l'altruismo reciproco: un vantaggio attuale viene sacrificato a favore di futuri vantaggi che potranno derivare da relazioni perduranti nel tempo.

L'altruismo reciproco funziona solo se è possibile identificare il partner, ci sono però esempi istruttivi di surrogati per l'identificazione: nel loro *"home port"*[4], alcuni grandi pesci predatori si fanno pulire i denti da piccoli pesciolini (operazione che fornisce a questi ultimi nutrimento). Fuori dallo *"home port"* i pesci predatori divorerebbero i loro amici perché non li identificherebbero. Entrambe le specie di pesci non possono identificarsi tra loro, ma possono identificare il posto nel quale un pacifico ed utile incontro ha luogo.

3. Il Dilemma del Prigioniero come gioco di filosofia morale.

I modelli di comportamento che ho descritto possono essere chiamati morali? Forse si potrebbe parlare di quasi-moralità o proto-moralità nei comportamenti animali e di morale artificiale nei programmi di computer. Peter Danielson, l'autore di *"artificial morality"* è contro queste etichette: tutti questi fenomeni in ultima analisi sarebbero per lui meccanismi d'egoismo e nulla più. Se il dilemma del prigioniero è appropriato per la descrizione di un comportamento morale, esso deve dimostrarsi tale in una singola situazione di scelta, dove nessun futuro guadagno è all'orizzonte. Io seguo l'approccio filosofico di Danielson, ma non la sua scelta terminologica. Moralità artificiale - analogo a vita artificiale - è di gran lunga troppo elegante e generale come termine per essere confinato ad un significato così limitato. D'altro canto egoismo e moralità sono connessi in così tanti luoghi, che un termine che li possa raggruppare è utile. Il principio di reciprocità è una delle radici biologiche della morale e sarebbe sbagliato escluderlo dai discorsi morali, nonostante la morale, in senso moderno, non possa più essere limitata alle sue origini.

[4] Letteralmente, per le navi: "porto di immatricolazione" (N.d.T.).

Il dilemma morale del prigioniero, nella versione di Danielson, differisce dal dilemma del prigioniero standard nel presupposto che i partecipanti conoscono la strategia della controparte e possono quindi predire le sue mosse. Ciò potrebbe ridurre l'interesse in termini di teoria dei giochi, ma non in termini filosofici. Supponiamo che io percepisca di trovarmi di fronte ad un partner cooperativo. Da egoista defezionerei allo scopo di ricavare il massimo profitto. Da altruista probabilmente coopererei. E' plausibile che le mie azioni possano essere considerate "immorali" nel primo esempio, e "morali" nell'altro. Naturalmente non è molto credibile che il partner, percependo la mia strategia egoistica, continui a cooperare, ma ciò non è impossibile – ed i valori morali raramente sono una questione di verosimiglianza.

Danielson ha scritto alcuni programmi in PROLOG che consentono ai computer di identificare interattivamente le rispettive strategie. All'interno di questo contesto - singoli incontri con mutue identificazioni - le strategie si evolvono secondo i principi Darwinistici. I risultati non restano a livello elementare e per certi versi sono sorprendenti.

Forse la strategia più degna di nota è quella del "cooperatore condizionato" introdotta da David Gauthier. Il cooperatore condizionato coopererà soltanto con coloro che coopereranno a loro volta. Lui o lei rifiuterà di cooperare con "massimizzatori incondizionati" che sfrutteranno la disponibilità del partner a cooperare. Perseguendo il suo profitto, il cooperatore condizionato promuoverà anche il bene comune. Gauthier dunque credeva di aver trovato un punto in cui razionalità e moralità coincidevano.

Danielson, tuttavia, trovò una lacuna di razionalità nella morale di Gauthier. Perchè, dice Danielson, non sfruttare quei partners che cooperano sempre e incondizionatamente? La ragione richiede di cooperare solo con quelli che cooperano condizionatamente – sotto la condizione che, a turno, essi cooperino (con noi). Danielson, dunque, introdusse il "cooperatore reciproco". Un esempio potrebbe mostrare le differenze tra i due cooperatori: un mercante 'reciprocamente cooperativo' vende ad un prezzo adeguato solo se il compratore contratta per questo; un mercante 'condizionatamente cooperativo' vende sempre al prezzo adeguato (ma non svenderà beni, naturalmente).

Nella replica di Gauthier il 'cooperatore reciproco' di Danielson è un "mostro morale". Danielson tuttavia non resta chiuso in una posizione rigorosamente razionale e formale, ma segue Gauthier nel campo dei ragionamenti sostanziali [Danielson 1992, pp. 61-123]. Egli suggerisce un 'test evolutivo' in cui senz'altro può essere notata l'influenza delle

ricerche di Axelrod. Come sarebbe una società nella quale sono presenti i 'cooperatori condizionati' di Gauthier? Questi non solo si sosterrebbero l'un l'altro ma sosterrebbero anche quei membri di buona indole che cooperano in ogni caso, i "cooperatori incondizionati". Questo, di per sé, non è un male, ma appoggiando i cooperatori incondizionati verrebbero indirettamente appoggiati anche gli immorali 'massimizzatori incondizionati', che, per natura, sfruttano i cooperatori incondizionati. Se, d'altra parte, si è adottata una posizione di cooperazione reciproca, non solo i 'cooperatori incondizionati' ma anche i 'massimizzatori incondizionati' saranno spinti fuori dalla società. Un esempio di questo andamento [Gauthier 1988]: un agricoltore ammazza tutti i conigli nei suoi campi per distruggere la fonte primaria di cibo delle volpi. Riportato in regola pratica, questo significherebbe: eliminare il debole allo scopo di togliere ai cattivi la loro preda naturale! Più specificamente: eliminare le persone negligenti per diminuire la fonte di reddito dei profittatori! O ancora, in modo più pertinente: sbarazzarsi di coloro che richiedono asilo per rimuovere il principale obiettivo di attacco dei neo-Nazisti!

E' interessante notare che concetto ed argomentazione correlata, simili al gioco, possono condurre ad un differente punto di vista. Alla fine del XVIII secolo, Jeremy Bentham scrisse "In difesa dell'usura". Bentham sostiene che l'esistenza dell'usura è benefica per la società. Gli usurai sono lucci in mezzo alle carpe: essi scuotono le persone dalla loro lentezza e creano un senso di auto-responsabilità nella popolazione. Il libro fu molto influente, anche in Germania. I liberali riuscirono ad abolire la perseguibilità penale per l'usura. Il codice penale tedesco (Strafgesetzbuch), infatti, non faceva menzione dell'usura nella sua versione originale del 1871. Poco dopo, comunque, fu cambiato. Il ragionamento di Bentham è strutturato come una teodicea[5]: il Male è considerato come uno strumento verso il Bene. Questo tipo di ragionamento è importante a tutt'oggi: molti disprezzano i neo-nazisti, ma sono felici che 'il lavoro sporco' lo stiano facendo loro.

E' interessante anche notare che non appena vengono utilizzati conigli e volpi per gli esempi, non appena sono ritratte persone reali, ruoli e situazioni reali, al posto di soggetti astratti, evidentemente fallisce il tentativo di dedurre la moralità dalla razionalità. Cosa

[5] Teodicea è la dottrina del diritto e della giustizia di Dio (dal greco, giustizia di Dio). Il termine fu coniato da Leibniz e posto come titolo di una sua opera, per provare come l'esistenza del Male non contraddica la bontà e la provvidenza di Dio, conciliando in pari tempo la libertà umana con la pre-scienza divina. La parola passò poi via via a significare lo studio esclusivamente filosofico di Dio, sui dati della sola ragione e così si identificò con quella che meglio deve chiamarsi "teologia naturale" (N.d.T.).

penseremmo di uno Stato, di un organo che legifera, e che volontariamente o involontariamente si consideri obbligato da principi di pura ragione? Che ha capitolato come soggetto morale. D'altra parte è anche vero che le ricerche di Danielson non sono un vuoto gioco intellettuale, ma qualcosa da prendere più seriamente. E' evidente che gli schemi di ragionamento in questione sono realistici e che vengono utilizzati ripetutamente nell'argomentazione morale, giuridica e politica.

4. Il *"Chicken Game"*[6] – Competizione e bene comune.

Per molto tempo, il dilemma del prigioniero ha rappresentato, nell'ambito della teoria dei giochi, l'unico esempio rilevante in campo giuridico ed etico. Danielson insiste giustamente sul fatto che da questi punti di vista il '*chicken game'* non risulta essere meno importante. Il '*chicken game'* viene giocato dai ragazzi americani in parecchie varianti; questa è una: due adolescenti si dirigono velocemente con le loro automobili l'uno contro l'altro. Il primo che devia dalla direzione della collisione è un *"Chicken",* e perde. Naturalmente è possibile che entrambi cambino direzione, o che nessuno dei due lo faccia.

Il '*chicken game'* differisce dal dilemma del prigioniero per il fatto che il peggior risultato possibile per entrambi i partecipanti si verifica in seguito a reciproca "defezione" (entrambi restano nella stessa corsia). Nel gioco reale uno dei possibili risultati sarebbe addirittura la morte. Se invece si evita l'altra macchina, si può perdere la faccia, ma rimanere vivi – questo sarebbe il secondo peggior risultato. Al contrario, la reciproca defezione nel dilemma del prigioniero spinge al secondo peggior risultato, e il peggior caso è sostenuto dal singolo cooperatore.

Nei libri di teoria dei giochi (sotto l'influenza del dilemma del prigioniero), è consuetudine contrassegnare il risultato di una defezione unilaterale con la lettera T (tentazione) e il risultato di una cooperazione unilaterale con la lettera S (premio del babbeo[7]); R (ricompensa) rappresenta la reciproca cooperazione e P (pena) la reciproca defezione. Per il dilemma del prigioniero, la scala di preferenza del risultato sarebbe: T > R > P > S. Per il '*chicken game'*, P e S sono invertiti e la scala delle preferenze è: T > R > S > P. Naturalmente i codici in lettere hanno un significato solo per il dilemma del prigioniero, ma vengono mantenuti nella descrizione di altri giochi per ragioni di comparabilità.

[6] Traducibile in italiano, con "gioco del pollo" (N.d.T.).
[7] Nel testo "Sucker's payoff" (N.d.T.).

Se il dilemma del prigioniero è particolarmente adatto a descrivere situazioni che possono essere regolate tramite accordi, il '*chicken game'*, invece, può adattarsi a due tipi differenti di situazioni. La prima è una situazione competitiva: due concorrenti possono distruggersi l'un l'altro – ma solo quello che abbandona, prima che ciò avvenga, sarà in svantaggio. La seconda situazione che può essere descritta usando il '*chicken game'* è quella che riguarda il mantenimento di un bene comune. Un buon esempio è il seguente [Taylor e Ward 1982]: i campi di due coltivatori olandesi sono situati dietro un canale di scolo. Prevedendo un'ondata di piena si devono rinforzare gli argini. Ogni coltivatore può decidere di lavorare sul canale (C) oppure no (D). Se nessuno di loro farà qualcosa (D/D), il canale si frantumerà ed accadrà la catastrofe. Se entrambi si metteranno al lavoro (C/C), il canale terrà e nessuno perderà molto tempo. Se solo uno lavorerà sul canale (D/C o C/D), questo terrà ancora, ma nel mentre che un coltivatore starà perdendo molto tempo per lavorare, il suo vicino guadagnerà sul tempo risparmiato. C'è un mutuo interesse alla preservazione di un bene comune, ma ciascun interesse privato mira a raggiungere l'obiettivo a spese dell'altro.

Il fatto che le situazioni competitive e il conseguimento di un bene comune possano entrambe essere rappresentate dal '*chicken game'*, evidenzia un alto livello di astrazione del gioco. Una differenza può essere vista nel fatto che, nelle situazioni competitive la defezione si manifesta in forma d'azione, mentre nell'altro contesto non competitivo essa si manifesta come omissione. L'esempio che segue mostrerà che competizione e bene comune sono comunque interdipendenti: nell'ora di punta le strade principali di una certa città sono troppo intasate per permettere a chicchessia di muoversi. Questa è una situazione di competizione paralizzata. Per questo motivo viene istituito un efficiente sistema di trasporto pubblico che porta verso una situazione di bene comune. Purtroppo molti evadono il pagamento del biglietto e conseguentemente il prezzo deve essere aumentato: se aumenta il numero delle persone che viaggia senza biglietto e se gli altri non saranno disposti a pagare i prezzi più alti che ne deriveranno, il sistema tracollerà.

Una differenza sostanziale tra il dilemma del prigioniero ed il '*chicken game'* consiste nel fatto che, mentre nel dilemma del prigioniero il modo migliore per indurre il partner a cooperare è quello di fargli delle promesse, nel '*chicken game'* è quello di minacciarlo. Quando qualcuno minaccia di non cooperare, nel dilemma del prigioniero, spingerà anche l'altro a non cooperare, salvo che costui sia un santo o sia fuori di testa;

fatta invece una promessa credibile, il soggetto coopererà qualora decida di comportarsi in modo moralmente sano. Nel '*chicken game'*, al contrario, è razionale la cooperazione sotto una credibile minaccia di defezione fatta dall'altra persona. Se promessa e minaccia siano manifestate seriamente o bluffando, naturalmente, è un'altra questione.

Una promessa è un impegno a fare per se stessi (io farò qualcosa di utile), una minaccia si riferisce principalmente all'altro (tu soffrirai un danno). Questa considerazione può essere generalizzata distinguendo quattro tipi fondamentali di protagonisti, a seconda di chi coopera con cooperatori o con defezionisti e di chi defeziona dai cooperatori o dai defezionisti. Il dilemma del prigioniero dà importanza all'"*ego*",il '*chicken game'* dà importanza all'"*alter*". La seguente matrice affronta una tale tipologia.

		Dilemma del prigioniero (tipi "ego"-basati)	*"Chicken Game"* (tipi "alter"-basati)
C / C	C / D	L'ingenuo	Il rammollito
C / C	D / D	L'affidabile	Il corretto
D / C	C / D	L'inaffidabile	Il bullo
D / C	D / D	Il cauto	Il duro

Dal punto di vista morale, il *chicken game* è molto più problematico del dilemma del prigioniero. L'*affidabile* che merita la nostra stima, è rimpiazzato dal *corretto* che entrerà in un incrocio con la luce verde anche se egli vede un'altra macchina che si avvicina dal lato. Io nutro sentimenti contrastanti per il *corretto*. E' anche inquietante che nei giochi del tipo '*chicken game'* prospereranno coloro che sono *rammolliti* nell'affrontare i *duri* e *duri* nell'affrontare i *rammolliti*. In un dilemma del prigioniero tale comportamento sarebbe irrazionale, ma essere *"The Bully"* è una strategia vincente per il '*chicken game'*.

A confronto con il dilemma del prigioniero, il '*chicken game'* genera un numero inferiore di comportamenti che siano nello stesso tempo sia morali che razionali. Potrebbe essere possibile concludere che il legislatore ha più ragione ad intervenire nelle situazioni che somigliano ai giochi del tipo '*chicken game'* che nelle situazioni come il dilemma del prigioniero. Simulazioni e giochi al computer potrebbero essere utili per chiarire tali questioni.

5. Preparare l'universo del discorso morale.

Tentiamo di ricostruire il mondo del comportamento morale, basandolo sull'interazione di regole morali.

Al più basso livello, si potrebbe agire senza riflettere, sia cooperativamente che con la defezione (C e D). Il comportamento cooperativo potrebbe essere considerato come *ingenuamente* morale, mentre la defezione potrebbe essere considerata come *ingenuamente* immorale.

Lo stato di assenza di riflessione non avrà possibilità di durare, specialmente appena a qualcuno sarà data la possibilità d'essere accostato ad un partner defezionista. Egli o lei svilupperà una regola per determinare il proprio comportamento. Questa regola lo/la porterà a cooperare con quelli che cooperano ed a defezionare da quelli che defezionano (Gauthier).

Dal momento che l'altro giocatore ha due opzioni (defezionare e cooperare), e poiché la reazione di ciascuno, a turno, può essere quella di defezionare o di cooperare, esistono quattro possibili condotte corrispondenti ai quattro tipi fondamentali di comportamento prima ricordati. Nelle seguenti espressioni, la seconda lettera rappresenta il presunto comportamento del partner, mentre la prima lettera indica la propria reazione. DC, per esempio, significa: io defezionerò dai partners che cooperano.

1 – CC, CD
2 – CC, DD
3 – DC, CD
4 – DC, DD

Primo: è possibile ancora cooperare, ma ora riflettendoci su, non importa se l'altro defeziona o coopera.

Secondo: la regola potrebbe essere di cooperare con chi coopera e di defezionare da chi defeziona.

Terzo: si potrebbe, al contrario, defezionare da chi coopera e cooperare con chi defeziona.

Quarto: si potrebbe defezionare sempre – non importa se il partner coopera o defeziona.

La prima e la seconda regola devono essere chiaramente considerate morali: la prima in maniera irrazionale, forse da santo, mentre la seconda regola di cooperare, pur non desiderando di essere ingannato, è razionale. Come già ricordato in precedenza, Gauthier ha suggerito questa regola.

Ma la riflessione può essere portata un gradino più avanti: la scelta tra cooperazione e defezione può essere basata sul comportamento effettivo del partner, invece di basarsi semplicemente sul comportamento presunto.

Giacché ci sono quattro regole di primo livello, una regola di secondo livello potrebbe essere rappresentata così (è stata scelta una rappresentazione tabulare per dimostrare la connessione strutturale con le regole primo livello, che sono o di cooperazione con o di defezione da):

C(CC, CD),
C(CC, DD),
D(DC, CD),
D(DC, DD)

Questa è la controversa regola che Danielson ha proposto come miglioramento rispetto a quella di Gauthier. Il giocatore coopera solo con quelli che altrimenti defezionerebbero. Come già chiarito, il nocciolo di questa soluzione è il tentativo di portare la razionalità un gradino più avanti, per trovare il punto dove moralità e razionalità coincidono e dare così una spiegazione razionale del comportamento morale.

Nell'animata disputa tra Gauthier e Danielson (il santo e l'ingenuo che sono incondizionatamente morali cadrebbero vittime della morale di Danielson), non è stata data molta attenzione ad una differenza strutturale: la regola di Gauthier è una regola di primo livello, basata semplicemente sul comportamento dell'altro. La regola di Danielson, in ogni caso, è una regola di secondo livello perché basata sulla regola di primo livello dell'altro. La regola di Danielson non può essere espressa come una regola di primo livello, ma la regola di Gauthier può comunque essere scritta anche come regola di secondo livello: C(CC, CD), C(CC, DD), D(DC, CD), D(DC, DD). Ogni regola, in generale, può essere espressa in termini di una riflessione di più alto livello.

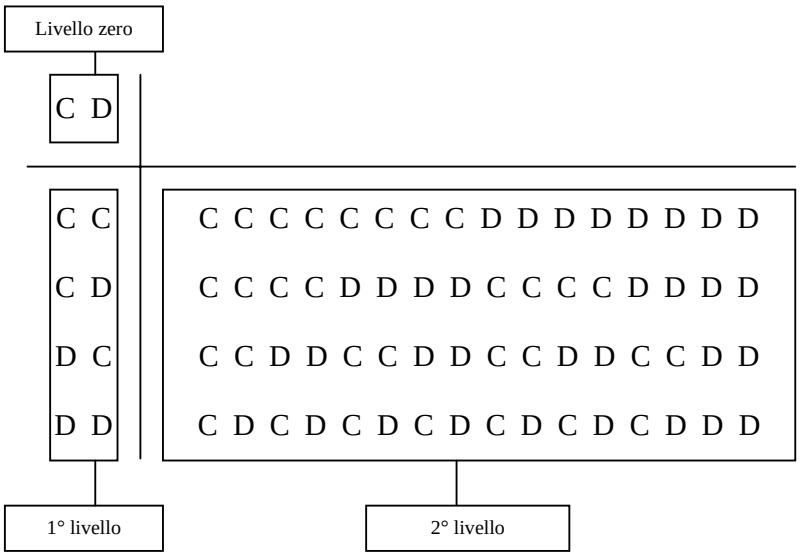
Il sottostante concetto strutturale può essere riassunto in questo modo:

- le regole del livello zero esprimono un semplice comportamento: C e D;
- le regole di primo livello basano il comportamento sulle regole di livello zero, cioè il presunto semplice comportamento C o D del partner;
- le regole di secondo livello sono conseguentemente basate sulle regole di primo livello;
- le regole del terzo livello sulle regole di secondo livello e così via.

Ad ogni livello di riflessione, cambia il soggetto della regola.

Le regole del livello zero sono singole espressioni e di esse ce ne sono due ($=2^1$); le regole del primo livello hanno due termini e di esse ce ne sono quattro ($=2^2$); ci sono sedici ($=2^4$) regole di secondo livello a quattro termini. Quindi il numero delle combinazioni esplode: ci sono 65.536 ($= 2^{16}$) possibili regole di terzo livello, con sedici termini ciascuno.

Tabella 1.



E' impossibile immaginarsi un simile numero di casi, ma le 4 possibili regole di primo livello e forse persino le sedici regole di secondo livello possono essere mentalmente raffigurate. Tutto quello che è richiesto è la familiarità con le tavole di verità della logica proposizionale.

Nella prima colonna a sinistra sono date le due espressioni elementari di livello zero: semplice cooperazione e semplice defezione. Nella seconda colonna, le regole di primo livello basate su modelli elementari di comportamento: quattro possibilità per reagire alla cooperazione o alla defezione, cooperando o defezionando. Nella colonna di destra, si trovano le sedici possibilità di secondo livello di reagire con la cooperazione o con la defezione alle regole date di primo livello[8].

[8] Mi piacerebbe richiamare un concetto degli anni 50 comparabile a questo: il filosofo [Gunther 1963] tentò di sviluppare una logica di riflessione che avrebbe dovuto oscillare tra i poli "I" e "You" ed un terzo polo "It". Il suo proposito era quello di creare una base per "macchine coscienti" – come Danielson vuole creare una base per "macchine morali".

Le lettere scritte nella tabella rappresentano soltanto la prima parte delle regole (scritta sulla parte più a sinistra nell'espressione formale). Quello cui fanno riferimento le lettere, può comunque essere ricostruito dalla tabella. Possono essere usate per la rappresentazione mnemonica delle regole le espressioni logiche "and" o "if … then", che sono contenute nella tabella,. Per esempio, la prima delle 16 colonne significa che uno coopererà sul secondo livello di riflessione, non importa quale regola il partner seguirà. Questa colonna nella logica corrisponde alla tautologia[9].

Una regola di livello più alto diventa gestibile nel caso sia possibile separare qualche termine decisivo e raggruppare tutti gli altri insieme. In questo modo la regola di secondo livello di Danielson, con i suoi quattro termini, può essere scritta come: C (CC, DD), *else* D.

Ma le regole di più alto livello possono ancora essere considerate realistiche? Almeno per le regole di terzo livello, la risposta è certamente affermativa, per la semplice ragione che deve essere possibile reagire con la cooperazione o con la defezione a regole di secondo livello come quella di Danielson. Si potrebbe respingere la regola di Danielson con la motivazione prima esposta del "mostro morale". La risposta di terzo livello sarebbe: D (C (CC, DD), else D), else C[10]. Una tale regola appare irragionevole, ma non è necessariamente così. E' concepibile che le regole si specializzino quando il modello diviene più differenziato. Alcune regole potrebbero, per esempio, essere limitate alla soppressione di altre specifiche regole, nel qual caso esse devono essere aiutate da altre regole per sopravvivere: in questo modo inizia la divisione delle funzioni.

Ogni nuovo livello di riflessione può essere collegato ad un incremento nella differenziazione: in termini di razionalità e di divisione delle funzioni. Perché le qualità di quelle regole che si sono dimostrate di successo possono essere combinate in una nuova regola. Il procedimento non è diverso da quello che si utilizza nell'incrocio tra specie per ottenere individui con caratteristiche superiori.

---

[9] Nella logica tradizionale la tautologia è una proposizione in cui il predicato "dice lo stesso" del soggetto: per es. "i quadrupedi hanno quattro zampe". Nella logica moderna, il termine designa le verità logiche del calcolo preposizionale (N.d.T.).

[10] L'impressione di "giudizio morale" che suscita un tale rifiuto, non è probabilmente una coincidenza. Non appena le regole non sono più basate sull'effettivo comportamento dell'altro, ma sulle regole dell'altro (secondo livello e superiori), esse sembreranno "morali" (o "immorali"). "Morale", in questo contesto, è preso nel senso che gli ha dato la tradizione filosofica di Thomasius e Kant, cioè che il comportamento "interno" è opposto alla "legge" come un comportamento "esterno" .

In un certo senso, la regola di Gauthier CC, DD può essere considerata come un incrocio di successo tra le regole elementari C e D. Un algoritmo genetico[Goldberg 1989] potrebbe gestirlo come segue: Primo, C e D sono portati al primo livello di riflessione. Le nuove espressioni possono essere considerate come filamenti di "cromosomi". Essi possono essere scissi in due parti e nuovamente ricombinati generando così un incrocio. Le due regole più differenziate, risultano essere: la regola di Gauthier ed il ruolo di "bullo" nel '*chicken game'*.

C ⇒ CC, CD ⇒ CC … CD ⇒ CC, DD (Gauthier)

D ⇒ DC, DD ⇒ DC … DD ⇒ DC, CD ("bullo" nel '*chicken game'*)

Possiamo continuare il processo di allevamento: innalzando le regole (Gauthier e "bullo") al successivo più alto livello ed incrociandole di nuovo. Se il "filamento di cromosomi" è scisso al primo "punto di rottura" saranno generate le seguenti regole: la regola di Danielson ed una seconda che probabilmente non sarà di successo:

C …, C …, D …, D … ⇒ C …, D …, D … ⇒ C …, D …, C …, C … (probabilmente non di successo)

D …, D …, C …, C … ⇒ D …, D …, C …, C … ⇒ D …, C …, D …, D … (Danielson)

Una volta che le nuove regole siano state progettate, meccanicamente o intellettualmente, esse potrebbero venire testate in gara. Giocherebbero l'una contro l'altra, come negli esperimenti di Axelrod. Si chiarirà quali regole cooperano (probabilmente a spese di terze parti), quali s'inseriscono nelle popolazioni delle altre regole e quali sono escluse da tali popolazioni.

I risultati non devono essere abbandonati a se stessi in accordo al principio Darwiniano di fitness. A mio avviso è necessario valutare le regole in riferimento ai concetti morali (che non derivano dalla simulazione al computer) e "allevarle". Ma sarà sempre essenziale porre attenzione al fatto che persino dove la qualità morale è accentuata, le regole siano sufficientemente robuste da prevalere nella "fatica quotidiana" e da primeggiare nella simulazione al computer. Altrimenti la moralità si tingerà di platonico pallore.

L'approccio (combinatorio, classificato in livelli, genetico ed in forma tabulare) qui suggerito – per le mie conoscenze - è nuovo. Danielson

usa un approccio differente: dapprima descrive una regola di comportamento informale senza darne il livello. Per formalizzare le regole, scrive un programma PROLOG, che gira sul computer ma non può essere afferrato intuitivamente. Il nostro metodo ha il vantaggio della trasparenza: livello riflessivo e complessità della regola, sono entrambi immediatamente comprensibili. Inoltre si può vedere se la regola è completa, ad es. se essa dà tutte le azioni e le regole possibili ad un certo livello di riflessione e se le opzioni sono elencate o alcune definite come residuali (…else…). Inoltre si può osservare, a colpo d'occhio, se due regole sono correlate e il grado di relazione può essere determinato con precisione.

Ma quest'approccio rivela un altro problema: il numero di casi indecidibili. L'impressione sarebbe che se due soggetti si trovano in una situazione di gioco, uno dei due deve trovarsi su un più alto livello di riflessione rispetto all'altro affinché ci possa essere una decisione determinabile. Quest'impressione non è molto corretta: ma è vero che il numero di casi che non possono essere decisi aumenta allo stesso livello di riflessione.

In ogni modo, l'esistenza di casi che non possono essere decisi, non comporta un errore nel nostro approccio ma piuttosto problemi relativi alla materia in sé. Sarebbe sbagliato cercare di evitarli applicando un elegante programma per computer. Un esempio:

due giocatori si fronteggiano l'un l'altro usando la regola di Gauthier (primo livello di riflessione):

CC, DD/CC, DD

Non c'è problema decisionale. Ciascuno può permettersi di cooperare per primo; l'altro "risponderà" di conseguenza. Non si verifica alcun problema se si sale al secondo livello di riflessione e si adotta la regola di Danielson:

C (CC, DD), else D/CC, DD

Formalmente il giocatore di sinistra è forzato a cooperare. Ma anche nella sostanza: egli non deve temere di cooperare per primo, perché sa che il giocatore di destra coopererà.

Ma cosa accade se i due giocatori si incontrano sul secondo livello di Danielson?

C (CC, DD), else D/C (CC, DD) , else D

Nessun giocatore può permettersi di cooperare per primo (un pagamento anticipato) perché l'altro ne approfitterebbe. Il fatto che l'altro è pronto ad approfittarne, giustifica la diffidenza.

Questo ci riporta al dissidio tra moralità e razionalità. Possiamo veramente affermare che la prontezza ad approfittare, trasformandosi in giustificata diffidenza, è più razionale che perdere consapevolmente alcune possibilità di sfruttamento e perciò instaurare la fiducia? Io credo che non sia possibile far coincidere razionalità e moralità una volta per tutte (come vorrebbe Danielson), perché sarà sempre possibile raggiungere un più elevato livello di riflessione dove entrambi divergeranno di nuovo. Sarebbe un importante risultato filosofico provare quest'ipotesi di mera congruità transitoria tra moralità e razionalità.

Ci troviamo di fronte non solo al problema formale dell'indecidibilità, ma anche al sostanziale problema della razionalità. E di nuovo vediamo che l'affermazione che la regola di Danielson possa essere immorale - ma in ogni caso più razionale di quella di Gauthier - non è fuori dubbio, almeno non in generale. Ci sono parecchie buone ragioni per sostenere questa tesi. Io accennerò solo ad una di esse: il mercanteggiamento porta all'inerzia ed a perdita di tempo.

Max Weber notava che l'introduzione negli Stati Uniti dei grandi magazzini diede un forte potere al capitalismo moderno. I compratori potevano fare affidamento sul fatto che i beni venivano offerti al più basso prezzo possibile. Non c'erano né il motivo né lo spazio per mercanteggiare. Per i nostri scopi è importante notare che, in accordo con Max Weber, i motivi per questa sorprendentemente razionale invenzione di successo erano in origine di natura morale e religiosa.

Un altro esempio, può essere preso dal *"chicken game"* questa volta, non dal dilemma del prigioniero: due "bulli" si fronteggiano l'un l'altro:

CD, DC/CD, DC

In questa situazione, il primo livello di riflessione già non riesce ad assicurare una decisione. Ciascun "bullo" imposta la sua azione attuale sull'azione dell'altro: egli sarà "rammollito" con il "duro" e "duro" con il "rammollito" ma se l'altro "bullo" sia "rammollito" o "duro" resta aperto.

Ciò vale solo fino a quando un "bullo" raggiunge il successivo livello di riflessione. Un "bullo" di secondo grado sarà "rammollito" con il "duro" e "duro" con il "rammollito", ma in più sarà ora "duro" con un semplice

"bullo" (e "rammollito", detto incidentalmente, con un "retto" - il secondo termine in parentesi – che non cede alla prepotenza dell'altro).

D(CC, CD), C(CC, DD), D(DC, CD), C(DC, DD)

Se l'altro "bullo" seguisse il primo nel secondo livello di riflessione, ne conseguirebbe una nuova situazione di stallo.

Sorprendentemente il "bullo" di secondo grado è guidato da quasi la stessa regola del "cooperatore reciproco" (il soggetto morale di Danielson) nel dilemma del prigioniero; solo gli ultimi termini delle regole sono differenti: D …, C …, D …, C … - ("bullo" di secondo grado) - D …, C …, D …, D … (Danielson). Questa rassomiglianza può essere spiegata "geneticamente", come abbiamo visto prima.

Forse avrebbe senso risolvere tali situazioni di stallo usando principi generali di diritto, morale o buon senso come regole di default. Il primo di tali principi da suggerire sarebbe quello della generalizzazione, forse l'imperativo categorico di Kant. In ogni modo, usare questo principio comporterebbe lunghe ricerche che, a loro volta, potrebbero essere aiutate dalle tecniche della morale artificiale. L'imperativo categorico non è affatto chiaro come sembra.

Se, per esempio, l'imperativo categorico dovesse essere interpretato come l'istanza di coerenza pratica (un punto di vista che è suggerito a volte da Kant), la domanda sarebbe: cosa rende meno coerente praticamente per due prigionieri "defezionisti" dover passare in prigione un periodo di tempo medio, che per due prigionieri "cooperativi" dover passare in prigione solo un breve periodo di tempo? E' meno piacevole. Questo è tutto.

Sorprendentemente, troviamo qui la versione teorica in termini di teoria dei giochi di un'argomentazione che Hegel usava per confutare Kant: "e se non ci fosse caparra, quale contraddizione ci sarebbe?" (Kant argomentava che se ogni trustee volesse appropriarsi indebitamente del denaro affidatogli, l'istituto del deposito non esisterebbe).

Davanti a tali difficoltà, suggerisco di cominciare con il principio, moralmente ambiguo ma facilmente formalizzabile, tipico dell'etica professionale: "Birds of a feather flock together"[11] (in tedesco, più

[11] Vale a dire letteralmente: "Uccelli di una sola specie fanno stormo insieme" traducibile nel proverbio italiano "ogni simile attrae il simile", oppure "chi si somiglia si piglia". Ma sempre nel genere, anche se meno in point dei precedenti: "a rubare a un ladro non è peccato", "ama chi t'ama rispondi a chi ti chiama", "amore con amor si paga" (N.d.T.).

drastico: “Eine Krähe hackt der anderen kein Auge aus“ [12]). Una tale regola di default non sarebbe irrealistica nemmeno con giocatori prevalentemente defezionisti. Dopo tutto, la solidarietà non è presente soltanto tra persone per bene, anche i disonesti hanno un po’ d’onore. In modo più restrittivo, una simile regola di reciprocità è stata già proposta da Danielson [1992, pp. 79 – 81].

La regola di default pertanto sarebbe che, se due giocatori si fronteggiano l’un l’altro sotto la stessa regola di comportamento e se la loro decisione non è determinata da questa regola, essi dovrebbero cooperare.

Una volta che questa semplice regola di default è testata, possiamo continuare…..

[12] Vale a dire letteralmente: “Una cornacchia non becca l’occhio dell’altra” che trova il corrispondente nel proverbio italiano “cane non mangia cane”, anche se il corrispondente inglese lascia implicare comportamenti cooperativi commissivi e non soltanto omissivi, come nelle traduzioni tedesca ed italiana corrispondente a quella tedesca (N.d.T.).