UNIVERSITY
OF TRENTO

DEPARTMENT OF INFORMATION AND COMMUNICATION TECHNOLOGY

38050 Povo – Trento (Italy), Via Sommarive 14
http://www.dit.unitn.it

IL RAGIONAMENTO CONTROFATTUALE: UN MODELLO E LA SUA APPLICAZIONE AL RAGIONAMENTO PRATICO

Roberta Ferrario

December 2002

Technical Report # DIT-02-101

Università degli Studi di Milano
Dottorato in Filosofia
Cotutela con l'Università di Strasburgo

Il ragionamento controfattuale: un modello e la sua applicazione al ragionamento pratico

Roberta Ferrario

Ciclo XV
A. A. 2002/2003

Coordinatore: Prof. Giambattista Gori
Relatore: Prof. Andrea Bonomi

Université Marc Bloch de Strasbourg
UFR PLISE
Cotutelle avec l'Université de Milan

Le raisonnement contrefactuel : un modèle et son application au raisonnement pratique

Roberta Ferrario

Doctorat Nouveau Régime
A. A. 2002/2003

Thèse préparée sous la direction du Prof. André Thibault

Indice

Alla mia famiglia

S'i' fosse fuoco, arderei 'l mondo;
s'i' fosse vento, lo tempestarei;
s'i' fosse acqua, i' l'annegherei;
s'i' fosse Dío, mandereil' en profondo;
s'i' fosse papa, allor serei giocondo,
ché tutti cristiani imbrigarei;
s'i' fosse 'mperator, ben lo farei:
a tutti tagliarei lo capo a tondo.
S'i' fosse morte, andarei a mi' padre;
s'i' fosse vita, non starei con lui:
similemente faria da mi' madre.
S'i' fosse Cecco, com' i' sono e fui, torrei le donne giovani e leggiadre:
le zop[p]e e vecchie lasserei altrui.

[Cecco Angiolieri, *sonetto V*]

Prefazione e Ringraziamenti

L'unica prefazione necessaria a questa tesi è un'avvertenza per gli eventuali lettori: la tesi tratta di controfattuali e razionalità, ma l'ipotesi principale da cui prende le mosse è che gli esseri umani siano degli agenti razionali; questa ipotesi è essa stessa completamente controfattuale.

La brevità della prefazione sarà invece compensata dalla lunghezza dell'elenco di persone da ringraziare, lunghezza causata, almeno in parte, dall'essenza stessa della tesi che, essendo un lavoro interdisciplinare inserito in un programma di cotutela, mi ha portato a trascorrere i tre anni del dottorato in città e istituzioni diverse.

Comincerei da Trento, la città dove ho trascorso la maggior parte del mio tempo e dove hanno sede l'università e l'istituto (l'ITC-IRST) che mi hanno a lungo ospitato mettendomi a disposizione strutture e risorse. A Trento lavorano alcune delle persone che hanno fornito il contributo maggiore al mio lavoro, primo fra tutti Paolo Bouquet, che l'ha seguito passo passo dalla proposta del progetto di ricerca al Collège Doctoral di Strasburgo fino alla redazione definitiva; a lui va un ringraziamento sentito per la pazienza e per i preziosi consigli. Insieme a lui vorrei ringraziare il Prof. Fausto Giunchiglia, che mi ha dato fiducia e mi ha permesso di lavorare con la sua logica, oltreché con le persone del suo gruppo di ricerca, tra cui vorrei ricordare Luciano Serafini, Chiara Ghidini, Stefano Zanobini, Matteo Bonifacio e Diego Ponte, i quali hanno sostanziato, con le loro osservazioni, parti importanti della tesi. Un altro ringraziamento va ad Achille Varzi, sperando di aver seguito degnamente il suo consiglio di "aprire molte porte, ma rimanendo sempre a camminare nel corridoio".

Sempre a Trento si trovano alcuni amici che vorrei ringraziare per avermi sopportato con pazienza, soprattutto nella parte finale del lavoro: tra questi

vorrei ricordare la mia compagna d'ufficio Roberta, le mie coinquiline Elisabetta e Anna Rita e Marco, che mi ha aiutato ad appianare le divergenze tra me e l'informatica.

Seconda tappa di questo viaggio è Strasburgo, dove ho trascorso un anno del mio dottorato e dove si trova l'istituzione che ha sovvenzionato questi tre anni di studio, il Collège Doctoral Européen. La prima persona a cui vanno i miei ringraziamenti è il Prof. André Thibault, mio direttore di tesi per parte francese, che ha accettato la sfida di dirigere una tesi in italiano e i cui commenti mi hanno spesso indotto a vedere le cose in una prospettiva diversa; vorrei inoltre ringraziare tutto lo staff del Collège, in particolare il Prof. Patrick Foulon, M.me Catherine Naud e M.me Béatrice Bader, sempre oltremodo disponibili. Vorrei infine ringraziare due persone che mi hanno praticamente "adottato" durante tutto il mio periodo strasburghese, aiutandomi a superare il momento di solitudine iniziale: Marie Pierre (che è stata anche la mia insegnante di francese e che mi ha aiutato nella redazione del résumé) e la mia compagna di viaggio Amaya.

Passiamo ora a Milano, che è in un certo senso il luogo da cui provengo, oltreché la seconda università, insieme alla Marc Bloch di Strasburgo, firmataria della cotutela. Qui vorrei ringraziare in primo luogo il Prof. Andrea Bonomi, mio tutor per parte italiana, i cui consigli hanno notevolmente influenzato la forma che ha assunto la tesi, soprattutto nella sua parte più filosofica. Assieme a lui vorrei ringraziare anche la Sig.ra Paola Maestri, che mi ha aiutato a districarmi con i problemi legati alla "doppia amministrazione" del mio dottorato.

Tappa finale del viaggio, l'università di Stanford, dove ho trascorso tre importanti mesi, grazie all'ospitalità del Prof. John Perry, e dove ho potuto seguire dei corsi che hanno contribuito moltissimo alla mia formazione, in particolare quelli tenuti dal Prof. Johan Van Benthem. Insieme a loro vorrei ringraziare Dikran Karagueuzian per la sua simpatia e per i suoi preziosi consigli e due amiche nonché attente lettrici delle bozze della mia tesi e – spero – future collaboratrici su molti progetti, Claudia Arrighi e Viola Schiaffonati.

Infine, un ringraziamento "ubiquitario" agli amici e amiche di sempre, che mi sono stati vicini nonostante questo mio continuo peregrinare. Cito

solo l'inizio di un elenco che in realtà è molto più lungo: Chiara, Edo, Gaia, Ivana, Lorella, Paola, Susanna, Viviana... a loro un grazie speciale.

Le raisonnement contrefactuel : un modèle et son application au raisonnement pratique. Résumé

Le sujet central de cette thèse est constitué par le raisonnement contrefactuel : à quoi sert-il, quand et pourquoi est-il utilisé ; quelles sont, en supposant qu'elles existent, sa signification et son utilité et, surtout, quelles sont les dimensions contrefactuelles que les agents rationnels développent ou peuvent développer au cours de procès de cognition et quel système théorique peut-il rendre compte de manière satisfaisante de ce phénomène si courant et aussi répandu.

Avant de commencer, il convient de clarifier ce que signifie contrefactuel et, plus particulièrement, ce que signifie le *raisonnement* contrefactuel.

Quelques définitions

L'*Oxford Companion for Philosophy* donne la définition suivante :

> A counterfactual is a conditional whose antecedent is false (typically, in philosophical practice, *known* to be false). The term is usually reserved for those (non-truth-functional) counterfactuals which are not true in virtue simply of their antecedent's falsity. Lawlike generalizations support counterfactuals: "Sugar dissolves in water" licenses "If this sugar cube were dropped in water

> it would dissolve"; but "All coins in my pocket are silver" does not yield "If this penny were in my pocket it would be silver"[1].
>
> [*The Oxford Companion for Philosophy*]

Il semblerait, donc, qu'il découle de cette définition que le terme *contrefactuel* se réfère à un phénomène lié à une forme grammaticale particulière, et plus spécifiquement, au conditionnel. Toutefois, il existe deux raisons pour lesquelles il n'est pas souhaitable de limiter l'analyse de la *contrefactualité* à la seule analyse de la logique des conditionnels.

La première raison tient au fait que la contrefactualité peut être exprimée de différentes façons qui n'exigent pas toutes l'utilisation du conditionnel, par exemple en conservant implicite la structure grammaticale de l'antécédent ou du conséquent, en l'occurrence:

En le sachant, je n'aurais pas été préoccupé (1)

J'ai raté le train, sinon j'aurais été ponctuel (2)

Si la coupole de Saint Pierre est mise à 1000 K, alors elle s'allumerait[2] (3)

Ni Hitler, ni bombe atomique.[3] (4)

Il est également possible d'exprimer la contrefactualité à travers un concept qui comprend non seulement la description de l'état dans lequel se trouve un certain objet, mais aussi la description de l'état dans lequel l'objet pourrait

---

[1]Un contrefactuel est un conditionnel dont l'antécédent est faux (spécifiquement dans la pratique de la philosophie, *retenue* faux). Le terme est habituellement réservé pour ces contrefactuels (non vrai-fonctionnels) qui ne sont pas vrais simplement en vertu du fait que l'antécédent est faux. Les contrefactuels se basent sur des généralisations en accord avec la loi: "Le sucre se dissout dans l'eau" autorise "Si ce morceau du sucre était mis dans l'eau, il se dissoudrait"; mais "Toutes les pièces contenues dans mes poches sont en argent" ne mène pas à la conclusion "Si ce penny était dans ma poche, il serait en argent". [*traduction de l'auteur*]

[2]Exemple pris de Dalla Chiara et Toraldo di Francia, en [34], p.68

[3]Exemple pris de D.K. Lewis en [99].

se trouver, étant donné les circonstances, comme le cas du prédicat dispositionnel. Un exemple en est l'adjectif "soluble", cité dans la définition de l'*Oxford Companion for Philosophy.*

Claudio Pizzi en [126] fournit une explication très claire de la raison par laquelle ces prédicats dispositionnels sont soutendus par une dimension contrefactuelle:

> In primo luogo, la presenza dei controfattuali nel linguaggio è poco apparente perché a volte usiamo costrutti linguistici in cui essi non compaiono in modo esplicito. Se dico che la zolletta che ho di fronte è solubile nel caffè ciò implica che se fosse stata messa nel caffè si sarebbe sciolta: dove l'ipotesi è certamente falsa in quanto, se si fosse verificata, non avrei nemmeno di fronte la zolletta di zucchero di cui sto parlando. Predicati come "solubile", "irascibile", "fragile" ecc. sono detti disposizionali perché descrivono la disposizione di un ente a reagire a determinati stimoli in circostanze possibili di qualche tipo[4].
>
> [Il Ragionamento Controfattuale, p.86]

La seconde raison – la plus importante – pour laquelle il est bon de ne pas s'arrêter à l'analyse linguistique est que, quelle que soit la forme avec laquelle ils sont exprimés, les contrefactuels représentent un type spécifique de forme de raisonnement et, comme tels, leur instances peuvent être combinées entre elles ou avec les instances d'autres formes de raisonnement, peuvent être itérées et "insérées" dans des processus de raisonnement plus vastes et plus complexes.

---

[4]Premièrement, la présence des contrefactuels dans le langage est peu apparent car nous utilisons parfois des constructions linguistiques dans lesquelles ils n'apparaissent pas de façon explicite. Si je dis que le cube de sucre que j'ai en face de moi est soluble dans le café, cela implique que s'il avait été mis dans le café, il se serait dissout : l'hypothèse est donc sûrement fausse car, si elle avait eu lieu, je n'aurais plus face à moi le cube de sucre dont je suis en train de parler. Les prédicats comme " soluble ", " irascible ", " fragile ", etc. sont dispositionnels car ils décrivent la disposition d'une entité à réagir à des stimuli dans certaines situations d'un type donné. [*traduction de l'auteur*]

Interprétée sous cette seconde acception, la contrefactualité se montre comme une dimension constituante de la rationalité humaine, incarnée par la capacité qu'ont les humains de faire abstraction de certains traits d'une situation qu'ils perçoivent comme réelle, d'imaginer des situations alternatives à celle-ci, de raisonner dans les confins de ces scénarios alternatifs en obtenant des informations qui relèvent de la situation réelle, mais qui ne pouvaient pas être directement inférées de celle-ci.

De façon similaire, John Pollock en [129] décrit ce qu'il entend par *raisonnement suppositionnel*, qui n'est rien d'autre que le raisonnement hypothétique, duquel le contrefactuel est une sub-partie spécifique :

> The employment of subsidiary arguments comprises *suppositional reasoning*, wherein we suppose something "for the sake of the argument", reason using the supposition in the same way we reason about beliefs and interests nonsuppositionally, and then on the basis of conclusions drawn using the supposition we draw further conclusions that do not depend upon the supposition. [...] Within the supposition, we reason as if the supposed propositions were beliefs, using all the rules for adoption and interest that were discussed in connection with linear reasoning[5].
>
> [Interest driven suppositional reasoning, p.427]

Les fonctions du raisonnement contrefactuel

Ainsi défini, le raisonnement contrefactuel est doué d'une série de fonctions aussi hétérogènes qu'importantes dans les inférences du sens commun nor-

[5] L'utilisation d'arguments subsidiaires comprend le *raisonnement suppositionnel*, dans lequel nous supposons quelque chose "par amour de la discussion" et nous raisonnons en utilisant la supposition de la même façon que nous raisonnons sur des croyances et intérêts non suppositionnels, et après, sur la base des conclusions obtenues avec la supposition, nous obtenons des conclusions qui ne dépendent pas de la supposition. [...] Dans la supposition, nous raisonnons comme si les propositions supposées sont des croyances, en utilisant toutes les règles pour l'adoption et l'intérêt qui ont été discutés en connexion avec le raisonnement linéaire. [*traduction de l'auteur*]

malement utilisées par les humains pour exécuter des devoirs qu'ils ont journellement à gérer. Ces fonctions appartiennent au domaine intellectuel, au domaine émotif ainsi qu'au domaine pratique.

Un exemple pris dans le domaine intellectuel est celui de la série de contre-exemples au raisonnement déductif, soit qu'ils soient entendus comme *reductio ad absurdum* dans un raisonnement formel, soit comme *falsification* pour un raisonnement empirique (scientifique ou non)[6] .

Un autre exemple intéressant est représenté par les formes de raisonnement ambiguës qui utilisent des métaphores, des double sens, de l'ironie et, en général, un raisonnement analogique, comme celles qui peuvent être appliquées aux contextes de fiction.

En ce qui concerne le domaine émotif, une importante littérature, théorique ou expérimentale en psychologie (cf. [114], [106], [149], [139], [86], mais surtout [115]) a soutenu l'hypothèse que le raisonnement contrefactuel accomplit des fonctions différentes selon qu'il institue une comparaison entre la réalité et un scénario meilleur (le soi-disant contrefactuel *upward*), ou entre la réalité et un scénario pire (contrefactuel *downward*).

Dans le cas *upward*, le raisonnement contrefactuel peut amener le repentir (quand le sujet perçoit qu'il n'a pas fait tout ce qu'il aurait dû faire pour rejoindre un objectif fixé), le regret (quand le sujet a fait quelque chose qui a porté préjudice à quelqu'un - soi-même ou un autre sujet - alors qu'en évitant d'exécuter l'action il aurait aussi évité le dommage).

Un autre effet du raisonnement contrefactuel *upward* est d'amplifier les émotions douloureuses quand le sujet comprend que, avec une petite modification du passé, sa situation actuelle serait vraiment meilleure. Le but suprême du contrefactuel *upward* serait alors de générer chez le sujet - au moment où celui-ci se trouve face à une situation analogue à celle sur laquelle

---

[6]Pizzi, en [126], dit: "Ragionamenti in cui si ipotizza qualcosa della cui verità non si è sicuri, o addirittura qualcosa della cui falsità si è sicuri, sono di uso corrente non solo nelle scienze formali ma anche nelle scienze empiriche e nella sfera del senso comune." ("Raisonnements dans lesquels on fait l'hypothèse qu'une chose dont on n'est pas sûr qu'elle soit vraie, ou, encore plus fort, dont on est sûr qu'elle soit fausse, sont communément utilisés non seulement dans les sciences formelles, mais également dans les sciences empiriques et dans le domaine du sens commun ") [*traduction de l'auteur*].

il a raisonné contrefactuellement - un rappel douloureux qui l'amène à éviter de répéter les erreurs du passé.

Les phrases suivantes donnent un exemple pour chacun des cas :

- **Contrefactuel *upward* exprimant le repentir** : "Si j'avais étudié davantage, j'aurais réussi mon examen ";

- **Contrefactuel *upward* exprimant le regret**: "Si je n'avais pas passé la fin de la semaine précédente en fêtes, j'aurais réussi mon examen";

- **Contrefactuel *upward* amplifiant les sensations douloureuses**: "Si j'étais arrivé ne serait-ce qu'une seconde avant, j'aurais pu prendre l'avion ".

Le contrefactuel *downward*, quant à lui, est source d'émotions positives telles que l'orgueil et la satisfaction ; c'est le cas lorsque, grâce à l'intervention de l'individu raisonnant, des événements négatifs, qui auraient pu arriver, sont évités ; une autre émotion positive est le soulagement quand on se trouve à un pas de la catastrophe mais que celle-ci, grâce à un détail apparemment insignifiant, se ne produit pas. Aussi dans ce cas, le but suprême devrait être de conduire l'individu à prendre des décisions plus adéquates grâce au rappel des émotions positives ressenties dans des situations similaires.

- **Contrefactuel *downward* exprimant l'orgueil** : "Si je n'avais pas étudié aussi fort, je n'aurais pas réussi mon examen ";

- **Contrefactuel *downward* exprimant le soulagement** : "Si j'étais resté sous l'arbre quelques instants de plus, la foudre m'aurait frappé".

Le troisième domaine, le pratique, comprend l'individualisation de sous-objectifs lorsque la démarche pour atteindre un objectif final est très complexe, la construction de schémas d'action alternatifs, le contrôle de plans formulés dans le passé et, par conséquent, la prévision de l'issue des plans futurs.

C'est essentiellement sur cette troisième dimension qu'est centrée notre étude et sa légitimation est établie par le fait que le raisonnement contrefactuel est considéré comme particulièrement important pour l'*agent* rationnel, plus encore que pour un *sujet* générique rationnel.

En d'autres termes, le raisonnement contrefactuel est particulièrement utile pour tous les processus cognitifs finalisés vers l'action, car il permet d'explorer différentes stratégies alternatives.

La *pensée* rationnelle acquiert un avantage énorme grâce au raisonnement contrefactuel car celui-ci rend explicite deux scénarios en même temps : le contrefactuel et le factuel (par contraste). Pour soutenir cette thèse, les psychologues Ruth Byrne et Alessandra Tasso ont conduit une étude empirique [29] dans laquelle, grâce à quatre expérimentations, elles ont montré le pouvoir d'explicitation des deux scénarios alternatifs que le raisonnement contrefactuel possède :

> Reasoners represent explicitly the case mentioned in the conditional, and they keep track of the possibility that there may be alternatives to it[7].
>
> [Deductive reasoning with factual, possible, and counterfactual conditionals, p.727]

De plus, les expérimentations semblent indiquer que le processus d'explicitation inhérent aux contrefactuels a des effets positifs sur la réalisation des devoirs assignés aux sujets.

> On this account, we can also make the further prediction that the initial understanding of a counterfactual is more difficult than the initial understanding of a factual conditional, because the counterfactual requires the construction of multiple models. Once this extra work is completed, however – as the results of these experi-

[7]Les sujets raisonnants représentent explicitement le cas mentionné par le conditionnel, et ils gardent trace de la possibilité qu'il y ait une alternatives à celui-ci. [*traduction de l'auteur*]

> ments have shown – it provides a richer basis for the subsequent tasks of deduction, verification, and falsification[8].
>
> [Deductive reasoning with factual, possible, and counterfactual conditionals, p.738]

Le système théorique que Byrne et Tasso utilisent pour rendre compte de résultats obtenus qui ne soient pas – du moins dans les intentions – formels, semble cependant inadapté pour le discours global que l'on veut tirer de cette thèse ; les évidences empiriques semblent conforter l'intuition, que nous partageons avec Byrne et Tasso, selon laquelle le contrefactuel fournit une valeur ajoutée spécifique au bagage cognitif d'un sujet rationnel.

En ce qui concerne l'*action* rationnelle, l'utilité du raisonnement contrefactuel découle de l'éventuelle possibilité ou – plus fréquemment – de l'impossibilité de vérifier dans la réalité les effets d'une action ; c'est le cas pour le pilote obligé de juger dans la réalité (plutôt qu'avec des hypothèses contrefactuelles) que toutes les manoeuvres de vol sont correctes ; c'est également la situation du savant obligé de vérifier " dans la réalité " la validité d'une loi scientifique qui fait abstraction de tout frottement (comme la loi d'inertie).

Toutefois, ce type d'avantage relatif à l'action, n'est pas l'apanage du raisonnement contrefactuel ; il appartient aussi à d'autres types de raisonnement hypothétique, par exemple au raisonnement hypothétique de la possibilité. En d'autres termes, le même pilote peut raisonner contrefactuellement : " Si je n'avais pas viré, j'aurais percuté les câbles de l'haute tension ", mais il peut également raisonner sur la possibilité : " Si je ne vire pas maintenant, je percuterai les câbles de l'haute tension ".

Quel avantage peut offrir le contrefactuel en comparaison du raisonnement hypothétique de la possibilité ? L'avantage d'avoir un point fixe, c'est-à-dire de savoir comment les choses se sont passées réellement.

---

[8]Sous cet aspect, nous pouvons aussi faire la prédiction ultérieure que la compréhension initiale d'un contrefactuel est plus difficile que la compréhension initiale d'un conditionnel factuel, car le contrefactuel demande la construction de plusieurs modèles. Quand ce travail préparatoire est réalisé, toutefois – ainsi que les résultats de ces expérimentations l'ont montré – il donne une base plus riche pour les actions de déduction, de vérification et de falsification qui lui succèdent. [*traduction de l'auteur*]

Pour reprendre l'exemple précédent, le pilote *sait* qu'en virant il a évité les câbles de haute tension et il peut utiliser cette information (ainsi que d'autres informations qu'il peut déduire de celle-ci) pour raisonner sur l'hypothèse contrefactuelle ; dans le second cas, l'agent peut seulement *croire* qu'avec cette manoeuvre il va éviter l'obstacle, mais il ne peut pas en être sûr[9].

Domaines d'application possibles pour le contrefactuel

Bien que l'étude des contrefactuels puisse apparaître au premier regard comme un simple exercice intellectuel, il semble que son utilité commence à être largement reconnue, tant il a acquis de légitimité dans un nombre toujours plus grand de disciplines pour lesquelles il est devenu un instrument fondamental.

Il existe par exemple une série d'études qui conjuguent jurisprudence et psychologie criminelle (cfr. [149], [33], [161], [118]), dans lesquelles le contrefactuel est utilisé pour évaluer les circonstances atténuantes ou aggravantes d'un délit ou pour comprendre si l'action de l'accusé est la véritable cause du dommage souffert par la victime (ce qui dans le jargon juridique est défini comme *conditio sine qua non*).

En économie, l'usage des contrefactuels a été essentiellement appliqué dans deux branches spécifiques : la théorie des décisions ([56], [119]) et la théorie des jeux ([15], [14], [13], [151], [80]), le contrefactuel y est devenu un puissant instrument heuristique, en particulier dans les situations d'information imparfaite.

En ce qui concerne la psychologie, en plus des travaux déjà signalés qui étudient les contrefactuels et les situations générant différents types de contrefactuels, il existe des études dans lesquelles les contrefactuels sont utilisés

[9]Quand nous disons que l'agent *sait* ce qui s'est passé, nous n'entendons pas que l'agent " croit que A et A est vrai ", comme on l'a souvent soutenu dans le cadre de la " tradition " philosophique, mais plutôt que, dans la théorie que l'agent utilise pour raisonner sur celle qu'il croit être la réalité A est vrai ; au contraire, dans le cas hypothétique de la possibilité, A ne prend pas une valeur de vérité définie dans la même théorie.

beaucoup plus comme instruments d'analyse que comme objets d'analyse ; c'est le cas de l'étude de psychoses telles que l'autisme, dans lequel, selon la lecture qui identifie en partie le syndrome autistique avec l'incapacité du sujet à élaborer une théorie du mental (cf. par exemple, [4] et surtout [30], qui recueille une longue série d'articles sur le sujet), le raisonnement contrefactuel peut être interprété comme un des moyens dont dispose l'individu pour élaborer une théorie du mental en direction des autres.

Dans le domaine de l'intelligence artificielle, il y a eu des études dirigées vers l'application du contrefactuel dans le *planning* et dans le diagnostic des erreurs (à ce propos cf. surtout l'avant-gardiste [67] ; d'intéressantes remarques existent aussi dans [87] et [41]).

A coté de ces développements qui sont pour la plus part académiques, il y a un courant, particulièrement fécond lors des dernières années, dans lequel recherche scientifique et production littéraire se confondent et créent une soi-disant " histoire virtuelle ". D'un côté, on étudie l'importance historique de certains événements, en imaginant des issues variées pour des batailles décisives ou des attitudes stratégiques différentes pour des généraux et des condottieres (c'est le cas d'ouvrages comme [157], [55] et [43]), de l'autre côté, on crée des récits et romans fantaisistes à partir de l'altération d'un fait historique, dans un cadre historiographique fidèle à la réalité – ou présumé tel (comme en [158], [148] et [84]). Pour saisir comment les reconstructions historiques soignées et les créations purement fantaisistes se mêlent dans ce genre de littérature, on peut se reporter à une partie de l'introduction de l'anthologie d'essais réalisée par Robert Cowley [43]:

> Et si une épidémie mystérieuse n'avait pas frappé les assaillants assyriens de Jérusalem en 701 avant J.C., aurait-on eu une religion hébraïque ? Ou le Christianisme ? Prenons des faits d'une durée de l'ordre de la fraction de seconde : Qu'est-ce qui se serait passé si la trajectoire d'une hache de guerre n'avait pas été interrompue et si Alexandre, âgé de 21 ans, avait été tué avant de devenir " Magnus " ? Ou si Cortés, qui fut quasiment capturé pendant le siège de Tenochtitlán, l'actuelle Mexico City, avait réellement été fait prisonnier ? Il est très probable que les jeunes États-Unis auraient trouvé un grand Empire indigène américain

> sur leurs confins méridionaux. Essayons aussi de considérer le rôle du hasard : si, dans la guerre civile américaine, le célèbre " ordre perdu " n'avait pas été vraiment perdu, il est probable que, comme James M. McPherson l'a écrit, les États Confédérés seraient restés indépendants. Un " ordre perdu " analogue a influencé l'issue de la bataille de la Marne en Septembre 1914 et, en conséquence, la Première Guerre Mondiale elle-même.
>
> [La storia fatta con i se, *traduction de l'auteur*]

Le fonctionnement du raisonnement contrefactuel

Tous les cas décrits jusqu'ici ont en commun un scénario centré sur un problème à résoudre et sur une série de solutions envisageables entre lesquelles on peut choisir. Le raisonnement contrefactuel devient alors un instrument pour analyser et juger la qualité des choix effectués. Cette dimension *pragmatique* du raisonnement contrefactuel est justement celle que nous voulons approfondir dans ce travail ; la dimension pragmatique doit caractériser notre description théorique non seulement pour rechercher comment fonctionne celui-ci, mais aussi comment il est intrinsèquement constitué.

Du côté du fonctionnement du raisonnement contrefactuel, notre hypothèse de travail est qu'il s'oriente vers deux directions, identifiées par deux formes de rationalité :

- **La rationalité moyens-buts** : à partir d'un ensemble d'assomptions et de préférences considérées comme fixes et immuables, elle détermine un objectif et applique sa fonction critique-dialectique aux différents moyens disponibles pour l'atteindre ;

- **La rationalité ex-post** : à partir d'un ensemble de moyens (capacités, ressources), considérés comme fixes et immuables, elle soumet à la critique dialectique les préférences/assomptions pour déterminer un objectif que l'on peut atteindre à partir des moyens à disposition.

Intuitivement, le premier type de rationalité semble réductible à un processus de révision des éléments *dans* une théorie, au contraire le second semble plutôt correspondre à la prise d'une nouvelle perspective, c'est-à-dire à l'attitude qui consiste à revoir le même problème à la lumière d'une théorie *différente*.

C'est pourquoi le système à adopter pour traiter le raisonnement contrefactuel doit être capable de représenter en même temps les opérations qui se passent *dans* les théories et celles qui se passent *entre* les théories, capable d'expliquer conjointement le caractère circonscrit de certains raisonnements et l'importance des relations qui existent entre les processus de raisonnement conduits sur la base d'assomptions différentes.

Une théorie contextuelle pour le raisonnement contrefactuel

L'idée de laquelle nous partons pour décrire notre théorie est que cette théorie devrait rendre compte le plus clairement possible de la façon dont un agent rationnel élabore un processus de raisonnement contrefactuel[10].

A notre avis, l'idéation d'une hypothèse contrefactuelle (et du raisonnement conséquent) est un processus de pensée qui en présuppose un autre : celui de la sélection de l'information que l'agent juge adéquat pour raisonner sur l'argument spécifique.

Ce processus de sélection est central et il est aussi crucial pour la détermination de l'issue du raisonnement contrefactuel car, selon les informations et les règles que l'agent *décide* d'utiliser, le processus cité donne des résultats différents. Nous considérons donc cette caractéristique comme une chose à laquelle on ne peut pas renoncer car elle permet d'expliquer d'une part le fait que des agents différents peuvent élaborer des raisonnements contrefactuels avec des résultats contraires en considérant le même problème et, d'autre part, elle rend compte du fait que le même agent, quand il a con-

[10]Une formulation préliminaire des intuitions qui nous ont conduit à la décision de traiter le raisonnement contrefactuel dans une perspective contextuelle est contenue en [57].

naissance de nouvelles informations, peut modifier largement ses processus de raisonnement, jusqu'à en fausser les résultats.

En outre, c'est toujours de ce processus de sélection que dépend le choix entre ce qui est factuel et ce qui est contrefactuel car, par exemple, si un agent se trompait sur le véritable état des choses, il pourrait considérer comme factuel ce qu'un " observateur externe " jugerait comme contrefactuel et vice-versa ; mais le effets pratiques de son raisonnement (par exemple, les actions qu'il ferait à partir de ce raisonnement), seraient toujours liés à son interprétation et non plus à la réalité observée par l'autre observateur (un exemple qui montre ce phénomène est décrit dans la section 4.6).

Cette relativité et cette flexibilité des concepts de factuel et contrefactuel sont très importantes pour l'étude des théories scientifiques, en raison de leur nature provisoire et de la fréquence des situations dans lesquelles les savants provenant de différentes communautés scientifiques attribuent des signifiés différents aux mêmes événements et où certains voient une variable là où les autres envisagent une constante et vice-versa. De quelque manière que ce soit, lorsqu'une théorie scientifique est consolidée dans une communauté, elle est traitée *comme si* elle décrivait les faits (c'est-à-dire comme si elle était factuelle) et les autres théories sont perçues comme contrefactuelles ; toutefois, cette perspective peut être changée à tout moments.

En d'autres termes, un agent possède, à notre avis, une base de connaissance très grande et articulée[11] mais qui n'est pas immédiatement disponible dans sa totalité lorsqu'il commence à résoudre un problème.

Cela résulte de plusieurs causes ; et prioritairement pour des causes " économiques " : si un agent devait prendre en considération tout ce qu'il sait avant de pouvoir formuler un plan ou établir une stratégie, les processus décisionnels seraient beaucoup plus lents et dispendieux. De plus, la base de connaissance d'un agent pourrait contenir des informations contradictoires qui auraient été apprises dans des circonstances variées. Nous voulons, au

[11]Compte tenu des objectifs de ce travail, il n'est pas indispensable de distinguer entre une base de connaissance énorme et indistincte et une base partagée en sous-domaines structurés entre eux ; si nous préférons la seconde thèse, c'est parce que cette image est plus proche de l'idée que nous soutenons, selon laquelle un agent situé face à un problème à résoudre, active chaque fois une portion spécifique de ses connaissances.

contraire, rendre compte du fait que l'agent peut choisir de considérer comme vrai un fait en raisonnant dans un contexte et comme vraie la négation du même fait en raisonnant dans un contexte différent[12]. En outre, comme on l'a déjà signalé, des agents différents peuvent avoir des perspectives divergentes relativement à un problème pris isolément, sans que cela conduise à tomber dans le relativisme absolu, car c'est la même sélection de l'information servant à identifier le contexte factuel duquel on part qui détermine les contraintes que le raisonnement contrefactuel doit satisfaire et la cohérence du raisonnement contrefactuel chez l'agent est donc subordonné à sa capacité d'identifier l'information appropriée pour raisonner sur un problème spécifique[13].

Comme Tetlock et Belkin remarquent en [156]:

> Different investigators will inevitably emphasize somewhat different criteria in judging the legitimacy, plausibility, and insightfulness of specific counterfactuals. It would be a big mistake, however, to confuse epistemic pluralism (which we accept up to a point) with an anything-goes subjectivism (which we reject and which would treat all counterfactual claims as equally valid in their own way)[14].
> [Counterfactual Thought Experiments in World Politics]

[12]Un exemple peut être utile pour préciser ce point. Supposons qu'un agent croit que la loi de Newton sur la gravitation universelle soit fausse d'après la théorie einsteinienne de la relativité ; une assertion qui découlerait de cette loi de Newton serait considérée par l'agent comme contrefactuelle par rapport à un problème géré par la relativité einsteinienne. Le même agent pourrait toujours juger vraie (et la fixer comme telle dans l'hypothèse du raisonnement) la loi de gravitation universelle newtonienne en résolvant un problème de physique à l'échelle " terrestre ". Dans ce contexte, les assertions qui dérivent de la loi de Newton seraient considérées, chez cet agent, comme " factuelles ".

[13]Nous ne pouvons pas exclure que notre position puisse être jugée, selon la citation de Tetlock et Belkin, comme " sujetivisme du tout va bien ", car pour elle le processus de raisonnement d'un psychotique qui raisonnerait à partir de ces fantaisies serait tout à fait légitime, si celles-ci gardaient leur cohérence interne. L'intérêt premier de notre théorie est qu'elle décrit une procédure de raisonnement correcte, au-delà de l'adéquation des prémisses desquelles elle part.

[14]Des investigateurs différents souligneront inévitablement les critères de façon diverse en jugeant la légitimité, la plausibilité et la capacité d'approfondissement de contrefactuels spécifiques. Ce serait vraiment une erreur grossière que de confondre le pluralisme

Pour terminer, s'il était vrai que, au moment de prendre une décision, un agent a à sa disposition tout ce qu'il sait, il serait difficile d'expliquer le fait que les agents commettent souvent des erreurs très banales, même dans des domaines où ils sont experts.

Pour toutes ces raisons, l'hypothèse que la base de connaissance dans sa totalité soit le point de départ des raisonnements nous semble inadaptée. Au contraire, nous pensons qu'il est plus sensé de soutenir que les agents, pour raisonner sur des problèmes spécifiques, " découpent " une portion de cette base de connaissance et l'utilisent pour construire la théorie partielle dont ils useront pour raisonner sur le problème, c'est-à-dire ce qu'on appelle, selon la définition fournie en [20], contexte ou théorie de travail (*working context*).

Cette opération est préliminaire par rapport à celle que le sujet réalise quand il formule une hypothèse contrefactuelle et développe, à partir de celle-ci, un raisonnement également contrefactuel.

Dans notre modèle, c'est à partir du choix initial des axiomes à utiliser dans le raisonnement spécifique que le contexte factuel et le contexte contrefactuel émergent, grâce au fait qu'ils sont situés dans une certaine relation de compatibilité (que nous appelons dans ce cas spécifique *relation de contrefactualité*). Selon les différents systèmes d'axiomes qui peuvent être assumés chez un agent, on déterminera différents couples de *contextes factuels* et *contextes contrefactuels* relatifs au même problème.

Au niveau intuitif, confronté avec un énoncé contrefactuel auquel il doit assigner une valeur de vérité, l'agent doit :

1. décider prioritairement, quelles sont les lois générales qu'il peut utiliser pour raisonner sur le problème qu'il doit résoudre ;
2. vérifier ce que se passe dans la " réalité " ;
3. poser une hypothèse contrefactuelle ;
4. déduire les conséquences de cette hypothèse à partir des lois générales qu'il a sélectionnées ;

épistémique (que nous acceptons jusqu'à un certain point) avec un sujetivisme du type " tout va bien " (que nous rejetons et qui traiterait les assertions contrefactuelles comme tout à fait valides à leur manière). [*traduction de l'auteur*]

5. estimer la valeur de vérité de l'énoncé contrefactuel.

Ces étapes, au niveau logique, correspondent aux opérations suivantes :

1. construire l'ensemble de tous les modèles possibles de la situation (à travers toutes les combinaisons envisageables des termes du langage, parmi lesquelles on éliminera toutes celles qui contredisent les axiomes) ;

2. construire le contexte factuel en choisissant, dans les interprétations restantes, celles où l'antécédent et le conséquent de l'énoncé contrefactuel sont tous les deux faux ;

3. construire le contexte contrefactuel, en choisissant, parmi les interprétations possibles, celles où l'antécédent de l'énoncé contrefactuel est vrai ;

4. vérifier la valeur de vérité du conséquent de ces interprétations : si dans toutes les interprétations restantes elle est vraie, alors l'énoncé contrefactuel sur lequel on était en train de raisonner est valable ; si dans toutes les interprétations elle est fausse, alors l'énoncé semifactuel correspondant est valable ; si, au contraire, dans certaines interprétations elle est vraie et dans autres fausse, alors la valeur de vérité de l'énoncé n'est pas déterminable à partir de ces axiomes, c'est-à-dire, dans cette théorie.

La conséquence la plus importante dérivant du fait d'assumer cette approche est que, sur une telle base, il n'est pas possible d'affirmer catégoriquement la vérité ou la fausseté d'un énoncé contrefactuel, parce que sur le même couple antécédent-conséquent on peut construire un nombre n, pouvant être infini, de couples de contextes factuels-contrefactuels, chacun d'entre eux étant individualisé à partir d'une relation de contrefactualité différente.

Ce scénario détermine aussi une conception différente de ce qu'est le " factuel " : ce n'est plus ce qui est vrai dans le monde réel (ou, en tous cas, dans un monde possible), mais ce que est vrai dans une théorie, c'est-à-dire dans tous les modèles locaux d'un contexte individualisé par une relation de compatibilité spécifique.

Cet approche évoque un courant de la philosophie de la science des années 1900 ([82], [58], [91], [47], [127]), qui insiste sur l'impossibilité absolue d'affirmer la vérité ou la fausseté d'un énoncé scientifique sans le replacer dans la tradition scientifique spécifique ou, encore mieux, dans le paradigme théorique particulier à partir duquel il a été généré.

Cette approche peut également être mise en rapport avec le précepte linguistique que l'on pourrait qualifier d'*holisme* et qui est largement partagé par notre notions de sens commun, selon lequel un mot gagne en signification uniquement dans le cadre d'un discours.

Le " cas scientifique " et le " cas linguistique " semblent tous deux souligner que la vérité et la fausseté des énoncés ou, en dernière analyse, leur signification, dépendent d'une série de règles qui ne sont pas générales et données une fois pour toutes, mais qui sont relatives à un domaine spécifique et c'est précisément ce domaine qui doit fournir une structure interprétative.

Le point de vue privilégié n'est donc plus celui de la réalité métaphysique, mais celui de la perspective cognitive particulière de l'agent rationnel, qui ne fait plus des opérations sur des énoncés dans une théorie qui lui est donnée au départ, mais, au contraire, qui agit directement sur les théories qu'il construit au fur et à mesure expressément pour résoudre les problèmes spécifiques. Ceci serait confirmé à la fois par la vélocité avec laquelle les agents font certains raisonnements, rapidité difficilement explicable si l'on considère que ces agents peuvent prendre en considération toute l'information qu'ils ont à leur disposition, et à la fois par le fait qu'ils arrivent souvent à des conclusions différentes quand ils obtiennent de nouvelles informations et, d'une certaine façon, construisent une théorie nouvelle.

Les différences entre ces deux types d'approche et les conséquences qui en découlent pour la représentation de la connaissance et pour les processus de raisonnement sont bien expliquées en [19].

Un autre avantage du formalisme que nous avons choisi d'utiliser est celui de rendre beaucoup plus facile la représentation du raisonnement, car celui-ci se déroule dans le contexte, qui est un *ensemble* de modèles, donc un objet *partiel*, qui n'assigne pas une valeur de vérité à tous les termes du langage qui le caractérisent, mais seulement à ceux desquels on veut parler dans la théorie. Le processus de raisonnement dans ce cas met uniquement en jeu

une partie limitée de l'information disponible et ceci rend les opérations plus rapides et faciles.

Pour mieux comprendre les différentes manières de passer de la sémantique des modèles locaux (SML) par rapport, par exemple, aux théories basées sur des objets complets tels que ceux de la logique modale avec les mondes possibles ; on peut citer un exemple devenu classique, celui proposé par Kit Fine relatif à Nixon et à l'holocauste.

Au cours d'une période de crise internationale, Nixon est assis dans la fameuse " chambre des boutons " ; nous savons que, heureusement, Nixon n'a pas appuyé sur le bouton, mais Fine demande, " si Nixon avait appuyé sur le bouton, il y aurait eu l'holocauste nucléaire " est-il un énoncé vrai ou faux ?

Pour répondre à cette question, les théoriciens des mondes possibles doivent comparer le monde duquel ils partent (par simplicité, on considère que celui-ci est le monde réel) et les différents mondes alternatifs. Pour cela, ils ont besoin d'assigner une valeur de vérité à tous les énoncés du langage concernant chaque monde possible. Une fois que cette assignation a été faite, il faut ordonner les mondes selon leur ressemblance avec le monde duquel on est parti.

Ici apparaît le problème remarqué par Fine : dans le cas de Nixon, les théoriciens des mondes possibles (en particulier Lewis, contre lequel Fine dirige sa critique) affirmeront que l'énoncé contrefactuel ci-dessus est vrai, mais est-ce que cela signifie qu'un monde dans lequel il y aurait l'holocauste serait *plus similaire* au monde réel qu'un monde dans lequel il y aurait une panne du circuit électrique et où il n'y aurait pas d'holocauste ?

Maintenant on néglige dans ce contexte la solution proposée par Lewis, qui implique des miracles, petits et grands, et on se concentre sur la nature des différentes réponses fournies par la SML.

En premier lieu, on construit dans la SML un contexte contenant tous les facteurs que l'agent retient comme nécessaires pour le raisonnement qu'il doit entreprendre et les axiomes qu'il entend utiliser : ce contexte consistera en un ensemble de toutes les interprétations possibles (résultant de la combinaison des termes expressément sélectionnés pour raisonner sur le problème) qui respectent les axiomes et les contraintes imposées par l'énoncé contre-

factuel sur lequel on est en train de raisonner (dans le cas où l'antécédent et le conséquent seraient faux). De la même façon, on construit le contexte contrefactuel qui contient les interprétations qui respectent les axiomes et les contraintes (dans le cas où l'antécédent de l'énoncé contrefactuel serait vrai). L'étape suivante consiste à vérifier quelle est la valeur de vérité du conséquent du contrefactuel dans toutes les interprétations contenues dans le contexte contrefactuel.

La solution qui a été fournie au problème de Fine est essentiellement différente de celle proposée par les théoriciens des mondes possibles : si la théorie de départ " dit quelque chose " de la panne du circuit électrique, la valeur de vérité de l'énoncé contrefactuel dépendra aussi du fait que la panne ait lieu ou non dans le contexte contrefactuel, autrement cela n'aura aucune influence. Si, par ailleurs, dans le contexte factuel on parle d'une panne électrique qui a effectivement eu lieu, mais que, dans le contexte contrefactuel il n'y a aucune hypothèse expressément soutenue relative à la panne, au contraire des sémantiques à mondes possibles on serait amené à considérer des mondes possibles dans lesquels la panne aurait eu lieu et aussi ceux dans lesquels ça se ne serait pas passé, car tous sont compatibles avec le raisonnement contrefactuel en objet.

A ce point, la question de la ressemblance disparaît complètement car, à travers le choix de la théorie sur laquelle on veut raisonner, on choisit aussi les facteurs qui doivent être pris en considération dans le raisonnement et, une fois que le choix a été fait, il n'y a pas d'ordonnancements à établir, il faut seulement aller regarder ce qui se passe dans les interprétations satisfaisant à certaines contraintes ; en d'autres termes, on n'établit aucune hiérarchie entre les modèles locaux (contrairement à ce que se passait avec les mondes possibles), on prend *tous* ceux et *seulement* ceux qui satisfont aux contraintes posées par la relation de compatibilité.

Ainsi, au-delà des éléments qu'on a sélectionnés et sur lesquels on est en train de raisonner, tout le reste peut être pareil ou différer énormément de la situation de départ : au de-là de ces éléments, les interprétations les plus similaires et celles qui divergent le plus ont le même poids dans la détermination de la valeur de vérité de l'énoncé contrefactuel, lequel est vrai ou faux *dans une théorie*.

En conclusion, ce sont deux aspects principaux – strictement liés entre eux – qui déterminent le choix pour la SML.

La première caractéristique est de permettre d'établir la valeur de vérité ou de fausseté d'un contrefactuel de façon différente selon la théorie qui sert de base pour effectuer l'évaluation. Cette caractéristique entraîne comme corollaire, la capacité d'exprimer la non-unicité du raisonnement contrefactuel, c'est-à-dire le fait que, à la différence du conditionnel matériel, le contrefactuel peut changer de valeur de vérité en ajoutant une prémisse[15]. Ce phénomène, dans la SML, correspond à la construction d'un contexte factuel différent et, conséquemment, d'un contexte contrefactuel également différent.

La seconde particularité est relative à la description de la façon dont se passe l'assignation de la valeur de vérité en énoncé contrefactuel, laquelle a lieu à travers une procédure aussi simple que faillible, mais qui n'oblige pas l'agent à considérer toutes les caractéristiques de la situation qu'il est en train de vivre et des situations dans lesquelles il aurait pu se retrouver, mais seulement les caractéristiques des situations sélectionnées sur la base de la théorie qu'on est en train d'utiliser.

Une formalisation du raisonnement contrefactuel basée sur la SML

L'un des buts principaux de ce travail est de trouver une systématisation formelle qui soit apte à représenter les caractéristiques du raisonnement contrefactuel que nous avons remarquées.

L'intuition de fond qui nous a conduits à ce choix est que le raisonnement contrefactuel est un type particulier de raisonnement contextuel, c'est-à-dire un raisonnement qui a lieu *dans* et *à travers* des domaines circonscrits. Si ces domaines, comme on l'a expliqué en [10], sont décrits comme étant caractérisés par trois propriétés : la partialité, l'approximation et la perspec-

[15]Comme dans l'exemple ci-dessus: " Si Nixon avait appuyé sur le bouton, il y aurait eu l'holocauste " peut être vrai et, en même temps, " Si Nixon avait appuyé sur le bouton et il s'il y ait eu une panne dans le circuit électrique, il y aurait eu l'holocauste " peut être fausse.

tive, les relations qui existent entre eux peuvent, en partie, être décrites comme des variations du niveau de partialité, du degré d'approximation ou de perspective.

Le raisonnement contextuel et, avec lui, le raisonnement contrefactuel peuvent être décrits et représentés en termes d'opérations *sur* et *entre* des domaines partiels, approximatifs et prospectifs.

À partir de cette idée, nous avons choisi d'utiliser, pour la construction du système formel, un système logique (et la sémantique qui s'y rapporte) ; ce système a été pensé dans le but de représenter le raisonnement contextuel et il s'est montré très efficace pour résoudre toute une série de problèmes qui apparaissent lors de l'étude de cette forme de raisonnement. Ce système logique prend le nom de Systèmes MultiContexte (ou *MultiContext Systems*) et la sémantique correspondante est la Sémantique à Modèles Locaux (ou *Local Model Semantics*).

Les Systèmes MultiContexte sont composés par théories (les contextes) liés l'un à l'autre par des types de liens particuliers. Du point de vue syntaxique, les contextes sont des théories caractérisées pour chacune d'entre elles par un langage, un ensemble d'axiomes et un ensemble de règles d'inférence à eux-mêmes. Le lien entre un contexte et un autre s'établit grâce à la présence de règles qui permettent d'importer et d'exporter l'information ; ces règles sont dites *règles pont*(ou *bridge rules*).

Sur le plan sémantique, un contexte est un ensemble de modèles locaux (un modèle local est un modèle classique à la Tarski), qui entretiennent entre eux des relations appelées *relations de compatibilité*, qui rendent explicite le type de contraintes qui doit exister entre deux contextes pour qu'ils puissent être déclarés compatibles selon la notion spécifique de compatibilité que l'on formalise à ce moment là.

Dans notre travail, nous avons essayé de fournir une série de définitions sémantiques pour le raisonnement contrefactuel en construisant un Système MultiContexte spécifique composé de couples de contextes factuel/contrefactuel, dans lesquels la relation existant entre ces couples est une relation de compatibilité spécifique, que nous avons qualifiée de *relation de contrefactualité*, construite en utilisant des contraintes spécifiques que ces contextes doivent satisfaire pour pouvoir être définis respectivement comme

factuel et contrefactuel.

Un raisonnement contrefactuel est donc un processus qui a lieu entre deux contextes déterminés par deux faits (qui sont l'antécédent et le conséquent de l'énoncé contrefactuel), qui sont tous les deux faux dans un cas (dans le contexte factuel) et dont l'un au moins est vrai dans l'autre cas (l'antécédent dans le contexte contrefactuel). Un processus d'inférence a lieu dans le contexte contrefactuel et la conclusion de ce processus est importée dans le contexte factuel et exprimée à travers un prédicat relatif aux deux faits (le prédicat dit que l'antécédent et le conséquent ont une relation contrefactuelle).

Naturellement, le même contexte factuel peut être connexe à plusieurs contextes contrefactuels à travers des relations de contrefactualité différentes et le même rapport factuel/contrefactuel existant entre deux contextes peut être inversé s'il est établi par une relation de contrefactualité différente et ces deux propriétés rendent le système fortement différent de ceux qui constituent la vision standard en philosophie.

Structure du travail

La thèse est structurée en trois parties ; une première partie dans laquelle on introduit le formalisme qui a été utilisé pour décrire la représentation et le fonctionnement du raisonnement contrefactuel et dans laquelle on fait la comparaison avec d'autres formalismes ou développements intuitifs fournis précédemment ; dans la seconde partie, le raisonnement contrefactuel est appliqué à certaines dimensions du raisonnement pratique, avec le but de montrer comme celui-ci peut constituer un instrument de raisonnement très efficace ; enfin, dans la troisième partie on trace les lignes générales des développements futurs possibles, essentiellement en direction des applications.

La première partie montre le chemin qui nous a conduits à l'élaboration de notre système formel en partant de l'analyse qui a été réalisée, surtout dans les trente dernières années, dans le domaine de la philosophie du langage, finalisée dans l'étude de la sémantique des conditionnels contrefactuels (chapitre 2), en passant en revue les théories, développées en premier dans

le domaine de la psychologie cognitive et de l'intelligence artificielle et qui interprètent le contrefactuel comme un phénomène de raisonnement mais, dans le premier cas, sans fournir aucun modèle formel et, dans le second cas, en fournissant des modèles qui n'utilisent que certaines des propriétés que nous considérons comme caractéristiques de ce type de raisonnement (chapitre 3). Le chapitre 4 présente le système formel que nous avons développé en utilisant la Sémantique à Modèles Locaux, une logique pour le raisonnement contextuel qui est déjà bien rodée dans la résolution de problèmes spécifiques, émergeant surtout dans la sphère de l'intelligence artificielle.

La deuxième partie, quant à elle, est structurée en trois chapitres : le chapitre 5 introduit les concepts-clé du raisonnement pratique et fournit les fondements de notre analyse, qui détermine une ligne directrice relative au problème de la rationalité et qui peut être suivie selon deux sens opposés : des préférences aux moyens à se procurer pour rejoindre un objectif et des moyens disponibles à de nouvelles préférences, qui déterminent la formation de nouveaux objectifs ; les deux chapitres suivants, le 6 et le 7 montrent comment le raisonnement contrefactuel peut s'appliquer aux deux formes de rationalité.

Enfin, la troisième partie, consacrée aux développements futurs, introduit brièvement des domaines d'application du raisonnement contrefactuel actuellement envisageables, tels que la rationalité scientifique, les agents artificiels et les scénarios multi-agents et elle fournit une description préliminaire d'une manière possible pour gérer ces domaines à partir du *framework* déterminé auparavant.

Capitolo 1

Introduzione: condizionali controfattuali e ragionamento controfattuale

"Contrariwise" continued Tweedledee, "If it was so, it might be; and if it were so, it would be; but as it isn't, it ain't. That's logic."[1]

[Lewis Carroll]

Il tema centrale di questa tesi è il ragionamento controfattuale: a cosa serve, quando e perché gli esseri umani lo utilizzano, quali sono (se esistono) il suo senso e la sua utilità ma, soprattutto, quali processi cognitivi che gli agenti razionali svolgono o possono svolgere sono caratterizzati da una dimensione controfattuale e quale sistema teorico può rendere conto in maniera soddisfacente di questo diffuso e pervasivo fenomeno.

Prima di cominciare, è utile fare chiarezza su cosa si intende per controfattuale e, più precisamente, per *ragionamento* controfattuale.

[1] "Al contrario" continuò Tweedledee, "Se così fosse, potrebbe essere; e se così fosse, sarebbe; ma, siccome non è, non è. Questa è logica." [*traduzione mia*]

Qualche definizione

Citiamo di seguito la definizione fornita dall'*Oxford Companion for Philosophy*:

> A counterfactual is a conditional whose antecedent is false (typically, in philosophical practice, *known* to be false). The term is usually reserved for those (non-truth-functional) counterfactuals which are *not* true in virtue simply of their antecedent's falsity. Lawlike generalizations support counterfactuals: 'Sugar dissolves in water' licenses 'If this sugar cube were dropped in water it would dissolve'; but 'All coins in my pocket are silver' does not yield 'If this penny were in my pocket it would be silver'[2].
>
> [*The Oxford Companion for Philosophy*]

Da questa definizione sembrerebbe quindi potersi inferire che per *controfattuale* si intende un fenomeno legato a una certa forma verbale, nella fattispecie, condizionale. Tuttavia, esistono due buone ragioni per le quali è sensato non limitare l'analisi della *controfattualità* all'analisi dei condizionali.

La prima è che esistono vari modi di esprimere la controfattualità che non richiedono di esplicitare per intero il condizionale, per esempio mantenendo implicita la struttura grammaticale dell'antecedente o del conseguente, come in:

Averlo saputo, non mi sarei preoccupato (1.1)

Ho perso il treno, altrimenti sarei arrivato puntuale (1.2)

---

[2]Un controfattuale è un condizionale il cui antecedente è falso (tipicamente, nella pratica filosofica, *ritenuto* falso). Il termine è abitualmente riservato a quei controfattuali (non verofunzionali) che *non* sono veri semplicemente in virtù della falsità del loro antecedente. I controfattuali si basano su generalizzazioni conformi alla legge: 'Lo zucchero si dissolve nell'acqua' autorizza 'Se questo cubetto di zucchero fosse immerso nell'acqua si dissolverebbe'; ma 'Tutte le monete nella mia tasca sono d'argento' non porta a concludere 'Se questo penny fosse nella mia tasca sarebbe d'argento'. [*traduzione mia*]

Se la cupola di San Pietro viene portata a 1000 K allora emette luce[3] (1.3)

Niente Hitler, niente bomba atomica[4] (1.4)

oppure addirittura esprimendo la controfattualità attraverso un singolo concetto che comprende sia la descrizione dello stato in cui un determinato oggetto si trova, sia la descrizione dello stato in cui si troverebbe date alcune (diverse) circostanze, come è il caso dei predicati disposizionali, di cui un esempio è l'aggettivo "solubile" citato nella definizione dell'*Oxford Companion for Philosophy*.

Claudio Pizzi in [126] fornisce una spiegazione molto chiara del motivo per il quale tali predicati disposizionali siano sottesi da una dimensione controfattuale:

> In primo luogo, la presenza dei controfattuali nel linguaggio è poco apparente perché a volte usiamo costrutti linguistici in cui essi non compaiono in modo esplicito. Se dico che la zolletta che ho di fronte è solubile nel caffè ciò implica che se fosse stata messa nel caffè si sarebbe sciolta: dove l'ipotesi è certamente falsa in quanto, se si fosse verificata, non avrei nemmeno di fronte la zolletta di zucchero di cui sto parlando. Predicati come solubile, irascibile, fragile ecc. sono detti disposizionali perché descrivono la disposizione di un ente a reagire a determinati stimoli in circostanze possibili di qualche tipo.
>
> [Il Ragionamento Controfattuale, p.86]

La seconda – e più importante – ragione per la quale è bene non fermarsi alla loro analisi linguistica è che, qualunque sia la forma grammaticale attraverso la quale essi vengono espressi, i controfattuali rappresentano un tipo specifico di forma di ragionamento (in altri termini, un'operazione cognitiva) e, come tali, le loro istanze possono essere combinate tra loro o con istanze

[3]Esempio portato da Maria Luisa Dalla Chiara e Giuliano Toraldo di Francia in [34], p.68.

[4]Esempio citato da David K. Lewis in [99].

di altre forme di ragionamento, possono essere iterati e inseriti in processi di ragionamento più ampi e complessi.

Interpretata secondo questa seconda accezione, la controfattualità appare come una dimensione costitutiva della razionalità umana, incarnata dalla capacità, che gli umani possiedono, di astrarre da alcuni tratti di una situazione che percepiscono come reale, di immaginare delle situazioni alternative a questa, di ragionare all'interno dei confini di tali scenari alternativi ricavando delle informazioni che sono rilevanti per la situazione reale, ma che da questa non potevano essere direttamente inferite.

In modo molto simile John Pollock in [129] descrive quello che definisce *ragionamento supposizionale*, che altro non è se non il ragionamento ipotetico, di cui quello controfattuale è una sottoparte specifica:

> The employment of subsidiary arguments comprises *suppositional reasoning*, wherein we suppose something "for the sake of the argument", reason using the supposition in the same way we reason about beliefs and interests nonsuppositionally, and then on the basis of conclusions drawn using the supposition we draw further conclusions that do not depend upon the supposition. [...] Within the supposition, we reason as if the supposed propositions were beliefs, using all the rules for adoption and interest that were discussed in connection with linear reasoning[5].
>
> [Interest driven suppositional reasoning, p.427]

[5]L'impiego di argomenti sussidiari comprende il *ragionamento supposizionale*, nel quale supponiamo qualcosa "per amor di discussione", ragioniamo usando la supposizione nello stesso modo in cui ragioniamo non supposizionalmente su credenze e interessi e quindi, sulla base delle conclusioni ottenute usando la supposizione, traiamo ulteriori conclusioni che non dipendono dalla supposizione. [...] All'interno della supposizione, ragioniamo come se le proposizioni supposte fossero credenze, usando tutte le regole per l'adozione e l'interesse che sono state discusse relativamente al ragionamento lineare. [*traduzione mia*]

Le funzioni del ragionamento controfattuale

Il ragionamento così definito svolge una serie di funzioni tanto eterogenee quanto importanti nelle inferenze di senso comune normalmente utilizzate dagli esseri umani in vista dell'esecuzione dei compiti che quotidianamente si ritrovano a dover affrontare.

Tali funzioni appartengono sia alla sfera intellettuale, sia a quella emotiva, sia a quella pratica.

Un esempio tratto dalla sfera intellettuale è la generazione di controesempi al ragionamento deduttivo, sia essa intesa come *reductio ad absurdum* in un ragionamento formale che come *falsificazione* per un ragionamento empirico (scientifico o meno)[6]. Un altro esempio interessante sono le forme di ragionamento "ambigue", cioè che fanno uso di metafore, doppi sensi, ironia e in generale di ragionamenti analogici, come quelli che possono essere applicati a contesti di *fiction*.

Per quanto riguarda la sfera emotiva, un'ampia letteratura sia teorica che sperimentale in psicologia (vedi [114], [106], [149], [139], [86], ma soprattutto [115]) ha sostenuto l'ipotesi che il ragionamento controfattuale assolva compiti diversi nel caso in cui istituisca il paragone tra la realtà e uno scenario migliore (i cosiddetti controfattuali *upward*), oppure tra la realtà e uno scenario peggiore (controfattuali *downward*).

Nel caso *upward*, il ragionamento controfattuale può indurre rimpianto (quando il soggetto percepisce di non aver fatto tutto il possibile per raggiungere un obiettivo che ha fallito), rimorso (quando il soggetto ha fatto qualcosa che ha danneggiato qualcuno – se stesso o un altro soggetto – ed evitando di compiere qualche azione che ha invece compiuto avrebbe potuto evitare insieme anche il danno). Altro effetto del ragionamento controfattuale *upward* è quello di amplificare le emozioni dolorose quando il soggetto realizza che, con una leggerissima modifica del passato, la sua situazione attuale sarebbe decisamente migliore. Il fine ultimo del controfattuale *upward* sarebbe allora quello di ingenerare nel soggetto – nel momento in cui questi

[6]Pizzi, in [126], afferma: "Ragionamenti in cui si ipotizza qualcosa della cui verità non si è sicuri, o addirittura qualcosa della cui falsità si è sicuri, sono di uso corrente non solo nelle scienze formali ma anche nelle scienze empiriche e nella sfera del senso comune."

si trovasse a fronteggiare una situazione analoga a quella su cui ha ragionato controfattualmente – un ricordo doloroso che lo porti a evitare di ripetere gli errori del passato.

Consideriamo alcuni esempi:

- **Controfattuale *upward* esprimente rimpianto**: "Se avessi studiato di più avrei passato l'esame". Il rimpianto nasce dall'aver omesso di compiere un'azione laddove ce ne sarebbe stato bisogno;

- **Controfattuale *upward* esprimente rimorso**: "Se non avessi trascorso il weekend precedente facendo feste, avrei passato l'esame". Il rimorso segue dall'aver compiuto un'azione dagli esiti nefasti laddove l'inazione sarebbe stata preferibile;

- **Controfattuale *upward* che amplifica le sensazioni dolorose**: "Se fossi arrivato solo un minuto prima, sarei riuscito a prendere l'aereo". L'amplificazione discende da situazioni in cui un piccolo cambiamento di partenza determina un esito finale decisamente peggiore rispetto a quello della situazione immaginata; questa forma si può applicare indifferentemente all'una o all'altra delle due forme precedenti.

All'opposto, il controfattuale *downward* è causa di emozioni positive, quali l'orgoglio o la soddisfazione, come quando, grazie all'intervento dell'individuo ragionante vengono evitati esiti nefasti che si sarebbero altrimenti prodotti, oppure il sollievo, quando ci si trova a un passo dalla "catastrofe" ma questa, grazie a un dettaglio in apparenza insignificante, non si produce. Anche in questo caso, il fine ultimo dovrebbe essere quello di guidare l'individuo verso gli esiti più favorevoli grazie al ricordo di emozioni positive provate in situazioni analoghe.

- **Controfattuale *downward* esprimente orgoglio**: "Se non avessi studiato così intensamente non avrei passato l'esame". L'orgoglio viene espresso attraverso l'enfatizzazione di un'azione ritenuta fondamentale per l'ottenimento dell'obiettivo raggiunto;

- **Controfattuale *downward* esprimente sollievo**: "Se mi fossi riparato sotto l'albero, il fulmine mi avrebbe colpito". Il sollievo consegue dal non aver compiuto un'azione che poteva dare luogo a effetti indesiderabili;
- **Controfattuale *downward* che amplifica le sensazioni positive**: "Se fossi uscito di casa solo due minuti dopo non ti avrei incontrato". Come nel caso *upward*, questa amplificazione si applica a entrambe le forme precedenti.

La terza sfera, quella pratica, comprende l'individuazione di sottobiettivi quando il piano per il conseguimento di un obiettivo finale è molto articolato, la costruzione di corsi di azione alternativi, il controllo di piani formulati in passato e, conseguentemente, la previsione dell'esito di piani futuri.

Soprattutto su questa terza dimensione è concentrata l'attenzione del presente studio e la sua legittimazione poggia sull'assunto che il ragionamento controfattuale sia particolarmente importante per l'*agente* razionale, più ancora che per un generico *soggetto* razionale. In altre parole, il ragionamento controfattuale è particolarmente utile per tutti i processi cognitivi finalizzati all'azione, poiché permette di esplorare varie strategie alternative.

Il *pensiero* razionale trae un enorme vantaggio dal ragionamento controfattuale poiché questo rende espliciti al tempo stesso due scenari: quello controfattuale e quello fattuale (per contrasto). A sostegno di questa tesi le psicologhe Ruth Byrne e Alessandra Tasso hanno condotto uno studio empirico [29] nel quale, attraverso quattro esperimenti, hanno mostrato il potere di esplicitazione dei due scenari alternativi posseduto dal ragionamento controfattuale:

> Reasoners represent explicitly the case mentioned in the conditional, and they keep track of the possibility that there may be alternatives to it[7].
>
> [Deductive reasoning with factual, possible, and counterfactual conditionals, p.727]

[7] I soggetti ragionanti rappresentano esplicitamente il caso menzionato nel condizionale e mantengono traccia della possibilità che ci siano alternative a esso. [*traduzione mia*]

Gli esperimenti sembrano inoltre indicare che questo processo di esplicitazione intrinseco ai controfattuali abbia degli effetti positivi sulla realizzazione dei compiti assegnati ai soggetti.

> On this account, we can also make the further prediction that the initial understanding of a counterfactual is more difficult than the initial understanding of a factual conditional, because the counterfactual requires the construction of multiple models. Once this extra work is completed, however – as the results of these experiments have shown – it provides a richer basis for the subsequent tasks of deduction, verification, and falsification[8].
>
> [Deductive reasoning with factual, possible, and counterfactual conditionals, p.738]

Al di là del sistema teorico che Byrne e Tasso utilizzano per rendere conto dei risultati ottenuti che, non essendo – nemmeno nelle intenzioni – formale, sembra inadeguato per il discorso globale che si vuole affrontare in questa tesi, le evidenze empiriche sembrano dare supporto all'intuizione, che condividiamo con Byrne e Tasso, che il controfattuale fornisca un valore aggiunto specifico al bagaglio cognitivo di un soggetto razionale.

Per quanto concerne l'*azione* razionale, invece, l'utilità del ragionamento controfattuale discende dalla poca convenienza o – spesso – dall'impossibilità di verificare nella realtà gli effetti di un'azione; è sufficiente pensare a un pilota costretto a giudicare nella realtà (piuttosto che attraverso ipotesi controfattuali) la correttezza di tutte le manovre di volo; oppure si può pensare a uno scienziato costretto a verificare "nella realtà" la validità di una legge scientifica che assume l'assenza di attrito (come la legge di inerzia).

Tuttavia, sempre in relazione all'azione, questo tipo di vantaggio non è esclusivo del ragionamento controfattuale, ma appartiene anche ad altri ti-

[8]Sotto questo rispetto, possiamo anche fare l'ulteriore predizione che la comprensione iniziale di un controfattuale è più difficile della comprensione iniziale di un condizionale fattuale, perché il controfattuale richiede la costruzione di più modelli. Una volta che questo lavoro extra è completato, tuttavia – come hanno mostrato i risultati di questi esperimenti – fornisce una base più solida per i successivi compiti di deduzione, verificazione e falsificazione. [*traduzione mia*]

pi di ragionamento ipotetico, per esempio al ragionamento ipotetico della possibilità. In altre parole, il pilota di cui sopra può ragionare controfattualmente: "Se non avessi virato, avrei centrato i cavi dell'alta tensione", ma può ugualmente ragionare sulla possibilità: "Se non virassi ora, centrerei i cavi dell'alta tensione".

Qual è il vantaggio che il controfattuale può offrire rispetto al ragionamento ipotetico della possibilità? Quello di avere un punto fermo, cioè di sapere come sono andate le cose in realtà. Per tornare all'esempio, in un caso il pilota *sa* che con la virata ha evitato i cavi dell'alta tensione e può utilizzare quell'informazione (e altre informazioni che può dedurre a partire da essa) per ragionare sull'ipotesi controfattuale, mentre, nel secondo caso, l'agente può solo *credere* che con quella manovra eviterà l'ostacolo, ma non può esserne certo[9].

Possibili domini di applicazione del controfattuale

Nonostante lo studio dei controfattuali possa apparire a prima vista un mero esercizio intellettuale fine a se stesso, sembra che la sua utilità cominci a essere ampiamente riconosciuta, tanto che ha acquistato legittima cittadinanza in un numero sempre maggiore di discipline, per le quali è diventato uno strumento fondamentale.

Per esempio, esistono una serie di studi che coniugano giurisprudenza e psicologia criminale (vedi [149], [33], [161], [118]), nei quali il controfattuale viene utilizzato per valutare le attenuanti o le aggravanti di un delitto, per esempio, o per capire se l'azione dell'imputato è la vera e unica causa del danno subito dalla vittima (quella che in gergo viene definita *conditio sine qua non*).

[9]Quando affermiamo che l'agente *sa* cosa è accaduto, non intendiamo asserire che l'agente "crede A e A è vero", come spesso è stato sostenuto nel quadro della "tradizione" filosofica, quanto piuttosto che, nella teoria che l'agente utilizza per ragionare su quella che egli crede essere la realtà A è vero, mentre, nel caso ipotetico della possibilità, A non assume un valore di verità definito nella stessa teoria.

In economia, invece, l'uso dei controfattuali ha riguardato soprattutto due branche specifiche, quali la teoria delle decisioni ([56], [119]) e la teoria dei giochi ([15], [14], [13], [151], [80]), dove il controfattuale assurge a potente strumento euristico, soprattutto in situazioni di informazione imperfetta.

Per quanto riguarda la psicologia, oltre ai lavori già segnalati, che studiano i controfattuali e i tipi di situazione che generano diversi tipi di controfattuali, si danno studi in cui i controfattuali, piuttosto che essere oggetto di analisi, sono usati come strumento di analisi; come ad esempio nello studio di psicosi quali l'autismo dove, secondo una lettura che identifica in parte la sindrome autistica con l'incapacità del soggetto di elaborare una teoria della mente (vedi, ad esempio, [4] e soprattutto [30], che raccoglie una lunga serie di articoli sull'argomento), il ragionamento controfattuale può essere interpretato come uno dei modi che un individuo ha a disposizione per elaborare una teoria della mente dell'altro.

Nel campo dell'intelligenza artificiale, ci sono stati alcuni studi tesi ad applicare il controfattuale nel *planning* e nella diagnosi degli errori (a tale proposito si veda soprattutto il pionieristico [67]; interessanti intuizioni sono presenti anche in [87] e [41]).

Accanto a questi sviluppi, che sono prettamente accademici, si trova un filone, particolarmente fecondo negli ultimi anni, in cui ricerca scientifica e produzione letteraria si intrecciano, cioè quello della cosiddetta "storia virtuale". Da un lato si indaga l'importanza storica di alcuni avvenimenti, immaginando esiti diversi di battaglie cruciali o atteggiamenti strategici diversi da parte di generali e condottieri (questo è il caso di libri come [157], [55] e [43]), dall'altro si creano racconti o romanzi di fantasia a partire dall'alterazione di un fatto storico, all'interno di un quadro storiografico fedele alla realtà – o presunto tale (come in [158], [148] e [84]).

Per mostrare come ricostruzioni storiche accurate e voli di fantasia si amalgamino in questo genere di letteratura, riportiamo una parte dell'introduzione della raccolta di saggi curata da Robert Cowley [43]:

> E se un'epidemia misteriosa non avesse colpito gli assedianti assiri di Gerusalemme nel 701 a.C.? Ci sarebbe stata una religione ebraica? O il cristianesimo? Prendiamo fatti della durata di frazioni di secondo: che cosa sarebbe successo se la traiettoria

> di un'ascia da guerra non fosse stata interrotta e il ventunenne Alessandro fosse stato ucciso prima di diventare "Magno"? O se Cortés, che per poco non fu catturato all'assedio di Tenochtitlán, l'odierna Mexico City, fosse caduto per davvero prigioniero? È molto probabile che i giovani Stati Uniti si sarebbero ritrovati un grande impero indigeno americano ai loro confini meridionali. Proviamo a considerare anche il ruolo del caso: se, nella guerra civile americana, il famoso "ordine perduto" non fosse andato perduto, è probabile che, come scrive James M. McPherson, gli Stati Confederati sarebbero rimasti indipendenti. Ma, di fatto, un analogo "ordine perduto" influenzò l'esito della battaglia della Marna nel settembre 1914 e, di conseguenza, la stessa prima guerra mondiale.
>
> [*La storia fatta con i se*, p.9]

Il funzionamento del ragionamento controfattuale

Tutti i casi finora descritti condividono uno scenario centrato su un problema da risolvere e su una serie di possibili soluzioni alternative tra cui scegliere. Il ragionamento controfattuale diventa allora uno strumento per poter analizzare e giudicare la bontà delle scelte fatte. Proprio questa dimensione *pragmatica* del ragionamento controfattuale è quella che vorremmo approfondire in questo lavoro ed è proprio la dimensione pragmatica che deve informare la nostra descrizione teorica non solo di come opera, ma anche di come è intrinsecamente costituito.

Sul lato del "come opera" il ragionamento controfattuale, la nostra ipotesi di lavoro è che lo faccia in due direzioni, identificate da due forme di razionalità:

- **Razionalità strumentale**: a partire da un insieme di assunzioni e preferenze considerate fisse e immutabili, individua un obiettivo e applica la propria funzione critico-dialettica ai possibili mezzi per raggiungerlo;

- **Razionalità ex-post**: a partire da un insieme di mezzi (capacità, risorse), considerati fissi e immutabili, sottopone a critica dialettica le preferenze/assunzioni, al fine di individuare un obiettivo raggiungibile a partire dai mezzi a disposizione.

Ora, intuitivamente, mentre il primo tipo di razionalità sembra riducibile a un processo di revisione di elementi *all'interno* di una teoria, il secondo sembra piuttosto corrispondere con l'assunzione di una nuova prospettiva, ossia con l'atteggiamento di rivedere lo stesso problema alla luce di una *differente* teoria.

Per questo motivo, il sistema da adottare per la trattazione del ragionamento controfattuale deve essere in grado di rappresentare sia le operazioni che hanno luogo *all'interno* delle teorie sia quelle che si compiono *tra* teorie, in grado di rendere ragione sia del carattere circoscritto di certi ragionamenti, sia dell'importanza delle relazioni che sussistono tra processi di ragionamento condotti sulla base di assunzioni diverse.

Una teoria contestuale per il ragionamento controfattuale

L'idea da cui partiamo per descrivere la nostra teoria è quella secondo la quale tale teoria dovrebbe rendere conto il più chiaramente possibile di come un agente razionale elabori un processo di ragionamento controfattuale[10].

A nostro modo di vedere, l'ideazione di un'ipotesi controfattuale (e del conseguente ragionamento) è un processo di pensiero che ne presuppone un altro: quello della selezione dell'informazione che l'agente giudica rilevante per ragionare su quello specifico argomento.

Questo processo di selezione è centrale ed è anche cruciale per la determinazione dell'esito del ragionamento controfattuale poiché, a seconda delle

[10]Una prima formulazione preliminare delle intuizioni che ci hanno guidato nella decisione di affrontare il ragionamento controfattuale nella prospettiva contestuale è contenuta in [57].

informazioni e delle regole che l'agente *decide* di utilizzare, il processo inferenziale darà risultati di volta in volta diversi. Riteniamo questa caratteristica irrinunciabile perché permette di spiegare da una parte il fatto che agenti diversi possano elaborare ragionamenti controfattuali con esiti opposti considerando lo stesso problema e dall'altra parte rende conto del fatto che uno stesso agente, quando viene a conoscenza di nuove informazioni, può modificare sensibilmente i suoi processi di ragionamento, fino a stravolgerne i risultati.

Inoltre, sempre da questo processo di selezione dovrebbe dipendere addirittura che cosa è fattuale e che cosa è controfattuale poiché, se ad esempio un agente si inganna sul reale stato delle cose, potrebbe considerare fattuale ciò che un "osservatore esterno" giudicherebbe controfattuale e viceversa, ma gli effetti pratici del suo ragionamento (per esempio, le azioni che compierebbe a partire da tale ragionamento), sarebbero comunque in linea con la sua interpretazione e non con la realtà osservata dall'altro (un esempio che mostra questo fenomeno è descritto nella sezione 4.6).

Questa relatività e flessibilità dei concetti di fattuale e controfattuale sono di particolare importanza per lo studio delle teorie scientifiche, data la loro natura provvisoria e data la frequenza con cui si verificano situazioni nelle quali scienziati provenienti da diverse comunità scientifiche assegnano significati diversi agli stessi eventi e, laddove gli uni vedono una variabile gli altri vedono una costante e viceversa. In un certo qual modo, quando una teoria scientifica è consolidata all'interno di una comunità, essa viene trattata *come se* descrivesse i fatti (ossia, come se fosse fattuale) e le teorie alternative vengono percepite come controfattuali; tuttavia, questa prospettiva può essere ribaltata in qualsiasi momento.

In altre parole, a nostro avviso, un agente possiede una base di conoscenza molto ampia e articolata[11], che però non gli è tutta immediatamente disponibile quando si accinge a risolvere un problema.

[11] Ai fini di questo lavoro non è di cruciale importanza distinguere tra un'enorme e indistinta base di conoscenza o una base ripartita in sottodomini fra loro strutturati, anche se noi propendiamo per la seconda tesi, poiché questa immagine è più rispondente all'idea, da noi sostenuta, che un agente, posto di fronte a un problema da risolvere, attivi di volta in volta una porzione specifica delle sue conoscenze.

Questo per svariati motivi; prima di tutto per ragioni "economiche": se un agente dovesse tenere in considerazione tutto ciò che sa prima di poter formulare un piano o di stabilire una strategia, i processi decisionali sarebbero molto lenti e dispendiosi. In secondo luogo, la base di conoscenza di un agente potrebbe anche contenere delle informazioni contraddittorie, che sono state apprese dall'agente in circostanze diverse. Noi vorremmo invece rendere conto del fatto che l'agente può scegliere di considerare vero un fatto quando ragiona in un contesto e vera la negazione dello stesso fatto quando ragiona in un contesto differente[12].

Inoltre, come già evidenziato prima, diversi agenti possono avere una prospettiva anche molto diversa riguardo un unico problema, senza che ciò comporti una ricaduta nel relativismo assoluto, poiché è proprio la selezione dell'informazione rilevante che identifica il contesto fattuale di partenza che determina i vincoli che deve soddisfare il ragionamento controfattuale e la coerenza del ragionamento controfattuale ideato dall'agente è quindi subordinata alla sua capacità di identificare l'informazione rilevante per ragionare su uno specifico problema[13].

Come fanno notare Tetlock e Belkin in [156]:

> Different investigators will inevitably emphasize somewhat different criteria in judging the legitimacy, plausibility, and insight-

[12]Un esempio a questo punto potrebbe essere utile. Supponiamo che un agente creda che la legge di gravitazione universale di Newton sia falsa alla luce della relatività einsteiniana; un asserto che seguisse da tale legge sarebbe considerato falso dall'agente e un ragionamento condotto a partire da tale asserto controfattuale in relazione a un problema affrontato con l'ausilio della teoria einsteiniana. Lo stesso agente potrebbe comunque ritenere vera (e fissarla come vera nelle ipotesi del ragionamento) tale legge di gravitazione universale newtoniana nell'affrontare un problema di fisica su scala 'terrestre'. In tale contesto, gli asserti derivati dalla legge di Newton sarebbero da lui considerati veri e i ragionamenti condotti su di essi "fattuali".

[13]Non possiamo escludere che questa nostra posizione sarebbe stigmatizzata da Tetlock e Belkin, secondo la citazione che segue, come "soggettivismo del va bene tutto", poiché in essa sarebbe del tutto legittimo il processo di ragionamento di uno psicotico che ragionasse a partire dalle sue fantasie, purché queste mantengano una loro coerenza interna. L'interesse primario della nostra teoria è che questa descriva una procedura di ragionamento corretta, al di là dell'appropriatezza delle premesse da cui parte.

> fulness of specific counterfactuals. It would be a big mistake, however, to confuse epistemic pluralism (which we accept up to a point) with an anything-goes subjectivism (which we reject and which would treat all counterfactual claims as equally valid in their own way)[14].
>
> [Counterfactual Thought Experiments in World Politics]

Infine, in ultima analisi, se fosse vero che, al momento di prendere una decisione, un agente ha a disposizione tutto ciò che sa, risulterebbe alquanto difficile spiegare il fatto che spesso gli agenti compiono errori, anche banali, perfino in domini dei quali sono molto esperti.

Per tutte queste ragioni l'ipotesi che l'intera base di conoscenza sia il punto di partenza dei ragionamenti ci pare inidonea. Più sensato ci sembra invece sostenere che gli agenti, per ragionare su specifici problemi, "ritaglino" una porzione di questa base di conoscenza e se ne servano per costruire la teoria parziale che useranno per ragionare sul problema, quello che, in base alla definizione fornita in [20], chiamiamo contesto o teoria di lavoro (*working context*).

Questa operazione è dunque preliminare rispetto a quella che compie un soggetto quando formula un'ipotesi controfattuale e, a partire da essa, sviluppa un ragionamento parimenti controfattuale.

Nel nostro modello è dalla scelta iniziale di quali assiomi utilizzare nel ragionamento specifico che emergono, in virtù del fatto di essere posti in una certa relazione di compatibilità (che definiremo nel caso specifico *relazione di controfattualità*), il *contesto fattuale* e il *contesto controfattuale*, come mostra la figura 4.6. A seconda dei diversi sistemi di assiomi che possono essere assunti da un agente, si determineranno quindi diverse coppie di contesti fattuali e controfattuali relativi a un medesimo problema.

---

[14]Investigatori differenti sottolineeranno inevitabilmente criteri in qualche modo differenti nel giudicare la legittimità, la plausibilità e la capacità di approfondimento di specifici controfattuali. Sarebbe un grosso sbaglio, tuttavia, confondere il pluralismo epistemico (che accettiamo fino a un certo punto) con un soggettivismo del tipo "va bene tutto" (che respingiamo e che tratterebbe le asserzioni controfattuali come ugualmente valide a loro modo).[*traduzione mia*]

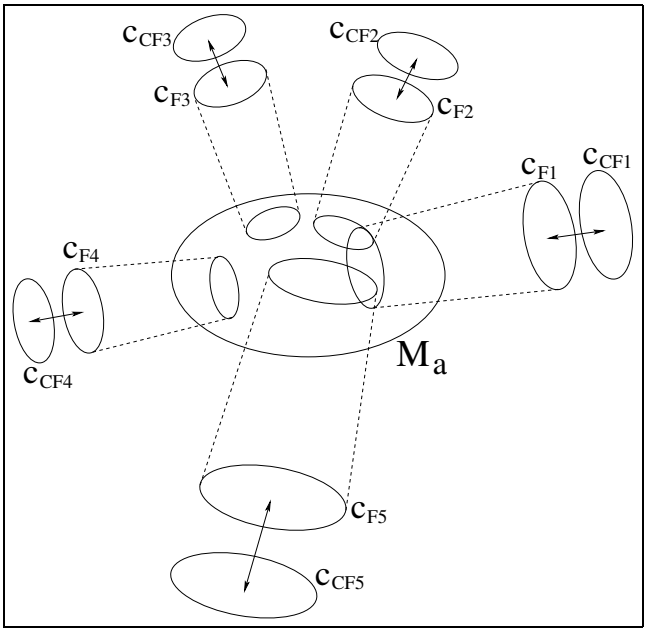

Figura 1.1: Costruzione di coppie di controfattualità

Consideriamo il classico esempio nel quale ci si interroga sulla verità dell'enunciato "Se quel fiammifero fosse stato sfregato, si sarebbe acceso"; se l'indagine viene condotta in un contesto (ossia una teoria) in cui gli assiomi *localmente validi* non parlino della presenza di ossigeno, ma solo del legame tra sfregamento e accensione di fiammiferi, verosimilmente l'esito del ragionamento sarà l'affermazione della verità del condizionale di cui sopra; se, invece, tra gli assiomi locali del contesto di ragionamento ce n'è almeno uno che parla di ossigeno, allora potrebbe succedere che, in alcuni dei modelli del contesto l'assenza di ossigeno impedisca l'accensione del fiammifero. Il controfattuale, nella forma presentata sopra, sarà dunque falso o almeno indecidibile. Questo è un esempio di enunciato che può essere vero o falso a seconda del contesto di ragionamento in cui è inserito.

A livello intuitivo, posto di fronte a un enunciato controfattuale al quale deve assegnare un valore di verità, l'agente deve decidere:

1. in primo luogo quali sono le leggi generali che gli servono per ragionare sul problema che si trova a dover risolvere;

2. verificare cosa succede nella "realtà";

3. porre un'ipotesi controfattuale;

4. trarre le conseguenze di tale ipotesi partendo dalle leggi generali che ha selezionato;

5. decidere il valore di verità dell'enunciato controfattuale.

Questi passi, a livello logico, corrispondono alle seguenti operazioni:

1. costruire l'insieme di tutti i possibili modelli della situazione, selezionando i termini del linguaggio che verrà utilizzato nello specifico caso (attraverso tutte le possibili combinazioni dei termini del linguaggio, dalle quali vanno eliminate quelle che contraddicono gli assiomi);

2. costruire il contesto fattuale, scegliendo, tra le interpretazioni rimaste, quelle in cui sia l'antecedente che il conseguente dell'enunciato controfattuale sono falsi;

3. costruire il contesto controfattuale, scegliendo, tre le possibili interpretazioni, quelle in cui l'antecedente dell'enunciato controfattuale è vero;

4. verificare il valore di verità del conseguente in tali interpretazioni: se in tutte le interpretazioni residue esso è vero, allora vale l'enunciato controfattuale sul quale si stava ragionando; se in tutte le interpretazioni è falso, allora varrà il corrispondente enunciato semifattuale[15]; se invece in alcune interpretazioni è vero e in altre falso, allora il valore di verità dell'enunciato non è determinabile a partire da quegli assiomi, ossia *in quella teoria*, ma sarà vero il corrispondente controfattuale della possibilità[16]: "Se fosse successo A sarebbe potuto succedere B".

[15]Per semifattuale si intende, in accordo con la letteratura, un condizionale avente antecedente falso e conseguente vero, del tipo "Se anche fosse successo A sarebbe comunque successo B".

[16]Quello che D. K. Lewis chiama *might counterfactual*.

La conseguenza più significativa derivante dall'assunzione di questo approccio è che in base a esso non è possibile affermare la verità o la falsità *tout court* di un enunciato controfattuale, ma il suo valore di verità dipenderà dagli assiomi della teoria a partire dalla quale il contesto controfattuale di cui è parte è stato costruito. In altri termini, sullo stesso enunciato controfattuale, ossia sulla stessa coppia antecedente-conseguente possono essere costruite un numero n possibilmente anche infinito di coppie di contesti fattuale-controfattuale, ognuna delle quali individuata da una relazione di controfattualità differente.

Questo scenario determina anche una diversa concezione di che cosa sia il "fattuale": esso passa da essere qualcosa che è vero nel mondo reale (o comunque in un mondo possibile) a ciò che è vero in una teoria, ossia in tutti i modelli locali di un contesto individuato da una specifica relazione di compatibilità.

Questo approccio risulta molto vicino a una certa parte della filosofia della scienza del Novecento ([82], [58], [91], [47], [127]), poiché in essa si afferma l'assoluta impossibilità di asserire la verità o falsità di un enunciato scientifico senza collocarlo nella specifica tradizione scientifica o, ancor meglio, nel particolare paradigma teorico dai quali è stato generato.

Ugualmente riconducibile a questo approccio è il precetto linguistico, che potremmo definire dell'*olismo*, largamente condiviso dalle nostre nozioni di senso comune, che una parola acquista significato solo all'interno di una frase e una frase, a sua volta, acquista significato solo all'interno di un discorso.

Sia il "caso scientifico" sia quello "linguistico" sembrano sottolineare che la verità o falsità degli enunciati o, in ultima analisi, il loro significato, dipendono da una serie di regole che non sono generalissime e date una volta per tutte, ma relative a uno specifico ambito ed è proprio questo ambito a fornire una struttura interpretativa.

Il punto di vista privilegiato non è più dunque quello "metafisico" della realtà, ma diventa quello della prospettiva cognitiva particolare dell'agente razionale, che non effettua più operazioni su enunciati all'interno di una teoria che gli è già data dall'inizio, ma opera invece direttamente sulle teorie che di volta in volta costruisce appositamente per risolvere problemi specifici. Questo sarebbe testimoniato sia dalla velocità con cui gli agenti compiono

certi ragionamenti, difficilmente spiegabile se si considera che essi debbano vagliare tutta l'informazione a loro disposizione, sia dal fatto che essi spesso giungano a conclusioni diverse quando acquisiscono nuova informazione e, in qualche modo, costruiscono una nuova teoria.

La differenza tra questi due tipi di approccio e le conseguenze da essi derivanti per la rappresentazione della conoscenza e per i processi di ragionamento è ben evidenziata in [19].

Un ulteriore vantaggio del formalismo che abbiamo scelto di utilizzare è quello di facilitare notevolmente la rappresentazione del ragionamento, poiché questo avviene nel contesto, che è un *insieme* di modelli, quindi un oggetto *parziale*, che non assegna valore di verità a tutti i termini del linguaggio che lo caratterizza, ma solo a quelli di cui si vuole parlare nella teoria. Il processo di ragionamento in questo caso coinvolge solo una parte ristretta dell'informazione disponibile e questo fa sì che le operazioni siano più veloci e agevoli.

Per comprendere meglio il diverso modo di procedere della semantica a modelli locali (rispetto, per esempio, alle teorie basate su oggetti completi come quelle della logica modale, come i mondi possibili), si pensi a un esempio ormai divenuto classico, quello proposto da Kit Fine relativo a Nixon e all'olocausto.

Abbiamo Nixon che, in un periodo di crisi internazionale sta seduto nella famigerata "stanza dei bottoni"; sappiamo che – fortunatamente – Nixon non ha premuto alcun bottone ma, si domanda Fine, "Se Nixon avesse premuto il bottone, ci sarebbe stato l'olocausto nucleare" è un enunciato vero o falso?

Per rispondere a questa domanda, i teorici dei mondi possibili devono fare un confronto tra il mondo da cui partono (assumiamo per semplicità che sia quello reale) e i diversi mondi possibili alternativi. Per fare questo, hanno bisogno di assegnare un valore di verità a tutti gli enunciati del linguaggio per ogni mondo possibile. Una volta effettuata questa assegnazione, dovranno ordinare i mondi a seconda della loro somiglianza con il mondo di partenza.

E qui emerge il problema messo in luce da Fine: nel caso di Nixon, i teorici dei mondi possibili (nella fattispecie Lewis, verso cui Fine dirige la sua critica) vorrebbero affermare che l'enunciato controfattuale di cui sopra è vero, ma come fanno a dire che un mondo in cui si abbia l'olocausto sia

più simile al mondo reale di un mondo in cui si abbia un guasto al circuito elettrico e si scampi dall'olocausto?

Ora, lasciamo da parte in questo contesto la soluzione abbozzata da Lewis, che chiama in causa miracoli piccoli e grandi e concentriamoci sulla diversa natura delle risposte fornite dalla semantica a modelli locali (SML).

In primo luogo, nella SML viene costruito un contesto contenente tutti i fattori che l'agente ritiene necessari per il ragionamento che deve intraprendere e gli assiomi che intende usare: questo contesto consisterà di un insieme di tutte le possibili interpretazioni (risultanti dalla combinazione dei termini appositamente selezionati per ragionare sul problema) che rispettano gli assiomi e i vincoli imposti dall'enunciato controfattuale sul quale si sta ragionando (ossia che l'antecedente e il conseguente siano falsi). In maniera speculare, viene costruito il contesto controfattuale, che racchiude le interpretazioni che rispettano gli assiomi e i vincoli (in questo caso, che l'antecedente dell'enunciato controfattuale sia vero). Il passo successivo è quello di verificare qual è il valore di verità del conseguente del controfattuale in tutte le interpretazioni contenute nel contesto controfattuale.

La risoluzione che viene fornita al problema di Fine è essenzialmente diversa da quella proposta dai teorici dei mondi possibili: infatti, se la teoria di partenza "dice qualcosa" del guasto al circuito elettrico, il valore di verità dell'enunciato controfattuale dipenderà anche dal suo verificarsi o meno nel contesto controfattuale, in caso contrario esso sarà del tutto ininfluente. Se, d'altro canto, nel contesto fattuale si parla di un guasto elettrico che si è verificato, ma nel contesto controfattuale non viene espressamente avanzata alcuna ipotesi relativa al guasto, mentre nelle semantiche a mondi possibili si sarebbe portati a considerare mondi possibili nei quali il guasto si fosse ugualmente verificato (per ragioni di somiglianza), nel nostro contesto controfattuale verrebbero mantenuti sia i modelli locali nei quali il guasto si fosse verificato, sia quelli in cui ciò non fosse accaduto, perché tutti compatibili col ragionamento controfattuale in oggetto.

A questo punto la questione della somiglianza scompare completamente poiché, attraverso la scelta della teoria entro la quale si vuole ragionare, si scelgono anche i fattori che devono essere presi in considerazione nel ragionamento e, una volta effettuata la scelta, non ci sono ordinamenti da stabilire,

solamente bisogna andare a guardare cosa succede nelle interpretazioni che soddisfano certi vincoli; in altri termini, non si stabilisce alcuna gerarchia tra i modelli locali (contrariamente a quanto accadeva prima per i mondi possibili), si prendono *tutti e soli* quelli che soddisfano i vincoli posti dalla relazione di compatibilità.

Quindi, al di là degli elementi che si sono selezionati e sui quali si sta ragionando, tutto il resto può rimanere uguale o differire massimamente dalla situazione di partenza: al di là di tali elementi, le interpretazioni massimamente simili e quelle massimamente divergenti hanno lo stesso peso nel determinare il valore di verità dell'enunciato controfattuale, il quale risulta vero o falso *in una teoria*.

In conclusione, sono principalmente due gli aspetti – strettamente concatenati tra loro – che hanno determinato la scelta in favore della SML.

La prima caratteristica è quella che permette di valutare la verità o falsità di un controfattuale diversamente a seconda della teoria dalla quale si parte per compiere la valutazione. Questa caratteristica porta con sé, come corollario, la capacità di esprimere la non monotonicità del ragionamento controfattuale, ossia il fatto che, a differenza del condizionale materiale, il controfattuale può cambiare di valore di verità con l'aggiunta di una premessa[17]. Questo fenomeno, nella SML, corrisponde alla costruzione di un diverso contesto fattuale e, conseguentemente, anche di un diverso contesto controfattuale.

La seconda peculiarità è relativa alla descrizione di come avviene l'assegnazione del valore di verità all'enunciato controfattuale, la quale ha luogo attraverso una procedura semplice quanto fallibile, ma che non impegna l'agente nella considerazione di tutte le caratteristiche della situazione che si trova a vivere e di quelle nelle quali si sarebbe potuto trovare, ma solo di quelle selezionate sulla base della teoria che sta utilizzando.

---

[17]Si ricordi l'esempio di cui sopra: 'Se Nixon avesse premuto il bottone, ci sarebbe stato l'olocausto' può essere vera e, al tempo stesso, può essere falsa: 'Se Nixon avesse premuto il bottone e si fosse verificato un guasto nel circuito elettrico, ci sarebbe stato l'olocausto'.

Una formalizzazione del ragionamento controfattuale basata sulla SML

Uno degli scopi centrali di questo lavoro è quello di trovare una sistematizzazione formale che sia appropriata a rappresentare quelle caratteristiche del ragionamento controfattuale che abbiamo messo in luce.

L'intuizione di fondo che ci ha guidati in questa scelta è che il ragionamento controfattuale sia un particolare tipo di ragionamento contestuale, un ragionamento cioè che ha luogo *dentro* e *attraverso* domini circoscritti. Se questi domini, come ben argomentato in [10], vengono descritti come caratterizzati da tre proprietà: parzialità, approssimazione e prospettiva, le relazioni che intercorrono tra di essi possono essere a loro volta descritte come variazioni di livello di parzialità, di grado di approssimazione o di prospettiva.

Il ragionamento contestuale e, con esso, il ragionamento controfattuale possono essere descritti e rappresentati in termini di operazioni *su* e *tra* domini parziali, approssimati e prospettici.

Sulla scia di questa idea, nel costruire il sistema formale abbiamo scelto di servirci di un sistema logico (e della relativa semantica) che sono stati ideati per rappresentare il ragionamento contestuale e che si sono rivelati particolarmente efficaci nel risolvere una serie di problemi che emergono dallo studio di questa forma di ragionamento. Tale sistema logico prende il nome di Sistemi MultiContesto (o *MultiContext Systems*) e la corrispondente semantica è la Semantica a Modelli Locali (o *Local Models Semantics*).

I Sistemi MultiContesto sono composti da teorie (i contesti, appunto), connesse l'un l'altra da particolari tipi di legami. Dal punto di vista sintattico, i contesti sono teorie caratterizzate ognuna da un linguaggio, un insieme di assiomi e un insieme di regole di inferenza loro proprie. Il legame tra un contesto e un altro avviene grazie alla presenza di regole che permettono di importare ed esportare informazione; tali regole sono dette *regole ponte* (o *bridge rules*).

Parallelamente, a livello semantico, un contesto è un insieme di modelli locali (un modello locale è un modello classico *à la* Tarski), i quali intrattengono fra di loro delle cosiddette *relazioni di compatibilità*, che esplicitano il tipo di vincolo che deve sussistere tra due contesti perché possano essere

definiti compatibili secondo la specifica nozione di compatibilità che si sta in quel momento formalizzando.

Nel nostro lavoro abbiamo tentato di fornire una serie di definizioni semantiche per il ragionamento controfattuale costruendo uno specifico Sistema MultiContesto, formato da coppie di contesti fattuale/controfattuale, in cui la relazione che intercorre tra queste coppie di contesti è una specifica relazione di compatibilità, che abbiamo definito *relazione di controfattualità*, costruita utilizzando dei vincoli specifici che tali contesti devono soddisfare perché possano essere definiti rispettivamente fattuale e controfattuale.

Un ragionamento controfattuale è quindi un processo che ha luogo tra due contesti determinati da due fatti (che sono l'antecedente e il conseguente dell'enunciato controfattuale), che sono entrambi falsi da un parte (nel contesto fattuale) e almeno uno vero dall'altra (l'antecedente nel contesto controfattuale). Un processo inferenziale ha luogo nel contesto controfattuale e la conclusione di questo processo viene importata nel contesto fattuale ed espressa attraverso un certo operatore modale relativo ai due fatti (il predicato dice che l'antecedente e il conseguente hanno una relazione controfattuale).

Naturalmente, uno stesso contesto fattuale può essere connesso a più contesti controfattuali attraverso relazioni di controfattualità diverse e lo stesso rapporto fattuale/controfattuale sussistente tra due contesti può essere invertito se stabilito da una diversa relazione di controfattualità e queste due proprietà rendono questo sistema profondamente diverso da quelli che costituiscono la visione standard in filosofia.

Struttura del lavoro

La tesi è strutturata in tre parti; una prima parte nella quale viene introdotto il formalismo che è stato utilizzato per descrivere la rappresentazione e il funzionamento del ragionamento controfattuale e viene posto a confronto con altri formalismi o trattazioni intuitive fornite in precedenza; nella seconda parte il ragionamento controfattuale viene applicato a varie dimensioni del ragionamento pratico, allo scopo di mostrare come esso possa costituire uno strumento di ragionamento particolarmente efficace; infine, nella terza parte

vengono tracciate le linee generali di possibili sviluppi futuri, soprattutto nella direzione delle applicazioni.

La prima parte mostra il cammino che ci ha portati all'elaborazione del nostro sistema formale, partendo dall'analisi che è stata compiuta, soprattutto negli ultimi trent'anni, nell'ambito della filosofia del linguaggio, volta a studiare la semantica dei condizionali controfattuali (capitolo 2), passando attraverso le teorie, sviluppate principalmente nell'ambito della psicologia cognitiva e dell'intelligenza artificiale, che interpretano il controfattuale come un fenomeno di ragionamento ma, nel primo caso, senza fornire alcun modello formale, nel secondo caso fornendo modelli che catturano solo alcune delle proprietà che riteniamo caratterizzare questo tipo di ragionamento (capitolo 3). Il capitolo 4, invece, presenta il sistema formale da noi sviluppato utilizzando la Semantica a Modelli Locali, una logica per il ragionamento contestuale già ampiamente "collaudata" nella risoluzione di problemi specifici, emergenti soprattutto dalla sfera dell'intelligenza artificiale.

La seconda parte, invece, è strutturata in tre capitoli: il capitolo 5 introduce i concetti-chiave del ragionamento pratico e fornisce i presupposti della nostra analisi, la quale individua una direttrice rispetto al problema della razionalità che può essere seguita in due sensi opposti: dalle preferenze ai mezzi da procurarsi per raggiungere un obiettivo e dai mezzi disponibili alle nuove preferenze, che determinano la formazione di nuovi obiettivi; i due capitoli seguenti, il 6 e il 7 mostrano come il ragionamento controfattuale sia attivo in entrambe le forme di razionalità.

Infine, la terza parte, sugli sviluppi futuri, introduce brevemente dei possibili domini di applicazione del ragionamento controfattuale, quali la razionalità scientifica, gli agenti artificiali e gli scenari multiagente e fornisce una descrizione preliminare di quale potrebbe essere il modo di affrontare tali domini a partire dal *framework* precedentemente delineato.

Parte I

Quale teoria per il ragionamento controfattuale

Capitolo 2

Teorie formali per i condizionali controfattuali

There are ever so many ways that a world might be; and one of these many ways is the way that this world is[1].
[David Lewis, *On the Plurality of Worlds*, p. 2]

Scopo di questo capitolo è di esporre una rassegna, che non ha la pretesa di essere esaustiva, di alcune posizioni a nostro avviso particolarmente significative delineatesi negli ultimi quarant'anni relativamente al problema dei condizionali controfattuali e alla determinazione del loro valore di verità o di un criterio di accettabilità.

Il capitolo è diviso in due parti nelle quali vengono presentati due raggruppamenti di approcci: il primo raggruppamento si propone di individuare le condizioni che devono soddisfare gli enunciati componenti un controfattuale perché questo possa essere valutato come vero, mentre il secondo raggruppamento indaga la derivabilità del conseguente di un controfattuale a partire dall'antecedente e da altri fattori.

Il primo raggruppamento di approcci, che abbiamo definito "vero-funzionale", utilizza prevalentemente i sistemi di logica modale – che può in un certo senso essere considerata attualmente la teoria standard per la risoluzione di questo tipo di problemi.

[1] Ci sono così tanti modi in cui un mondo potrebbe essere fatto; e uno di questi molti modi è il modo in cui questo mondo è fatto.

Il secondo raggruppamento di approcci l'abbiamo chiamato "consequenzialista", seguendo la definizione fornita da Pizzi in [126], ma è altrove stato definito "meta-linguistico".

Nessuno di questi due raggruppamenti di approcci può essere considerato sotto alcun rispetto una scuola o una corrente unitaria. Il criterio di ordinamento che abbiamo scelto di utilizzare nasce dall'esigenza di identificare una caratteristica che sia sufficientemente differenziante e che, al tempo stesso, permetta di segnare l'inizio di un percorso che condurrà progressivamente verso la nostra proposta.

Tuttavia, quale criterio unificante, come vedremo alla fine del capitolo, tutti questi approcci ricercano quello che potremmo definire un *criterio di rilevanza globale* per determinare quali fatti vadano utilizzati nella valutazione o nella derivazione del controfattuale. Nel nostro approccio, invece, non esiste *un solo* criterio di rilevanza, poiché verità e derivabilità sono definite sempre localmente e non in modo universale.

2.1 Approcci vero-funzionali

L'idea di applicare la logica modale allo studio dei condizionali e, in particolar modo, dei controfattuali, discende – probabilmente – dall'idea di Robert Stalnaker di dare un'interpretazione in termini di mondi possibili del test formulato da Frank Ramsey in [134] per giudicare gli enunciati condizionali. Il test consisteva in un processo in tre passi:

1. aggiungere ipoteticamente l'antecedente al *corpus* delle proprie credenze;

2. rivedere il proprio *corpus* di credenze il minimo necessario per poter assumere l'antecedente del condizionale;

3. valutare l'accettabilità del conseguente a partire dal *corpus* così modificato.

Questa revisione del *corpus* di credenze può essere tradotta, secondo Stalnaker e secondo chi ne ha seguito l'esempio, in un'ipotetica minima revisione del mondo reale necessaria a rendere l'antecedente vero.

Questa traduzione di insiemi di credenze in mondi possibili favorisce l'utilizzo della logica modale come strumento per determinare le condizioni di verità degli enunciati controfattuali.

2.1.1 La funzione di somiglianza di Stalnaker

La teoria di Stalnaker, enunciata solo in forma intuitiva in [150] e formalizzata in un articolo scritto a quattro mani con Rich Thomason [152], è basata su una funzione di selezione su mondi possibili.

Come rilevato in precedenza, Stalnaker parte dall'interpretazione che Ramsey dà del condizionale e cerca un "analogo ontologico" del *corpus* di credenze ipotetiche di cui quest'ultimo parlava. Tale analogo ontologico sono, appunto, i mondi possibili.

Così Stalnaker fornisce una semantica basata sui mondi possibili comprendente anche una funzione di selezione che discrimina tra i mondi possibili; ciò permette di effettuare la valutazione dell'enunciato controfattuale solo su alcuni di questi o, più precisamente, come vedremo meglio in seguito, su uno di questi in particolare.

Stalnaker afferma dunque che un controfattuale è vero in un mondo se il conseguente è vero in un mondo possibile individuato da una funzione di selezione, la quale dovrebbe scegliere il mondo possibile in cui l'antecedente è vero e che differisce il meno possibile dal mondo di partenza.

La funzione di selezione dell'approccio di Stalnaker prende come argomenti una proposizione e un mondo possibile e restituisce un altro mondo possibile. La funzione seleziona, per ogni antecedente, un mondo possibile nel quale tale antecedente è vero. L'intero controfattuale è vero nel mondo di partenza se il suo conseguente è vero nel mondo individuato dalla funzione.

La selezione è fondata su un ordinamento dei mondi possibili secondo la somiglianza al mondo di partenza.

In [150] troviamo una spiegazione intuitiva del modo in cui Stalnaker ricava il valore di verità dei controfattuali:

> In addition to a model structure, our semantical apparatus includes a *selection function*, f, which takes a proposition and a

> possible world as its value. The *s*-function selects, for each antecedent A, a particular world in which A is true. [...] I shall use the following terminology for talking about the arguments and values of *s*-functions: where $f(A, \alpha) = \beta$, A is the *antecedent*, α is the *base world*, and β is the *selected world*.
>
> 1. For all antecedents A and base worlds α, A must be true in $f(A, \alpha)$.
> 2. For all antecedents A and base worlds α, $f(A, \alpha) = \lambda$ only if there is no world possible with respect to α in which A is true.
>
> [...] The informal truth conditions that were suggested above required that the world selected *differ minimally* from the actual world[2].
>
> [A theory of conditionals, pp.34-35]

λ nella citazione rappresenta il mondo assurdo. In sostanza, la funzione f prende il mondo di partenza α e l'antecedente del controfattuale, A e restituisce un mondo possibile, β, nel quale valutare il conseguente; nel caso in cui la funzione non riesca a individuare nessun mondo possibile, allora renderà λ, il mondo assurdo.

Ma quando un controfattuale del tipo $A > B$ (Stalnaker usa il simbolo $>$ per indicare il connettivo controfattuale) può essere valutato come vero?

---

[2]Oltre alla struttura modello, il nostro apparato semantico comprende anche una *funzione di selezione*, f, che prende come suoi valori una proposizione e un mondo possibile. La *s*-funzione seleziona, per ogni antecedente A, un mondo particolare in cui A è vero. [...] Userò la seguente terminologia per parlare di argomenti e valori delle *s*-funzioni: dove $f(A, \alpha) = \beta$, A è l'*antecedente*, α è il *mondo base* e β è il *mondo selezionato*.

1. Per tutti gli antecedenti A e mondi base α, A deve essere vero in $f(A, \alpha)$.
2. Per tutti gli antecedenti A e mondi base α, $f(A, \alpha) = \lambda$ solo se non c'è nessun mondo possibile rispetto ad α in cui A è vero.

[...] Le condizioni di verità informali che sono state suggerite sopra richiedevano che il mondo selezionato *differisca minimamente* dal mondo attuale. [*traduzione mia*]

> $A > B$ is true in α if B is true in $f(A, \alpha)$;
>
> $A > B$ is false in α if B is false in $f(A, \alpha)$[3].
>
> [A theory of conditionals, p.35]

Quindi il valore di verità assunto dall'enunciato conseguente nel mondo selezionato dalla funzione determina il valore di verità che l'intero enunciato controfattuale assume nel mondo di partenza.

Se in [150] sono contenute spiegazioni intuitive sui meccanismi di valutazione dei condizionali, [152] elenca una serie di definizioni che caratterizzano la semantica. Vediamone solo alcune e cominciamo con la definizione di struttura modello, già utilizzata senza esplicitarla in [150]:

> A **CQ** *model structure* (**CQ**ms) is a structure $M = \langle K, R, \lambda, D, D' \rangle$ where $\lambda \in K, K' = K - \{\lambda\}$ is a non-empty set, R is a binary reflexive relation on K', D is a function taking members α of D' into possibly empty sets D_α, and D' is a set disjoint from $\bigcup_{\alpha \in K'} D_\alpha$[4]
>
> [A semantic analysis of conditional logic, p.25]

Una struttura è quindi composta da un insieme di mondi, K, da una relazione di accessibilità, R, dal mondo assurdo, λ, da una funzione, D e da un insieme, D'.

Una volta definita la struttura modello, Thomason e Stalnaker, passando attraverso le nozioni di sequenza e valutazione, definiscono la *s*-funzione:

> A *sequence* σ on a morphology **M** and **QM**ms $\langle K, R, \lambda, D, D' \rangle$ is a function taking members of V_M (individual variables) x into members $\sigma(x)$ of D.

---

[3] $A > B$ è vero in α se B è vero in $f(A, \alpha)$;
$A > B$ è falso in α se B è falso in $f(A, \alpha)$. [*traduzione mia*]

[4] Una **CQ** *struttura modello* (**CQ**ms) è una struttura $M = \langle K, R, \lambda, D, D' \rangle$ dove $\lambda \in K, K' = K - \{\lambda\}$ è un insieme non vuoto, R è una relazione riflessiva binaria su K', D è una funzione che applica elementi α di K' in insiemi eventualmente vuoti D_α, e D' un insieme disgiunto da $\bigcup_{\alpha \in K'} D_\alpha$. [*trad. it. a cura di Claudio Pizzi:* [153], p.217]

> [...] A *valuation* of a morphology **M** on a **QM**ms $\langle K, R, \lambda, D, D' \rangle$ is a function v assigning, for each member α of K', *(i)* a value $v_\alpha(P)$ in $\{T, F\}$ to each 0-ary predicate letter P of **M**; *(ii)* a subset $v_\alpha(Q)$ of the cartesian product D^n to each n-ary predicate letter Q of **M**; *(iii)* to each individual constant a of **M** a member $v_\alpha(a)$ of D.
>
> [...] An *s-function* on a **QM**ms $M = \langle K, R, \lambda, D, D' \rangle$ and morphology **M** is a function f which assigns to each wff A, each $\alpha \in K'$, and each sequence σ on **M** and M a member $f(A, \alpha, \sigma)$ of K' meeting the following condition: for all A, α, and σ, if $f(A, \alpha, \sigma) \neq \lambda$, then $\alpha R f(A, \alpha, \sigma)$[5].
>
> [A semantic analysis of conditional logic, pp.26–27]

Quindi la funzione di selezione parte da un mondo, α, un enunciato (l'antecedente A) e una sequenza, σ, che, come abbiamo visto, associa variabili individuali a elementi di D e restituisce un altro mondo, connesso ad α attraverso la relazione di accessibilità R. Questo mondo è, tra i mondi accessibili da α in cui sia vero l'antecedente A, quello (l'unico) più simile ad α.

Una teoria fondata su presupposti molto simili è quella avanzata da David K. Lewis, che sarà presentata nel prossimo paragrafo.

---

[5]Una *sequenza* σ su una morfologia **M** e una **QM**ms $\langle K, R, \lambda, D, D' \rangle$ è una funzione che applica elementi di V_M (variabili individuali) x in elementi $\sigma(x)$ di D.

[...] Una *valutazione* di una morfologia **M** su una **QM**ms $\langle K, R, \lambda, D, D' \rangle$ è una funzione v che assegna, per ciascun elemento α di K', *(i)* un valore $v_\alpha(P)$ in $\{V, F\}$ a ciascuna lettera per predicati 0-adici P di **M**; *(ii)* un sottoinsieme $v_\alpha(Q)$ del prodotto cartesiano D^n a ciascuna lettera predicativa n-adica Q di **M**; *(iii)* a ciascuna costante individuale a di **M** un elemento $v_\alpha(a)$ di D.

[...] Una *s-funzione* su una **QM**ms $M = \langle K, R, \lambda, D, D' \rangle$ e una morfologia **M** è una funzione f che assegna a ciascuna fbf A, ciascun $\alpha \in K'$, e ciascuna sequenza σ su **M** e M un elemento $f(A, \alpha, \sigma)$ di K', che soddisfa la condizione seguente: per ogni A, α e σ, se $f(A, \alpha, \sigma) \neq \lambda$ allora $\alpha R f(A, \alpha, \sigma)$. [*trad. it. a cura di Claudio Pizzi:* [153], pp.218–219]

2.1.2 Le sfere di mondi di Lewis

Il più celebre libro di Lewis sull'argomento, *Counterfactuals* [100], parte dalla definizione dei connettivi che utilizza per esprimere i due diversi tipi di controfattualità che ha individuato:

- $\Box\!\!\rightarrow$: individua una forma più forte di controfattuale. $A\Box\!\!\rightarrow C$ significa: "Se fosse successo A sarebbe successo C", che indica che né A né C si sono verificati, ma il verificarsi di A avrebbe portato con sé il verificarsi di C;
- $\Diamond\!\!\rightarrow$: individua una forma più debole di controfattuale. $A\Diamond\!\!\rightarrow C$ significa: "Se fosse successo A sarebbe potuto succedere C", che indica che né A né C si sono verificati, ma il verificarsi di A avrebbe comportato che si desse almeno una possibilità che anche C si realizzasse.

Vediamo ora le definizioni di Lewis:

> My methods are those of much recent work in possible-world semantics for intensional logic. I shall introduce a pair of counterfactual conditional operators intended to correspond to the various counterfactual conditional constructions of ordinary language; and I shall interpret these operators by saying how the truth value at a given possible world of a counterfactual conditional is to depend on the truth values at various possible worlds of its antecedent and consequent. [...]
>
> $$\Box\!\!\rightarrow$$
>
> read as '*If it were the case that ___, then it would be the case that ...*', and
>
> $$\Diamond\!\!\rightarrow$$
>
> read as '*If it were the case that ___, then it might be the case that ...*'. [...]
>
> The two counterfactual operators are to be interdefinable as follows.

$$\phi\Diamond\!\!\rightarrow \psi =^{df} \sim(\phi\Box\!\!\rightarrow\sim\psi),$$
$$\phi\Box\!\!\rightarrow \psi =^{df} \sim(\phi\Diamond\!\!\rightarrow\sim\psi)\ ^{6}$$

[*Counterfactuals*, pp.1–2]

Come nel caso di Stalnaker, anche qui la prima cosa da definire è come vengano scelti i mondi possibili nei quali andare a controllare il valore di verità degli enunciati componenti il controfattuale.

Laddove Stalnaker aveva utilizzato una funzione di somiglianza, Lewis individua un ordinamento dei mondi basato su una somiglianza globale (*overall similarity*); tale ordinamento gli permette di raggrupparli in sfere centrate attorno al mondo nel quale si vuole valutare il controfattuale e ordinate secondo questo criterio: la sfera più interna contiene i mondi possibili che sono più simili al mondo di partenza e più ci si allontana dal centro più i mondi contenuti nelle sfere si differenziano dal mondo di partenza.

> Let \$ be an assignment to each possible world i of a set $\$_i$ of sets of possible worlds. Then \$ is called a (*centered*) *system of*

---

[6] I miei metodi sono quelli di molto del lavoro recente nella semantica a mondi possibili per la logica intensionale. Introdurrò un paio di operatori condizionali controfattuali che corrispondano alle varie costruzioni condizionali controfattuali del linguaggio ordinario; e interpreterò questi operatori dicendo come il valore di verità di un condizionale controfattuale in un dato mondo possibile dipenda dal valore di verità del suo antecedente e conseguente in vari mondi possibili. [...]

$$\Box\!\!\rightarrow$$

letto come '*Se si desse il caso che ___, allora si darebbe il caso che ...*', e

$$\Diamond\!\!\rightarrow$$

letto come '*Se si desse il caso che ___, allora potrebbe darsi il caso che ...*'. [...]
I due operatori controfattuali sono interdefinibili nel modo seguente.

$$\phi\Diamond\!\!\rightarrow \psi =^{df} \sim(\phi\Box\!\!\rightarrow\sim\psi),$$
$$\phi\Box\!\!\rightarrow \psi =^{df} \sim(\phi\Diamond\!\!\rightarrow\sim\psi)$$

[*traduzione mia*]

> *spheres*, and the members of each $\$_i$ are called *spheres* around i, if and only if, for each world i, the following conditions hold.
>
> (C) $\$_i$ is *centered on* i; that is, the set $\{i\}$ having i as its only member belongs to $\$_i$.
>
> (1) $\$_i$ is *nested*; that is, whenever S and T belong to $\$_i$, either S is included in T or T is included in S.
>
> (2) $\$_i$ is *closed under unions*; that is, whenever S is a subset of $\$_i$ and $\bigcup S$ is the set of all worlds j such that j belongs to some member of S, $\bigcup S$ belongs to $\$_i$.
>
> (3) $\$_i$ is *closed under (nonempty) intersections*; that is, whenever S is a nonempty subset of $\$_i$ and $\bigcup S$ is the set of all worlds j such that j belongs to every member of S, $\bigcup S$ belongs to $\$_i$[7].
>
> [*Counterfactuals*, pp.14–15]

Le condizioni appena enunciate servono a descrivere l'ordinamento secondo il quale sono disposte le sfere.

A questo punto, non è difficile intuire quando un enunciato controfattuale sarà giudicato vero: premesso che si prendono in considerazione soltanto antecedenti "ragionevoli" (cioè che siano veri in almeno un mondo di queste sfere), se in tutti i mondi appartenenti alla sfera più prossima in cui l'antecedente è vero, anche il conseguente è vero, allora tutto l'enunciato controfattuale è pure vero:

---

[7] Sia $\$$ un assegnamento a ciascun mondo possibile i di un insieme $\$_i$ di insiemi di mondi possibili. Allora $\$_i$ è detto un *sistema di sfere* (*centrato*), e i membri di ciascun $\$_i$ sono detti *sfere* attorno a i se e solo se, per ogni mondo i, valgono le seguenti condizioni.

(C) $\$_i$ è *centrato su* i; ossia, l'insieme $\{i\}$ che ha i come suo unico membro appartiene a $\$_i$.

(1) $\$_i$ è *annidato*; ossia, ogni volta che S e T appartengono a $\$_i$, o S è incluso in T oppure T è incluso in S.

(2) $\$_i$ è *chiuso rispetto alle unioni*; ossia, ogni volta che S è un sottoinsieme di $\$_i$ e $\bigcup S$ è l'insieme di tutti i mondi j tali che j appartiene a qualche membro di S, $\bigcup S$ appartiene a $\$_i$.

(3) $\$_i$ è *chiuso rispetto alle intersezioni (non vuote)*; ossia, ogni volta che S è un sottoinsieme non vuoto di $\$_i$ e $\bigcup S$ è l'insieme di tutti i mondi j tali che j appartiene a ogni membro di S, $\bigcup S$ appartiene a $\$_i$. [*traduzione mia*]

> $\phi \Box\!\!\rightarrow \psi$ is true at a world i (according to a system of spheres \$) if and only if either
>
> 1. no ϕ-world belongs to any sphere S in $\$_i$, or
> 2. some sphere S in $\$_i$ does contain at least one ϕ-world, and $\phi \supset \psi$ holds at every world in S[8].
>
> [*Counterfactuals*, p.16]

La prima possibilità si riferisce ai controfattuali banalmente veri, quelli il cui antecedente non è vero in nessuno dei mondi possibili accessibili dal mondo di partenza (e corrisponde allo stratagemma del mondo assurdo λ nel sistema di Stalnaker). La seconda possibilità equivale a dire che, se esiste almeno un mondo accessibile a quello di partenza in cui l'antecedente è vero, allora perché il controfattuale sia vero il conseguente deve essere vero in tutti i mondi della sfera più interna in cui sia vero l'antecedente.

Un'analisi sostanzialmente uguale viene fornita anche in [105], nel quale si ritrova anche la definizione di un modello assiomatico per la logica dei controfattuali:

> In general, we may define a *model* as any quadruple $\langle I, R, \leq, [\ ] \rangle$ such that:
>
> 1. I is a nonempty set (regarded as playing the role of the set of worlds);
> 2. R is a binary relation over I (regarded as the accessibility relation);
> 3. $\leq$ assigns to each i in I a weak ordering $\leq_i$ of I (regarded as the comparative similarity ordering of worlds from the standpoint of i) such that whenever $j \leq_i k$, if iRk then iRj;

---

[8] $\phi \Box\!\!\rightarrow \psi$ è vero in un mondo i (secondo un sistema di sfere \$) se e solo se o

1. nessun ϕ-mondo appartiene ad alcuna sfera S in $\$_i$, o
2. qualche sfera S in $\$_i$ contiene almeno un ϕ-mondo e $\phi \supset \psi$ vale in ogni mondo in S.

[*traduzione mia*]

4. [ ] assigns to each sentence A a subset $[A]$ of I (regarded as the set of worlds where A is true);
5. $[-A]$ is $I - [A]$, $[A\&B]$ is $[A] \cap [B]$, and so on;
6. $[A \prec B]$ is $\{i \epsilon I$: for some j in $[A]$ such that iRj, there is no k in $[B]$ such that $k \leq_i j\}$[9]

[Counterfactuals and Comparative Possibility, p.26]

Gli elementi del modello sono dunque l'insieme di mondi, I, la relazione di accessibilità, R, un'operazione, $\leq_i$, che ordina i mondi secondo la somiglianza a un mondo $i \in I$ e un'operazione [ ], che ricava dei sottoinsiemi all'interno di I nei quali valgano determinati enunciati.

Una differenza evidente tra l'approccio di Lewis e quello di Stalnaker è dato dal fatto che, mentre Stalnaker deve "andare a vedere come stanno le cose" in un solo mondo, quello più simile al mondo di partenza nel quale l'antecedente è vero, Lewis ha davanti a sé un insieme di mondi. Questo perché Lewis rifiuta la cosiddetta Assunzione di Unicità (*Uniqueness Assumption*), che determina il fatto che esista sempre un solo e unico mondo che sia il più simile al mondo reale secondo un certo rispetto. L'Assunzione di Unicità

[9]In generale, possiamo definire un *modello* come una qualsiasi quadrupla $\langle I, R, \leq, [\ ] \rangle$ tale che:

1. I è un insieme non-vuoto (considerato come tale da giocare il ruolo dell'insieme di mondi);
2. R è una relazione binaria su I (considerata come la relazione di accessibilità);
3. $\leq$ assegna a ciascun i in I un ordinamento debole $\leq_i$ di I (inteso come l'ordinamento di somiglianza comparativa dei mondi dal punto di vista di i) tale che ogniqualvolta $j \leq_i k$, se iRk allora iRj;
4. [ ] assegna a ciascun enunciato A un sottoinsieme $[A]$ di I (inteso come l'insieme di mondi a cui A è vero);
5. $[-A]$ è $I - [A]$, $[A\&B]$ è $[A] \cap [B]$, e così via;
6. $[A \prec B]$ è $\{i \epsilon I$: per qualche j in $[A]$ tale che iRj, non c'è nessun k in $[B]$ tale che $k \leq_i j\}$.

[*trad. it. a cura di Claudio Pizzi:* [104], p.258].

deriva dal principio del Terzo Escluso Condizionale (*Conditional Excluded Middle*):

$$(\phi \mathbin{\Box\!\!\rightarrow} \psi) \vee (\phi \mathbin{\Box\!\!\rightarrow} \neg\psi)$$

Ovvero, in un mondo o un enunciato ϕ implica controfattualmente un enunciato ψ, oppure implica la sua negazione $\neg\psi$ e quindi tra due mondi che differiscano per almeno un enunciato, ce ne sarà sempre uno che è più simile a un terzo mondo rispetto all'altro. Per apprezzare la differenza, si prenda uno dei classici esempi, che può essere fatto risalire a Nelson Goodman:

Se New York fosse in Georgia, allora New York sarebbe nel Sud (2.1)

Secondo il principio del terzo escluso condizionale, o è vero che se New York fosse in Georgia, allora NY sarebbe nel Sud, oppure è vero che se NY fosse in Georgia, allora NY non sarebbe nel Sud. La conseguenza dell'affermazione di questo principio è però che debba essere possibile discriminare tra un mondo in cui NY si trovasse all'interno dei confini dell'attuale Georgia e un mondo in cui la Georgia si trovasse attorno al punto dove si trova NY nel mondo reale. Secondo Stalnaker è sempre possibile, nelle circostanze specifiche, decidere quale di questi due mondi debba essere considerato più simile al mondo reale, mentre, nell'interpretazione di Lewis, questi due mondi si troverebbero nella stessa sfera.

Questa differenza, che formalmente si traduce nell'assunzione di unicità (presente nel sistema di Stalnaker e assente in quello di Lewis), a livello concettuale è piuttosto rilevante, perché per Lewis non solo esistono diversi sensi in cui i mondi possono essere somiglianti, ma anche rispetto a un senso specifico esistono dei casi in cui è impossibile decidere quale di due mondi sia più somigliante a un terzo.

Riteniamo che esistano e siano anche piuttosto frequenti nel ragionamento situazioni in cui non si hanno sufficienti elementi per essere in grado di dire se sia vero un controfattuale o la sua negazione; per questo motivo giudichiamo più plausibile la scelta di Lewis di valutare il controfattuale su più mondi, indistinguibili l'uno dall'altro rispettivamente alla loro somiglianza al mondo di partenza.

2.1.3 La *situation semantics* e i controfattuali

La *situation semantics* è un *framework* semantico alternativo alle teorie estensionali e alle semantiche a mondi possibili, sviluppato a partire dai primi anni '80 e nato da un'idea di John Perry e Jon Barwise, che lo hanno esposto in maniera sistematica in [6].

Il concetto da cui parte è quello di *situazione*. Una situazione è una parte (o un aspetto) della realtà, più precisamente quella parte che ci è accessibile in quanto agenti limitati e sulla quale ragioniamo; può essere una regione spazio-temporalmente connessa, un contesto di predicazione, una collezione di condizioni di sfondo per un vincolo e altro ancora. Una situazione è modellata da un insieme di fatti atomici positivi e negativi ed è specificata dalla collezione di fatti che soddisfa. Analogamente, un tipo di situazione (*type of situation*) è specificato dai fatti e dai tipi di fatti (*types of fact*) che le sue istanze soddisfano. Le situazioni servono per rendere conto del ruolo cruciale che hanno il contesto e le assunzioni di sfondo nei comportamenti. Di fronte a una proposizione del tipo:

$$s \models \sigma$$

dove s è una situazione e σ è una proposizione, maggiore è il contenuto informativo di σ, maggiore è l'informazione che abbiamo anche sulla situazione s. Consideriamo il seguente esempio, proposto da Keith Devlin:

Jon Barwise è stato il primo direttore del CSLI (2.2)

Se questo enunciato compare in un discorso tra parlanti che sappiano perfettamente chi sia Jon Barwise e che tipo di istituto sia il CSLI, la proposizione sarà formalizzata nel seguente modo e avrà lo scopo di comunicare la semplice informazione che proprio Jon Barwise (che l'interlocutore conosce) è stato il primo direttore del CSLI (che l'interlocutore conosce):

$$s \models \langle\langle \text{primo-direttore-di, JON BARWISE, CSLI, } 1\rangle\rangle$$

e tutta l'informazione necessaria per comprendere l'affermazione è già presente nelle assunzioni di sfondo.

Al contrario, se l'affermazione viene rivolta a un interlocutore che non abbia mai sentito parlare né di Barwise né del CSLI, comunicherà qualcosa

di diverso: che esiste una persona a che si chiama Jon Barwise e che questa persona è stata il primo direttore di un istituto che si chiama CSLI. La proposizione assumerà questa forma:

$$s \models \langle\langle \text{di-nome}, a, \text{JON BARWISE}, 1\rangle\rangle \wedge \langle\langle \text{umano}, a, 1\rangle\rangle \wedge \langle\langle \text{maschio}, a, 1\rangle\rangle \wedge \langle\langle \text{primo-direttore-di}, a, CSLI, 1\rangle\rangle$$

Naturalmente l'interlocutore continuerà a non sapere qual è la professione di Jon Barwise o qual è la sua nazionalità, o che tipo di istituto sia il CSLI, però dall'affermazione e da alcune delle conoscenze di sfondo che possiede può inferire che Barwise è un essere umano (nella sua esperienza gli istituti sono diretti da esseri umani) e un maschio (se l'interlocutore ha qualche cognizione sui nomi propri inglesi).

Come si vede, a seconda di quali siano le conoscenze di sfondo in cui viene recepita un'asserzione, il contenuto informativo di tale asserzione varia.

Le situazioni possono essere classificate in base a proprietà, relazioni e stati di fatto complessi; per esempio, due tipi di situazione possono essere legati da un vincolo che fa sì che, quando si presenta una situazione del primo tipo, se ne presenta anche una del secondo.

Le situazioni possono poi essere classificate direttamente, grazie allo stato di fatto che supportano, oppure indirettamente, attraverso ciò che significano, ossia attraverso il tipo di situazione che implicano relativamente a un dato vincolo. Inoltre, tra i modi di classificazione indiretta, si distinguono le classificazioni che avvengono grazie al contenuto informativo, ossia quelle relative alle leggi di natura e alle generalizzazioni, oppure quelle che vengono fatte sulla base del contenuto intenzionale, ossia quelle relative a convenzioni, abitudini, piani e altre regole comportamentali umane.

Per quanto riguarda i controfattuali, Barwise in [5] traccia un'interessante distinzione tra l'aggettivo "controfattuale" e l'aggettivo "congiuntivo", entrambi riferiti ai condizionali. "Controfattuale" si applica alle asserzioni, mentre "congiuntivo" si applica agli enunciati, dove l'enunciato è un ente dotato di significato, significato che ha delle connessioni con il valore di verità delle asserzioni che si fanno attraverso gli enunciati.

Barwise riconosce l'importanza dell'elemento contestuale nell'interpretazione dei controfattuali e individua nella volontà di eliminare l'elemento con-

testuale, "traducendo" gli enunciati in altri enunciati dove il contesto non giochi alcun ruolo, l'errore comune alla maggior parte degli approcci alternativi al suo.

Su questa questione riteniamo che Barwise sia nel giusto: i tentativi di operare una decontestualizzazione assoluta sono destinati a fallire, poiché ogni enunciato dipende da un numero infinito di parametri che sono *solo parzialmente* esplicitabili; una volta accettato il carattere intrinsecamente contestuale degli enunciati, è possibile compiere operazioni sui contesti per ricavarne informazione.

Un altro aspetto rilevante di questo approccio risiede nell'impegno a dare un'unica sistematizzazione che valga sia per il condizionale indicativo che per quello congiuntivo, sia per il condizionale matematico che per quello controfattuale.

Barwise presenta due tipologie di condizionali, quelli generici e quelli specifici; questi ultimi possono essere visti come istanze dell'interpretazione dei primi, così:

> I propose to interpret general conditional statements as describing 'parametric constraints' and specific conditionals as describing instances of the constraints where parameters are fixed[10].
>
> [The Situation in Logic-II: Conditionals and Conditional Information, p.4]

Secondo Barwise e Perry, l'informazione che un condizionale – e quindi anche un controfattuale – comunica è l'esistenza di un determinato vincolo condizionale sul mondo. I condizionali più generici indicano l'esistenza di un vincolo generale, che vale tra tipi di situazioni, mentre i condizionali specifici individuano le singole situazioni legate dal vincolo. Questo vale anche, più in generale, per tutti i tipi di asserzione:

---

[10]Propongo di interpretare le asserzioni condizionali generali come descrizioni di 'vincoli parametrici' e i condizionali specifici come descrizioni di istanze dei vincoli in cui i parametri sono fissati. [*traduzione mia*]

> Meaning consists in constraints between types of situations, and it is such constraints that allow a situation to contain information[11].
>
> [The Situation in Logic-II: Conditionals and Conditional Information, p.21]

Tornando all'interpretazione dei condizionali, Barwise mette in evidenza l'importanza delle assunzioni di sfondo e del fattore contestuale, sottolineando ancora una volta la componente parametrica dei condizionali generici:

> Thus, the interpretation of a general conditional statement is a parametric constraint $C \mid \mathbf{B}$, where $\mathbf{B}$ is a parameter anchored to the prevailing background, and where C is $S \Rightarrow S'$, these types being the interpretations of the antecedent and consequent, respectively. As such, this will not provide a complete proposition, but only a parametric proposition, a proposition relative to the background conditions B – the proposition that $C \mid B$ is actual. [...] This makes the exact information content of a statement of a general conditional highly context dependent, which seems right[12].
>
> [The Situation in Logic-II: Conditionals and Conditional Information, p.27]

Quindi, anche le leggi generali, che legano due tipi di situazione, sono valide solo date certe condizioni di sfondo, ossia in un contesto.

---

[11] Il significato consiste di vincoli tra tipi di situazioni e sono tali vincoli che permettono a una situazione di contenere informazione. [*traduzione mia*]

[12] Così, l'interpretazione di un'asserzione condizionale generale è un vincolo parametrico $C \mid \mathbf{B}$, dove $\mathbf{B}$ è un parametro ancorato allo sfondo prevalente e dove C è $S \Rightarrow S'$, dove questi tipi sono le interpretazioni dell'antecedente e del conseguente, rispettivamente. Di per sé questo non fornisce una proposizione completa, ma solo una proposizione parametrica, una proposizione relativa alle condizioni di sfondo B – la proposizione che $C \mid B$ è reale. [...] Questo rende l'esatto contenuto di informazione di un'asserzione di un condizionale generale fortemente dipendente dal contesto, cosa che sembra essere corretta. [*traduzione mia*]

Prendiamo l'esempio, anch'esso dovuto a Devlin, di un vincolo generale secondo il quale se un uovo viene lasciato cadere si rompe; il legame sembra semplice e ragionevole ma, a esaminarlo bene, ci si rende conto che non è un legame universale, ma presuppone una certa dose di assunzioni di sfondo: che l'uovo a cui si applica si trovi nel campo gravitazionale terrestre, che venga lasciato cadere da una certa distanza minima, che non sia stato bollito (altrimenti la suddetta distanza minima aumenta), la superficie sulla quale viene lasciato cadere deve avere certe caratteristiche di durezza e anelasticità ecc.

Le assunzioni di sfondo nella maggior parte dei casi non vengono esplicitate, ma devono tutte valere affinché il vincolo valga; diventa necessario esplicitarle solamente quando un vincolo fino a quel momento affidabile porta a errori in una situazione che è un'istanza di un tipo di situazione in cui normalmente il vincolo vale. Come detto in precedenza, i condizionali specifici sono il risultato dell'operazione che consiste nel fissare i valori dei parametri costituenti i rispettivi condizionali generici:

> The speaker is talking about a specific, highly limited, situation, say s_u. Usually just a few things and some relations between them are involved. He is saying that this is a situation where a conditional constraint $S \Rightarrow S' \mid \mathbf{B}$ applies, where $\mathbf{B}$ is anchored to the background conditions. S is the interpretation of ϕ, S' is the interpretation of ψ. Thus, his utterance will be informational relative to B if there is an anchor f for the parameters of B such that $s_u : B(f)$, and if he has the information, relative to B, that $S \Rightarrow S' \mid \mathbf{B}$ is actual. He may have such information simply by being in that type of situation and knowing how things work there. The propositional content of his utterance is just that $S(f) \Rightarrow S'(f)$ is actual[13].

[13]Il parlante sta parlando di una situazione specifica, fortemente limitata, diciamo s_u. Abitualmente solamente poche cose e poche relazioni tra di esse sono coinvolte. Sta dicendo che questa è una situazione in cui si applica il vincolo condizionale $S \Rightarrow S' \mid \mathbf{B}$, dove $\mathbf{B}$ è ancorato alle condizioni di sfondo. S è l'interpretazione di ϕ, S' è l'interpretazione di ψ. Quindi, la sua espressione sarà informativa relativamente a B se c'è un'ancora f per i parametri di B tale che $s_u : B(f)$ e, se ha questa informazione, relativa a B, che

> [The Situation in Logic-II: Conditionals and Conditional Information, p.28]

Il condizionale specifico dice che sussiste il vincolo specifico $S(f) \Rightarrow S'(f)$, che è un'istanza del vincolo generale $S \Rightarrow S'$, valido rispetto alle condizioni di sfondo **B**.

Un lavoro che prende le mosse da questa analisi accurata ma assolutamente non formale e si propone di fornire una base logica formale e uno sguardo alle applicazioni (in intelligenza artificiale) è l'articolo di Wayne Wobcke [162].

Un primo elemento di differenza tra [5] e [162] è che, laddove Barwise descriveva i vincoli in relazione ad assunzioni di sfondo, Wobcke considera che essi valgano relativamente a un tipo "base" di situazione, contestualmente determinata. Wobcke assume inoltre l'esistenza di una gerarchia tra le situazioni e il posizionamento del tipo "base" di situazione determina lo sfondo rispetto al quale il condizionale viene valutato. Anche in questo caso, però, la determinazione della situazione base è un problema di ordine pragmatico e non logico.

In generale, per Wobcke, la determinazione di un vincolo avviene in questo modo:

> On our account, a constraint $A \Rightarrow B$ holds at some type of situation σ if B holds in the most general type of situation subsumed by σ that satisfies A[14].
>
> [A Theory of Conditionals based on Hierarchies of Situations, p.6]

Per parlare di condizionali, Wobcke recupera dalla *situation semantics* la nozione di "opzione significativa" (*meaningful option*), che in [6] era stata definita solo relativamente a situazioni, la applica ai tipi di situazione e la

---

$S \Rightarrow S' \mid \mathbf{B}$ è reale. Può avere questa informazione semplicemente trovandosi in quel tipo di situazione e sapendo come funzionano lì le cose. Il contenuto proposizionale della sua espressione è solo che $S(f) \Rightarrow S'(f)$ è reale. [*traduzione mia*]

[14]Nel nostro resoconto, un vincolo $A \Rightarrow B$ sussiste in un tipo di situazione σ se B sussiste nel tipo di situazione più generale inclusa in σ che soddisfi A. [*traduzione mia*]

reinterpreta alla luce della gerarchia di situazioni: la situazione σ' è detta "opzione significativa" della situazione σ rispetto al vincolo $\Phi \Rightarrow \Psi$ se il tipo di fatto Φ sussiste in σ, il tipo di fatto Ψ sussiste in σ' e vale il vincolo $\Phi \Rightarrow \Psi$. Le opzioni significative di un tipo di situazione sono i tipi di situazione che stanno sotto di essa nella gerarchia.

Anche nel caso dei condizionali, si ragiona in maniera analoga a come si era fatto per i vincoli in generale, aggiungendo però all'apparato logico una funzione che ordini i tipi di situazione in una gerarchia:

> [...] an SC interpretation includes a selection function f which for each type of situation σ and fact formula Φ, defines a most general subtype of σ which satisfies Φ (if there is one). Intuitively, the selection function specifies the most normal course of events given the information Φ. [...] More formally, these desired interpretations are those in which the types of situations accessible to σ that satisfy Φ are arranged in a partial pre-order, with the selection function choosing one of the minimal elements in this order[15].
>
> [A Theory of Conditionals based on Hierarchies of Situations, p.20]

Anche in questo caso, l'elemento pluralistico pare ineliminabile, poiché, ancora una volta, il tipo di situazione su cui viene valutato il condizionale viene selezionato da una funzione che non è data una volta per tutte nella logica, ma presenta un indubbio carattere pragmatico.

---

[15] [...] una SC interpretazione include una funzione di selezione f che per ogni tipo di situazione σ e formula di fatto Φ, definisce un sottotipo più generale di σ che soddisfa Φ (se ce n'è uno). Intuitivamente, la funzione di selezione specifica il corso di eventi più naturale data l'informazione Φ. [...] Più formalmente, queste interpretazioni desiderate sono quelle in cui i tipi di situazioni accessibili a σ che soddisfano Φ sono sistemati secondo un pre-ordine parziale, con la funzione di selezione che sceglie uno degli elementi minimali in questo ordine. [*traduzione mia*]

2.2 Approcci "consequenzialisti"

Anche gli approcci consequenzialisti hanno il loro punto di partenza nel test di Ramsey [134], ma in questo caso il giudizio sul controfattuale viene compiuto piuttosto attraverso la verifica che sussista un certo tipo di connessione tra antecedente e conseguente, tale per cui il conseguente sia deducibile da un insieme di premesse che contenga l'antecedente.

Il problema più pressante per i teorici che si riconoscono in questo tipo di analisi è dunque quello di capire quali altre premesse, oltre all'antecedente, debba contenere l'insieme dal quale si dovrebbe dedurre il conseguente.

2.2.1 Goodman e la cotenibilità

La prima posizione che andremo ad analizzare all'interno del paradigma consequenzialista è quella di Nelson Goodman, che è anche stata una delle prime in ordine di tempo a essere proposta.

Già nei primi passaggi del suo celebre articolo [78], Goodman caratterizza la sua analisi in un modo che la contrappone agli approcci che sorgeranno in seguito all'interno degli studi sulle semantiche a mondi possibili:

> [...] the truth of statements of this kind [...] depends not upon the truth or falsity of the components but upon whether the intended connection obtains[16].
>
> [The problem of counterfactual conditionals, p.10]

Questo è un importante punto di distinzione rispetto agli approcci verofunzionali: si afferma che non è possibile in alcun modo costruire una tavola di verità per i controfattuali, ma che la loro essenza va ricercata nel tipo di legame che stabiliscono tra antecedente e conseguente.

Il secondo compito di Goodman è quindi quello di scoprire in che cosa consiste questa connessione. Essa viene identificata con il legame che sussiste tra il conseguente del controfattuale (C) da una parte e l'antecedente (A) e

[16] [...] la verità di asserzioni di questo tipo non dipende dalla verità o falsità dei componenti ma dal fatto che si dia o meno la connessione intesa. [*traduzione mia*].

un insieme di asserzioni esprimenti le condizioni rilevanti che devono valere affinché C consegua da A.

Goodman parte allora alla ricerca di un criterio per identificare tali condizioni rilevanti:

> It seems that we must elaborate our criterion still further, to characterize a counterfactual as true if there is some set S of true statements such that $A \cdot S$ is self-compatible and leads by law to the consequent, while there is no such set S' such that $A \cdot S'$ is self-compatible and leads by law to the negate of the consequent[17].
>
> [The problem of counterfactual conditionals, pp.15–16]

Tuttavia è necessario escludere la possibilità che sia S a discriminare tra C e la sua negazione e specificare che deve essere *la congiunzione di A con S* a essere compatibile con C e non con $-C$:

> Our rule reads that a counterfactual is true if and only if there is some set S of true sentences such that S is compatible with C and with $-C$, and such that $A \cdot S$ is self-compatible and leads by law to C; while there is no set S' compatible with C and with $-C$ and such that $A \cdot S'$ is self-compatible and leads by law to $-C$[18].
>
> [The problem of counterfactual conditionals, pp.16–17]

Aggiungendo pezzo dopo pezzo, Goodman arriva a specificare completamente i vincoli che deve rispettare un controfattuale per essere giudicato vero, introducendo il concetto di *cotenibilità*:

[17] Sembra che dobbiamo elaborare ancora meglio il nostro criterio, per caratterizzare un controfattuale come vero se c'è qualche insieme S di asserzioni vere tali che $A \cdot S$ è coerente e porta per legge al conseguente, mentre non c'è un insieme S' tale che $A \cdot S'$ è coerente e porta per legge alla negazione del conseguente. [*traduzione mia*]

[18] La nostra regola dice che un controfattuale è vero se e solo se c'è qualche insieme S di enunciati veri tali che S è compatibile con C e con $-C$ e tale che $A \cdot S$ è coerente e porta per legge a C; mentre non c'è un insieme S' compatibile con C e con $-C$ e tale che $A \cdot S'$ è coerente e porta per legge a $-C$. [*traduzione mia*]

> S, in addition to satisfying the other requirements already laid down, must not be merely compatible with A but 'jointly tenable' or *cotenable* with A. A is contenable with S, and the conjunction $A \cdot S$ self-cotenable, if it is not the case that S would not be true if A were[19].
>
> [The problem of counterfactual conditionals, p.18]

Dalla definizione di cotenibilità in termini di legame controfattuale tra A e S, è facile intuire la ragione che ha spinto Goodman a parlare del *problema* dei controfattuali poiché, come si può notare, la nozione di cotenibilità, sulla quale Goodman basa l'assegnazione del valore di verità ai controfattuali è a sua volta definita per mezzo di un controfattuale, con l'ovvia conseguenza di intrappolare l'argomento in un regresso all'infinito.

Lo stesso Goodman l'aveva di fatto già segnalato nel suo articolo:

> But in order to determine whether or not a given S is cotenable with A, we have to determine whether or not the counterfactual 'If A were true, then S would not be true' is itself true. [...] Thus we find ourselves involved in an infinite regressus or a circle; for cotenability is defined in terms of counterfactuals, yet the meaning of counterfactuals is defined in terms of cotenability[20].
>
> [The problem of counterfactual conditionals, p.19]

Un aspetto a nostro avviso molto importante dell'analisi di Goodman è l'aver riconosciuto che la ricerca dei fatti rilevanti che costituiscono il solo tribunale

---

[19] S, oltre a soddisfare gli altri requisiti già elencati, non solo deve essere compatibile con A ma 'unitamente tenibile' o *cotenibile* con A. A è cotenibile con S e la congiunzione $A \cdot S$ auto-cotenibile, se non si dà il caso che S non sia vero quando A lo è. [*traduzione mia*]

[20] Ma per poter determinare se un dato S sia o meno cotenibile con A, dobbiamo determinare se il controfattuale 'Se A fosse vero, allora S non sarebbe vero' sia esso stesso vero o meno. [...] Così ci troviamo avvolti in un regresso all'infinito o in un circolo; poiché la cotenibilità è definita in termini di controfattuali, e ancora il significato dei controfattuali è definito in termini di cotenibilità. [*traduzione mia*]

per la valutazione di un controfattuale è più centrale rispetto alla ricerca di criteri di somiglianza per la classificazione di mondi possibili.

Se l'analisi di Goodman sotto un certo punto di vista sembra arenarsi sul regresso all'infinito, d'altra parte egli ha pur sempre il merito di aver segnalato l'importanza della conoscenza di sfondo (rappresentata dagli enunciati cotenibili) nella valutazione del controfattuale, aprendo la strada a una linea di ricerca di criteri alternativi alla cotenibilità e non coreferenziali.

2.2.2 La teoria inferenzialista: Kvart

Un approccio che sembra prendere le mosse da assunti molto simili a quelli analizzati da Goodman è quello sostenuto da Igal Kvart, il quale descrive il suo approccio come metalinguistico e basato sulle nozioni di probabilità oggettiva e di rilevanza causale. Il lavoro in cui la sua analisi dei controfattuali è sviluppata nei minimi dettagli è il libro *A Theory of Counterfactuals* [92], ma nell'articolo [93] si trovano espresse chiaramente le linee-guida della sua posizione.

Anche per lui la questione centrale è quella di determinare quali sono le asserzioni che, unite all'antecedente, permettono di ricavare il conseguente. Tali asserzioni saranno identificate da una funzione f, che viene definita appunto da Kvart "funzione delle premesse implicite".

Ecco come in [93] presenta lo schema inferenziale per i controfattuali:

> Thus, using the sign '$\rightarrow$' for the logical consequence relation, a counterfactual $A > B$ is thus true if and only if
>
> $$f(A > B, \ldots) \cup \{A\} \rightarrow B$$
>
> thus manifesting what is to be called the *Inferential Schema* for counterfactuals. This schematic characterization of truth-conditions for counterfactuals therefore reduces the problem to the determination of the function f. This function will be called the *implicit premises function* (for short: **i.p. function**), and its

> values for fixed arguments will be called the **implicit premises** (for those arguments)[21].
>
> [Counterfactuals, pp.139–140]

Il ruolo della funzione delle premesse implicite è quello di descrivere gli eventi che non subiscono alcuna influenza in seguito al passaggio dallo stato di cose in cui vale $\neg A$ (quello di partenza) a quello in cui vale A.

L'analisi di Kvart parte da quei controfattuali che lui stesso definisce della *divergenza naturale* (i cosiddetti *n.d. counterfactuals* da *natural divergence*), ossia quei controfattuali che presuppongono che il cambiamento nello stato dei fatti ipotizzato dal controfattuale abbia luogo nel lasso temporale che va dal verificarsi dell'evento di cui parla l'antecedente (t_A) al verificarsi dell'evento di cui parla il conseguente (t_B), mentre tutti gli eventi che si sono verificati prima di t_A restano del tutto indipendenti dall'ipotesi controfattuale.

Il motivo della scelta di partire proprio da questo tipo di controfattuali è da ascriversi al fatto che essi sono quelli più largamente utilizzati nel ragionamento pratico, qualora per esempio si voglia parlare delle azioni umane. Inoltre, secondo Kvart, i problemi maggiori che possono presentarsi nell'analisi dei controfattuali sono già presenti nel tipo della divergenza naturale e l'estensione dell'analisi anche ai casi diversi non è particolarmente problematica.

Perché i controfattuali del tipo della divergenza naturale siano veri, secondo Kvart è necessario stabilire che l'evento espresso dall'antecedente, la storia del mondo precedente al verificarsi dell'evento-antecedente e gli

---

[21] Così, utilizzando il simbolo '$\rightarrow$' per la relazione di conseguenza logica, un controfattuale $A > B$ è quindi vero se e solo se

$$f(A > B, \ldots) \cup \{A\} \rightarrow B$$

manifestando in questo modo quello che viene detto *Schema Inferenziale* per i controfattuali. Questa caratterizzazione schematica delle condizioni di verità per i controfattuali si riduce in questo modo al problema della determinazione della funzione f. Questa funzione sarà detta *funzione delle premesse implicite* (in breve: **p.i. funzione**) e i suoi valori per argomenti fissati saranno detti **premesse implicite** (per quegli argomenti). [*traduzione mia*]

eventi verificatisi nel periodo intercorrente tra t_A e t_B che non siano influenzati (o siano influenzati solo positivamente) dal verificarsi di A, tutti questi elementi insieme portino, grazie alle leggi di natura, al verificarsi dell'evento-conseguente B.

Resta dunque da capire come possano esser ricavati questi eventi compresi nel periodo $t_A - t_B$ che non subiscono l'influenza negativa di A; tali eventi costituiscono alcune[22] delle premesse implicite della funzione $f(A > B) \ldots$, che può ora essere riscritta come $f(A, t_B)$.

Per identificare tali premesse, Kvart introduce i cosiddetti *irril-semifattuali* (*irrel-semifactuals*), semifattuali irrilevanti, nel senso che il loro antecedente è irrilevante ai fini del verificarsi del conseguente e *p.p.-semifattuali* (*p.p.-semifactuals*), semifattuali puramente positivamente rilevanti, nel senso che l'unica influenza che il loro antecedente può avere sul conseguente è un'influenza positiva, ossia che ne aumenta la probabilità:

> [...] *irrel-semifactuals* are semifactuals whose antecedent-events are causally irrelevant to their consequent-events, and *p.p.-semifactuals* – semifactuals whose antecedent-events are purely positively causally relevant to their consequent-events. [...] the roles of irrel-semifactuals and p.p.-semifactuals is to determine (via their consequents) the portions of the actual course-of-events (in (t_A, t_B)) which are to constitute the background on which the effects of the transition from $\sim A$-to-A (for an antecedent A) are to be evaluated[23].
>
> [Counterfactuals, p.149]

[22] Gli altri costituenti della funzione sono le leggi di natura e lo stato del mondo anteriore ad A.

[23] [...] gli *irril-semifattuali* sono semifattuali i cui eventi-antecedente sono causalmente irrilevanti per i loro eventi-conseguente e i *p.p.-semifattuali* – semifattuali i cui eventi-antecedente sono puramente positivamente causalmente rilevanti per i loro eventi-conseguente. [...] i ruoli degli irril-semifattuali e dei p.p.-semifattuali è di determinare (attraverso i loro conseguenti) le porzioni del corso-di-eventi reale (in (t_A, t_B)) che costituiscono lo sfondo sul quale vengono valutati gli effetti della transizione da $\sim A$-ad-A (per un antecedente A). [*traduzione mia*]

In seguito decide di ammettere nell'insieme delle premesse implicite anche i conseguenti dei cosiddetti *n.r.-semifattuali*, i semifattuali negativamente rilevanti, ossia i semifattuali il cui antecedente ostacola il verificarsi del conseguente, purché essi siano veri, giungendo così alla seguente analisi:

> A counterfactual $A > B$ (of the n.d.type) is true if and only if
>
> $\{A\} \cup W_A \cup \{$the consequents of true semifactuals $A > C$ with $t_C \subseteq (t_A, t_B)\}$ $-L\rightarrow B$[24]
>
> [Counterfactuals, p.153]

dove W_A è lo stato del mondo prima del verificarsi di A e il simbolo "$-L\rightarrow$" indica l'implicazione via leggi di natura.

Ovvero, un controfattuale del tipo della divergenza naturale è vero se il suo antecedente, unito alla storia del mondo prima del suo verificarsi, unito a tutti gli eventi che si verificano prima dell'evento-conseguente, ma dopo l'evento-antecedente, non influenzati da quest'ultimo, danno come risultato grazie alle leggi di natura, il suo conseguente.

Esistono però controfattuali che richiedono di considerare anche cambiamenti dovuti a processi cominciati prima del tempo t_A dell'evento-antecedente.

Kvart estende la sua analisi anche a questo tipo di controfattuali; perché essi siano veri, è necessario che il conseguente segua dalle leggi, dall'evento-antecedente (come nel caso del tipo della divergenza naturale), ma anche dalla storia del mondo prima del processo in questione – chiamiamolo P – e da una serie di fatti relativi al periodo del processo e all'intervallo (t_A, t_B) che vanno aggiunti all'informazione "fattuale".

Inoltre, per evitare che nelle premesse implicite vengano inclusi anche fatti "inverosimili", Kvart introduce una strumentazione probabilistica che in [92] elabora e spiega nei dettagli, ma della quale qui enunciamo solamente la versione intuitiva fornita in [93]:

---

[24] Un controfattuale $A > B$ (del tipo d.n.) è vero se e solo se

$\{A\} \cup W_A \cup \{$i conseguenti dei semifattuali veri $A > C$ con $t_C \subseteq (t_A, t_B)\}$ $-L\rightarrow B$

[*traduzione mia*]

> [...] an important requirement a process specified by 'P' must meet in order for the process to qualify as one that could have 'led' to the A-event is that
>
> $$P(A/P\&W_P) > P(A/W_P)$$[25]
>
> [Counterfactuals, p.164]

Ossia la probabilità del verificarsi dell'evento A dato il processo P e lo stato del mondo anteriore a P è maggiore della probabilità del verificarsi di A dato solo lo stato del mondo anteriore a P.

Secondo Kvart, proprio il ricorso a un'analisi indipendente (in termini probabilistici) di un sottogruppo di controfattuali (gli irril-semifattuali e i p.p.-semifattuali) scongiura per la sua teoria la minaccia del regresso all'infinito che aveva invece colpito la teoria di Goodman.

2.2.3 La teoria coerentista: Rescher

Nicholas Rescher affronta il problema dei controfattuali in [137] all'interno di una teoria coerentista che si contraddistingue per il fatto di ricercare, più che una definizione della nozione di verità, un criterio per affermare che un enunciato sia vero. Tale criterio consiste nella coerenza dell'enunciato in questione con gli altri enunciati accettati in precedenza come veri.

L'idea da cui parte Rescher è che l'antecedente del controfattuale introduca un elemento di incoerenza nell'insieme di credenze o conoscenze accettate dall'agente ragionante. Il conseguente del controfattuale sarebbe allora uno dei risultati dell'operazione compiuta dall'agente per ristabilire la coerenza.

In altre parole, secondo Rescher, bisogna partire da un insieme S di credenze o conoscenze consistente e aggiungere l'ipotesi controfattuale a tale insieme, rendendolo in tal modo inconsistente. All'interno di un insieme

---

[25] [...] un importante requisito che un processo specificato come 'P' deve soddisfare per poter essere qualificato come un processo che abbia 'condotto' all'A-evento è che

$$P(A/P\&W_P) > P(A/W_P)$$

[*traduzione mia*]

inconsistente, tuttavia, si possono rintracciare dei sottoinsiemi consistenti massimali (*s.c.m.*) in cui sia vero l'antecedente controfattuale.

A questo punto Rescher introduce tre diverse nozioni di conseguenza per mettere in relazione le proposizioni con questi insiemi di credenze/conoscenze:

- **W-conseguenza**: p è una W-conseguenza dell'insieme S se c'è qualche **s.c.m.** $S' \subseteq S$ tale che $S' \models p$;
- **I-conseguenza**: p è una I-conseguenza dell'insieme S se per ogni **s.c.m.** $S' \subseteq S$, $S' \models p$;
- ***P*-conseguenza** (o conseguenza *plausibile* o *preferibile*): p è una P-conseguenza dell'insieme S se per ogni **s.c.m.** $S' \subseteq S$ tale che S' è un **s.c.m.** preferito di S, $S' \models p$.

Questo terzo tipo di nozione di conseguenza è quello che sarà utilizzato nel caso dei controfattuali.

Ecco come Rescher illustra la questione in [137]:

> A 'solution' of the Problem of Counterfactual Conditionals is at hand when the particular W-consequence Q_1 of **S'** which is in question is also a consequence of all the *P-preferred* $\sim P_1$-containing m.c.s. of **S'** (for some appropriate criterion of preference P), with the result that Q_1 – unlike its logically compatible competitors Q_2, Q_3, etc. – is a 'natural' consequence of **S'** (with respect to the preferential criterion at issue)[26].
>
> [*The Coherence Theory of Truth*, p.286]

Questo passo è importante perché mostra che nella teoria di Rescher esiste normalmente più di un candidato tra i modelli massimamente consistenti

[26]Una "soluzione" del problema dei condizionali controfattuali è a portata di mano quando la particolare W-conseguenza Q_1 di **S'** che è in questione è pure una conseguenza di tutti i s.c.m. di **S'** contenenti $\sim P_1$ che sono *P-preferiti* (per qualche criterio appropriato di preferenza P), con il risultato che Q_1 – a differenza dei suoi concorrenti incompatibili Q_2, Q_3, ecc. – è una conseguenza "naturale" di **S'** (rispetto al criterio preferenziale in questione).

[*tr. it. di Claudio Pizzi:* [138], p.117]

che contengono l'antecedente e quindi di norma esiste più di un conseguente possibile per il controfattuale. Solo attraverso la definizione di un criterio di preferenza è possibile capire quale sia il conseguente controfattuale "più naturale" per quell'antecedente.

Ovviamente, resta il problema di determinare questi criteri preferenziali; per far questo Rescher propone un indiciamento di plausibilità, che assegna valori diversi a leggi, generalizzazioni universali o fatti atomici. L'indiciamento, poi, non è dato una volta per tutte, ma si può decidere, a seconda delle circostanze, di valutare, ad esempio, una generalizzazione come più plausibile di un fatto atomico o viceversa. Questo consentirà di ricavare conclusioni variabili dallo stesso antecedente controfattuale a seconda del criterio di indiciamento scelto.

La pluralità di criteri di preferibilità, ben lungi dal costituire una limitazione, è vista da Rescher come una peculiarità degli approcci coerentisti:

> No attempt will be made here to provide one solitary monolithic solution. We view the situation as fundamentally pluralistic: there is no one single criterion of m.c.s. preference that by itself provides the sole rationally viable and invariably appropriate procedure. A variety of methods for establishing alethic eligibility is available, each with its own distinctive points of advantage and disadvantage and each peculiarly fitted for application to a certain range of uses[27].
>
> [*The Coherence Theory of Truth*, p.99]

Tale criterio di indiciamento non sarà dunque individuato dalla logica, ma dall'euristica che il singolo individuo decide di adottare:

---

[27]Nessun tentativo sarà fatto in questa sede di fornire una soluzione monolitica solitaria. Consideriamo questa situazione come fondamentalmente pluralistica: non c'è un singolo criterio di preferenza dei s.c.m che in sé fornisca la sola procedura razionalmente ottenibile e invariabilmente appropriata. Per stabilire l'eleggibilità aletica è disponibile una varietà di metodi, ognuno con i suoi punti distintivi di vantaggio e svantaggio e ognuno particolarmente studiato per l'applicazione a una certa gamma di usi. [*traduzione mia*]

> Such conceivable but more far-fetched and less palatable-seeming counterfactuals as 'If this match had been struck, it would not have been dry', which call for a rejection of a statement different from p_5 (specifically p_2), are ruled out – not by the 'logic' of the situation (**S'**$_2$ is, after all, a perfectly good m.c.s.) – but by the policies adopted in implementing the natural plausibilities of the case[28].
>
> [*The Coherence Theory of Truth*, p.292]

Non esiste quindi un modo univoco di determinare quale sia la conseguenza più naturale di un enunciato. Questo è molto interessante perché permette di spiegare da una parte perché agenti diversi a volte ricavino conseguenze diverse dallo stesso enunciato e, quando invece si ha l'accordo tra più agenti, ciò significa che questi agenti condividono tacitamente il criterio di preferenza o, perlomeno, i loro criteri di preferenza sono compatibili.

La posizione di Rescher racchiude in sé due elementi di relativismo: uno legato al criterio di indiciamento, l'altro alla preponderanza della nozione di coerenza su quella di verità; questo secondo aspetto la rende piuttosto adatta a rappresentare i processi cognitivi di "riaggiustamento" dell'informazione disponibile che soggetti limitati nelle loro facoltà conoscitive sono costretti a mettere costantemente in atto per aggiornare le loro conoscenze.

2.2.4 La revisione di credenze

La teoria basata sulla *belief revision* – o revisione delle credenze – parte dall'assunto che i condizionali non sono portatori di valore di verità, anche se è possibile fornirne delle condizioni di accettabilità o rifiuto. Ai condizionali così intesi viene dato il nome di *condizionali epistemici*.

Ecco come Horacio Arló Costa definisce i condizionali epistemici in [37]:

---

[28]Controfattuali concepibili ma più stravaganti e all'apparenza meno digeribili come "Se questo fiammifero fosse stato sfregato non sarebbe stato asciutto", che richiedono la reiezione di un enunciato diverso da p_5 (nella fattispecie p_2), sono esclusi, non dalla "logica" della situazione (**S'**$_2$, dopo tutto, è un s.c.m. perfettamente in ordine), ma dalle strategie adottate nel soddisfare le naturali plausibilità del caso.

[*tr. it. di Claudio Pizzi:* [138], p.124]

> These conditionals are not part of the stock of X's "first order" beliefs, but they are part of X's metabeliefs about X's own beliefs, and the ways that they may change. We call these conditionals *epistemic conditionals*[29].
>
> [Epistemic Conditionals, Snakes, and Stars, p.204]

dove X è un agente ragionante qualsiasi.

Gli autori che si richiamano alla revisione di credenze sostengono che l'interpretazione del test di Ramsey fornita da Stalnaker in [150] non è fedele alle intenzioni dello stesso Ramsey, intenzioni che Horacio Arló Costa e Isaac Levi in [40] riassumono con un elenco di condizioni:

> 1. The conditionals considered acceptable according to Ramsey test are neither truth-value bearers nor objects of belief.
> 2. The conditionals 'If A, then B', and 'If A, then $\neg B$', cannot be simultaneously acceptable relative to the epistemic state of any agent that is in suspense about A.
> 3. The conditionals delivered by the Ramsey test are to be understood as expressions of *suppositional reasoning*.
> 4. An agent who is in suspense about A, accepts 'If A, then B' with respect to his epistemic state K iff B belongs to the belief state obtained after adding A to K[30].
>
> [Two notions of epistemic validity, p.219]

---

[29] Questi condizionali non sono parte della base di credenze "del primo ordine" di X, ma sono parte delle metacredenze di X sulle sue proprie credenze e sui modi in cui possono cambiare. Chiamiamo questi condizionali *condizionali epistemici*. [*traduzione mia*]

[30]

1. I condizionali considerati accettabili secondo il test di Ramsey non sono né portatori di valore di verità né oggetti di credenza.
2. I condizionali 'Se A, allora B' e 'Se A, allora $\neg B$' non possono essere simultaneamente accettabili relativamente allo stato epistemico di un agente che sospenda il giudizio su A.

Nella versione fornita da Stalnaker del test di Ramsey i condizionali sono vero-funzionali e quindi non rispondenti ai requisiti avanzati dallo stesso Ramsey lungo tutto il corso dei suoi studi.

Secondo Peter Gärdenfors [63], il significato dei condizionali non risiede in una supposta corrispondenza col mondo reale, ma con un sistema di credenze caratterizzato da una classe di modelli degli stati epistemici, una funzione di valutazione per la determinazione degli atteggiamenti epistemici, una classe di input epistemici e una funzione che assegni a ogni stato di credenza e input epistemico, un nuovo stato di credenza.

In [63] Gärdenfors formalizza gli stati di credenza attraverso insiemi di enunciati deduttivamente chiusi e sono individuati tre possibili atteggiamenti epistemici: accettazione, rifiuto e sospensione del giudizio. Dati un enunciato A e un insieme di credenze K:

1. A è accettato rispetto a K sse $A \in K$;
2. A è rifiutato rispetto a K sse $\neg A \in K$;
3. su A è sospeso il giudizio rispetto a K sse $A \notin K, \neg A \notin K$.

Nel caso di enunciati condizionali, come $A > B$, è necessario introdurre la nozione di "impegno epistemico" (*epistemic commitment*), rappresentata con il simbolo $*$. L'espressione $K * A$ rappresenta dunque lo stato di credenze K aggiornato dell'informazione A. L'atteggiamento epistemico di accettazione per i condizionali diventa quindi:

$$A > B \text{ è accettato rispetto a } K \text{ sse } B \in K * A$$

La nozione di accettazione di Gärdenfors è però ancora in parte legata alla verofunzionalità, poiché afferma che un atteggiamento di accettazione è corretto se la proposizione alla quale si applica è vera. Di conseguenza:

---

3. I condizionali resi dal test di Ramsey devono essere compresi come espressioni del *ragionamento supposizionale*.
4. Un agente che ha sospeso il giudizio su A accetta 'Se A, allora B' rispetto al suo stato epistemico K sse B appartiene allo stato di credenza ottenuto in seguito all'aggiunta di A a K.

[*traduzione mia*]

$$A > B \in K \text{ sse } B \in K * A$$

Un approccio un po' diverso è quello proposto da Arló Costa e Levi che affermano che, mentre da una parte i modelli epistemici di Gärdenfors forniscono criteri di accettazione per condizionali che restano verofunzionali, dall'altra il loro approccio produce criteri di accettazione per condizionali *à la* Ramsey, ossia completamente epistemici.

Pur non essendo verofunzionali, i condizionali sono nondimeno importanti per esprimere atteggiamenti cognitivi molto rilevanti ed è quindi necessario caratterizzarne i criteri di accettabilità attraverso quella che gli autori definiscono una "teoria stratificata".

Dati L_0, un linguaggio booleano senza operatori modali o epistemici, K, l'insieme di enunciati di L_0 accettati da un agente a un tempo t (chiuso rispetto alla conseguenza logica), tutti gli enunciati di K sono, dal punto di vista di tale agente, veri al tempo t. Quando si ragiona sull'accettabilità di un condizionale $A > B$, in realtà non si sta prendendo in considerazione l'appartenenza a K, ma la possibilità di avere B in quella che è una trasformazione di K (seguita all'aggiunta di A).

Sarà allora necessario identificare un linguaggio più esteso di L_0, chiamiamolo $L_>$, con il quale sia possibile esprimere tutti quegli enunciati accettabili sulla base di K e dell'impegno al cambiamento espresso dall'agente al tempo t. Sia allora $s(K)$ l'"insieme di supporto", ossia l'insieme che raggruppa tuti i condizionali accettati dall'agente al tempo t. $s(K)$ ha le seguenti caratteristiche:

- $s(K) \subseteq K$;
- $s(K)$ è chiuso rispetto alla conseguenza logica;
- ogni enunciato $A \in L_0$ che appartiene a $s(K)$ appartiene anche a K.

Il test di Ramsey assume dunque in [40] la seguente forma:

Se $A, B \in L_0$, allora $A > B \in s(K)$ sse $B \in K * A$ con K consistente.

In una serie di articoli ([38], [39], [37], [40]), Arló Costa e altri suoi collaboratori hanno fornito le definizioni di soddisfacibilità, validità e implicazione

per i loro modelli epistemici e ne hanno mostrato un ampio spettro di applicazioni, le più importanti delle quali possono essere brevemente elencate: iterazione, condizionali ontici e "dogmatici" (*opinionated*), modelli preferenziali, logiche non monotone, condizionali "annidati" (*nested*) e interpretazioni probabilistiche.

Tutto il discorso portato avanti all'interno della teoria della revisione di credenze relativamente ai condizionali in genere vale ovviamente anche con i controfattuali, con la differenza che il soggetto ragionante, invece che introdurre un input nuovo all'interno dell'insieme di credenze accettate, aggiunge un'informazione che si pone in conflitto con una o più credenze accettate e questo determina una revisione che comporti il minimo cambiamento necessario ad accomodare il nuovo input all'interno della base di credenze.

2.3 Considerazioni conclusive

Nella nostra esposizione siamo partiti da quello che può essere considerato l'approccio standard in filosofia, ossia quello basato sulla semantica a mondi possibili, fino a giungere alle teorie basate sulla revisione di credenze che, pur essendo state ampiamente utilizzate nell'ambito degli studi di intelligenza artificiale per implementare negli agenti artificiali alcuni processi di calcolo paragonabili al nostro ragionamento condizionale e controfattuale, non sono ancora state completamente recepite in ambito filosofico, anche se gli autori di cui abbiamo parlato stanno fornendo un contributo decisivo in questo senso.

Le teorie che abbiamo chiamato "vero-funzionali" hanno un carattere più "metafisico" rispetto alle teorie consequenzialiste, che sono invece più attente alla dimensione cognitiva, nel senso che, mentre le prime si pongono come obiettivo di assegnare un valore di verità ai controfattuali rispetto a ciò che accade nel mondo, le seconde si preoccupano di definire dei criteri affinché i controfattuali possano essere assunti a pieno titolo nelle credenze o conoscenze di un agente ragionante.

Rispetto a questa alternativa, noi propendiamo per la prospettiva cognitiva, soprattutto per via delle applicazioni del ragionamento che intendiamo sviluppare nella direzione del ragionamento finalizzato all'azione, dove ciò

che è allo studio sono proprio i processi mentali messi in atto dagli agenti cognitivi, che presentano delle problematiche per risolvere le quali gli approcci consequenzialisti sono stati espressamente pensati e sono quindi naturalmente portati ad affrontare.

Un'altra tematica, collegata a questa, ma in un certo senso trasversale alle due classi di approcci è quella della parzialità: nella valutazione dei controfattuali ci si deve confrontare con oggetti completi (come i mondi possibili o l'insieme di tutte le leggi di natura), oppure con oggetti parziali (come le situazioni o gli stati epistemici)?

Normalmente sono i consequenzialisti a sottoscrivere la tesi della parzialità, poiché la loro attenzione è concentrata su agenti con capacità intellettive limitate e fallibili e non con entità assolute come "la verità"; tuttavia non mancano, anche sul versante vero-funzionale, teorie che si confrontano con la parzialità, come la *situation semantics*.

Anche in questo caso, riteniamo che per poter fornire una rappresentazione verosimile del modo di ragionare tipico di agenti limitati, la caratteristica della parzialità sia ineliminabile.

La posizione che presenteremo e difenderemo nel capitolo 4 ha molto in comune sia con la *situation semantics* che con le teorie coerentiste e con la revisione di credenze, ma cerca di far confluire la vero-funzionalità e i criteri di accettabilità nella nozione di *verità locale*, che esprime ciò che un agente valuta come vero all'interno di una teoria che egli stesso costruisce per ragionare su un determinato problema.

Capitolo 3

Dai condizionali al ragionamento controfattuale

Questo capitolo segna un passaggio importante nella nostra disamina del lavoro che è già stato fatto sul "fenomeno della controfattualità", poiché mostra l'esistenza di una prospettiva di analisi alternativa rispetto a quella tradizionalmente compiuta dalla filosofia del linguaggio, che si proponeva di condurre un'indagine limitata al rapporto tra costrutti linguistici e valori di verità. Questa nuova prospettiva permette di interpretare la controfattualità come un fenomeno di ragionamento e quindi di indagare la relazione tra costrutti linguistici e procedure di ragionamento.

Nella sezione 3.1 si mostra come, nell'ambito degli studi in intelligenza artificiale, alcuni autori comincino a percepire la funzionalità del ragionamento controfattuale come strumento cognitivo utilizzabile nel quadro di un più ampio processo di ragionamento.

Le sezioni 3.2 e 3.3 hanno come oggetto due teorie, la prima sviluppata nell'ambito delle scienze cognitive, la seconda dell'intelligenza artificiale, che hanno il merito di avere tentato una rappresentazione sistematica non solo della dinamica del ragionamento controfattuale, ma anche di come questa dinamica si integri in una teoria avente lo scopo di spiegare come funziona tutto il ragionamento, a partire da come è strutturata la conoscenza, fino ad arrivare a come viene utilizzata e a come evolve.

3.1 Alcuni approcci in intelligenza artificiale

Nell'ambito dell'intelligenza artificiale sono stati compiuti una serie di studi che hanno messo in evidenza l'importanza del ragionamento controfattuale nell'elaborazione di determinati processi cognitivi che possono essere riprodotti dalle macchine e, parallelamente, si è assistito al tentativo di utilizzare alcuni tipi di calcoli logici per rappresentare il fenomeno.

Il primo articolo che presenta una trattazione abbastanza esaustiva in questo senso è [67], articolo in cui Matthew Ginsberg da una parte presenta il suo approccio basato sulle teorie elaborate da Lewis e Stalnaker in filosofia e lo utilizza per tentare di risolvere problemi di intelligenza artificiale, in particolar modo problemi legati al *planning* e alla diagnosi di errori, dall'altra mostra delle applicazioni a un *planner* e a un sistema diagnostico reali (rispettivamente, STRIPS e DART).

Ginsberg ha visto con chiarezza l'utilità del ragionamento controfattuale nella risoluzione di questi specifici problemi:

> The interest of AI researchers in nonmonotonic inference techniques is quite pragmatic; these techniques have been shown to be useful in addressing problems in a variety of areas where conclusions may be tentative. Our intention in this section is to demonstrate that counterfactual inference neatly captures the nonmonotonicity encountered in planning and diagnosis problems[1].
>
> [Counterfactuals, p.55]

Le condizioni di verità per i controfattuali fornite da Ginsberg sono un riadattamento da Lewis ; dato un insieme consistente di enunciati, S, che descrivono il mondo, e una premessa controfattuale p, si "indebolisca" S rimuovendo tutti i fatti che contribuiscono a provare $\neg p$; formalmente si prendano tutti i

[1] L'interesse dei ricercatori in IA per le tecniche di inferenza nonmonotona è abbastanza pragmatico; queste tecniche si sono mostrate utili nell'affrontare problemi in una varietà di aree nelle quali le conclusioni possono essere tentative. È nostra intenzione in questa sezione dimostrare che l'inferenza controfattuale cattura nettamente la nonmonotonicità incontrata nei problemi di planning e diagnosi. [*traduzione mia*]

sottoinsiemi di S che non implicano $\neg p$, si ordinino questi sottoinsiemi e sia $W(p, S)$ la classe dei sottoinsiemi massimi di S.

Ecco la definizione formale di [67]:

> $W(p, S) \equiv \{T \subseteq S \mid T \not\models \neg p \text{ and } T \subset U \subseteq S \Rightarrow U \models \neg p\}$
>
> We will define a counterfactual $p > q$ to be *true* in a world S if, and only if, for every $T \in W(p, S), T \cup \{p\} \models q$, so that the conclusion follows in every possible world where p holds[2].
>
> [Counterfactuals, p.44]

È abbastanza semplice rilevare l'equivalenza tra la soluzione offerta da Ginsberg e quella di Lewis. Questo lavoro, tuttavia, ha il duplice merito di avere per primo individuato la funzionalità del ragionamento controfattuale nella risoluzione di alcuni tipi di problemi e di aver esplorato alcune formalizzazioni logiche (logiche a più valori e *situation semantics*) per vedere quali tra esse potessero essere considerate più idonee per questi tipi di applicazioni del controfattuale.

Qualche anno più tardi, John McCarthy e Tom Costello in [41] hanno ribadito l'interesse che può suscitare lo studio dei controfattuali in intelligenza artificiale, evidenziandone la funzionalità.

Un primo contributo interessante di [41] è l'affermazione dell'importanza di valutare i controfattuali in riferimento alla teoria nella quale sono formulati, sostenuta attraverso un'analogia con i sistemi cartesiani, all'interno dei quali i punti assumono coordinate diverse. Questo accostamento sfocia nella definizione della nozione di *controfattuale cartesiano*, secondo gli autori il più immediato da trattare e anche il più utile:

> The most straightforward and possibly the most useful counterfactuals are what we call *cartesian counterfactuals*. A situation is described by the values of a number of parameters. The premise of the counterfactual is that one of the parameters has a

[2]Definiremo un controfattuale $p > q$ *vero* in un mondo S se, e solo se, per ogni $T \in W(p, S), T \cup \{p\} \models q$, così che la conclusione segua in ogni mondo possibile in cui valga p. [*traduzione mia*]

> different value than in the actual situation and that the other parameters have the same values. [...] If there are two systems that propagate changes to theories in two different ways, they may give different truth values for some $p \succ q$. This corresponds to the idea that we can choose different co-ordinate systems for the same space. In this case the meaning of the counterfactual depends on the co-ordinate frame. Indeed in some theories the counterfactual may not have a meaning at all. We see the world through the lenses of theories/frames and so must our robots[3].
>
> [Useful Counterfactuals, p.2]

Secondo gli autori i controfattuali sono importanti perché da essi è possibile imparare nuove cose sul mondo, per esempio quando ci si presenta nella realtà una situazione sufficientemente simile a un'altra su cui si sia ragionato controfattualmente, dal momento che la somiglianza ci autorizza a utilizzare in questa nuova situazione la stessa *teoria approssimata* che avevamo impiegato per il controfattuale:

> In so far as our knowledge is incomplete, new sentences can tell us more about the world. Every counterfactual we are told gives us more information about how the world would be, if things were only slightly different, relative to some unstated *approximate theory*. This information can later be used in a situation with only a small number of differences between it and the present, so that the *approximate theory* is applicable to both[4].

[3] I più semplici e forse utili controfattuali sono quelli che chiamiamo *controfattuali cartesiani*. Una situazione è descritta dai valori di un numero di parametri. La premessa del controfattuale è che uno dei parametri abbia un valore diverso rispetto a quello che ha nella situazione reale e che gli altri parametri abbiano gli stessi valori. [...] Se due sistemi propagano i cambiamenti alle teorie in due modi differenti, possono dare valori di verità differenti per qualche $p \succ q$. Questo corrisponde all'idea che possiamo scegliere diversi sistemi di coordinate per lo stesso spazio. In questo caso il significato del controfattuale dipende dal frame di coordinate. Addirittura in alcune teorie il controfattuale potrebbe non aver alcun significato. Vediamo il mondo attraverso le lenti di teorie/frame e lo stesso deve valere per i nostri robot. [*traduzione mia*]

[4] Nella misura in cui la nostra conoscenza è incompleta, nuovi enunciati possono dirci

> [Useful Counterfactuals, p.4]

McCarthy e Costello forniscono poi un'assiomatizzazione basata sul *situation calculus*, nella quale riescono sia a derivare i controfattuali che a inferire informazione non controfattuale a partire dai controfattuali.

Tuttavia, questa trattazione importa dal *situation calculus* una caratteristica che nella nostra prospettiva appare piuttosto come una limitazione, ossia la necessità di postulare sempre un linguaggio gerarchicamente superiore, quello che l'agente usa per parlare del mondo. In qualche modo, questo linguaggio è di natura differente rispetto a quelli utilizzati nelle teorie approssimate dei controfattuali, che devono sempre, in ultima istanza, essere riferiti e riportati a esso. Questo, oltre a complicare notevolmente la logica, costringe, in un certo senso, a imporre a priori che cosa sia fattuale e che cosa controfattuale, mentre noi vorremmo poter rappresentare il fatto che ciò che è controfattuale in una prospettiva possa essere fattuale in un'altra e viceversa.

Altri approcci molto interessanti sviluppati nell'ambito dell'intelligenza artificiale sono quelli fondati sulla *belief revision* e *update*, quello di Joseph Halpern basato sulla logica epistemica [81] e quello del gruppo di Judea Pearl [2], che si basa sulle reti bayesiane.

Un articolo di Thomas Eiter e Georg Gottlob, [49], espone in maniera molto chiara l'idea di fondo che sottende tutti i diversi approcci che identificano il problema della valutazione di un controfattuale con quello della valutazione di una revisione o di un aggiornamento di una base di conoscenza. Come scritto in [49]:

> The "implication problem" is as follows: given a knowledge base T, an update p, and a formula q, decide whether q is derivable from $T \circ p$, the updated (or revised) knowledge base. [...] Note that the implication problem we consider exactly corresponds to

---

qualcosa di nuovo sul mondo. Ogni controfattuale che ci viene detto ci dà più informazione circa come il mondo sarebbe, se le cose fossero leggermente differenti, relativamente a qualche inespressa *teoria approssimata*. Questa informazione può in seguito essere usata in una situazione che abbia solo un piccolo numero di differenze rispetto a quella presente, così che la *teoria approssimata* sia applicabile a entrambe. [*traduzione mia*]

> evaluating a counterfactual according to the particular revision or update semantics. Counterfactuals are conditional statements of the form "if p were true, then q would hold", where p is assumed to be false in the actual world. According to the *Ramsey Test*, evaluating such a counterfactual in a given knowledge base T is equivalent to test whether q is a logical consequence of $T \circ p$. [...] Given T and change operator $\circ$, define that "if p, then q (denoted by $p > q$) is true over T iff $T \circ p \models q$ holds [...][5].
>
> [On the Complexity of Propositional Knowledge Base. Revision, Updates, and Counterfactuals, pp.228, 240]

L'approccio di Halpern è interessante soprattutto dal punto di vista delle applicazioni poiché mostra come, combinando l'uso degli enunciati controfattuali con la logica epistemica ed elementi temporali, possono essere risolti problemi nel campo della programmazione [81] o delle decisioni nella teoria dei giochi [80].

In [80], per esempio, Halpern combina l'operatore controfattuale con l'operatore di conoscenza per esprimere l'operatore introdotto da Dov Samet per la conoscenza ipotetica; tale operatore, scritto $K^H(E)$, viene interpretato come "Se fosse stato H, allora avrei saputo E". Halpern propone di interpretare tale operatore come "se avessi considerato H possibile, avrei saputo E", esprimendo quindi l'operatore $K^H(E)$ come $L(H) > K(E)$, dove $>$ è l'operatore controfattuale standard (*à la* Lewis).

Come spiega Halpern:

---

[5]Il "problema dell'implicazione" è il seguente: data una base di conoscenza T, un aggiornamento p e una formula q, decidere se q è derivabile da $T \circ p$, la base di conoscenza aggiornata (o revisionata). [...] Da notare che il problema dell'implicazione corrisponde esattamente a quello di valutare un controfattuale secondo una particolare semantica di revisione o aggiornamento. I controfattuali sono enunciati condizionali della forma "se p fosse vero, allora varrebbe q", dove si assume che p sia falso nel mondo reale. Secondo il *Test di Ramsey*, valutare un tale controfattuale in una data base di conoscenza T equivale a testare se q è una conseguenza logica di $T \circ p$. [...] Data T e l'operatore di cambio $\circ$, si definisce che "se p, allora q (denotato da $p > q$) è vero rispetto a T sse $T \circ p \models q$.[*traduzione mia*]

> This reading suggests that we can then represent $K^H(E)$ as $L(H) > K(E)$, where $>$ is the standard counterfactual operator (so that $H > E$ can be read as "if H were the case, then E would be true"), K is the standard knowledge operator, and L is its dual (i.e., $L(E) = \neg K(\neg E)$, where $\neg$ denotes complementation)[6].
>
> [Hypothetical Knowledge and Counterfactual Reasoning, p.316]

Halpern suggerisce poi che questa traduzione in termini di controfattuali dell'operatore di Samet si rivela molto utile nell'analisi dei giochi a informazione imperfetta.

In [81], invece, Halpern mostra che, introducendo i controfattuali in un linguaggio di programmazione, è possibile implementare programmi che eseguano azioni supplementari che assicurino il raggiungimento di un obiettivo quando non hanno la sicurezza di raggiungerlo senza l'aggiunta di tali azioni. Secondo Halpern l'unico modo per svolgere questo compito è quello di riformulare il programma in termini controfattuali, ovvero come se dicesse: "se non sai se, nel caso tu non facessi nulla più, l'obiettivo sarebbe raggiunto, allora esegui delle azioni aggiuntive".

La trattazione che Halpern dà del fenomeno dei controfattuali è in sostanza una rielaborazione della semantica di Lewis, ma l'aspetto veramente interessante è costituito dall'idea di fondo che il controfattuale può essere integrato con altre forme di ragionamento e in questo modo costituire uno strumento molto potente per la soluzione di problemi pertinenti a diversi ambiti disciplinari.

Per concludere questa breve esposizione di alcune trattazioni dei controfattuali nell'ambito degli studi di intelligenza artificiale, prendiamo ora in considerazione la posizione di Judea Pearl, che affronta il problema dei controfattuali collocandolo all'interno della teoria della causalità e dell'azione da lui sviluppate.

---

[6]Questa lettura suggerisce che possiamo rappresentare $K^H(E)$ come $L(H) > K(E)$, dove $>$ è l'operatore controfattuale standard (di modo che $H > E$ può essere letto come "se fosse stato H, allora E sarebbe stato vero"), K è l'operatore di conoscenza standard e L è il suo duale (cioè, $L(E) = \neg K(\neg E)$, dove $\neg$ denota l'operazione complemento). [*traduzione mia*]

Un articolo in cui vengono messi bene in luce i legami tra controfattuali e causalità è [121] e una caratterizzazione assiomatica viene fornita in [62]; tuttavia, la trattazione più esaustiva è contenuta nel libro *Causality: Models, Reasoning, and Inference*, [122].

Pearl, cercando di fornire una base più solida alla nozione di somiglianza fornita da Lewis in vari lavori, tra cui [100], [101] e [103], affronta il problema della valutazione del controfattuale servendosi di una teoria bayesiana e fornendo un supporto probabilistico al processo che Lewis in [103] aveva definito *imaging*.

Il risultato è un processo in tre passi che Pearl illustra in [122]:

> *Step 1 (abduction)*: Update the probability $P(u)$ to obtain $P(u \mid e)$.
>
> *Step 2 (action)*: Replace the equations corresponding to variables in set X by the equations $X = x$.
>
> *Step 3 (prediction)*: Use the modified model to compute the probability of $Y = y$.
>
> In temporal metaphors, this three-step procedure can be interpreted as follows. Step 1 explains the past (U) in light of the current evidence e; step 2 bends the course of history (minimally) to comply with the hypothetical condition $X = x$; finally, step 3 predicts the future (Y) based on our new understanding of the past and our newly established condition, $X = x$[7].
>
> [*Causality: Models, Reasoning, and Inference*, p.63]

Vedremo ora come in un'altra disciplina, le scienze cognitive, siano stati

[7] *Passo 1 (abduzione)*: Aggiornare la probabilità $P(u)$ per ottenere $P(u \mid e)$. *Passo 2 (azione)*: Sostituire le equazioni corrispondenti alle variabili nell'insieme X con le equazioni $X = x$. *Passo 3 (predizione)*: Usare il modello modificato per computare la probabilità di $Y = y$. Nelle metafore temporali, questo procedimento in tre passi può essere interpretato come segue. Il passo 1 spiega il passato (U) alla luce dell'evidenza corrente e; il passo 2 modifica il corso della storia (minimamente) per accomodare la condizione ipotetica $X = x$; infine, il passo 3 predice il futuro (Y) sulla base della nostra nuova comprensione del passato e la nostra condizione appena stabilita, $X = x$. [*traduzione mia*]

elaborati modelli (non formali) sulla base di presupposti pratici molto simili, ma ponendo l'accento sulla dimensione della parzialità del ragionamento.

3.2 Gli spazi mentali di Fauconnier

Il primo, ma forse più fondamentale, passo nella direzione che vogliamo intraprendere è stato compiuto pochi anni fa nell'ambito delle scienze cognitive, quando Gilles Fauconnier in [52] ha deciso di applicare la sua teoria basata sugli spazi mentali al fenomeno dei controfattuali.

Il primo importante spostamento di prospettiva è probabilmente dovuto alla natura stessa della disciplina di cui si occupa Fauconnier, le scienze cognitive. Infatti, mentre la logica generalmente si occupa di studiare i rapporti esistenti tra linguaggio e verità, le scienze cognitive si pongono come obiettivo quello di esaminare i modelli mentali che gli agenti cognitivi sviluppano in seguito alla ricezione di determinati stimoli, tra cui anche quelli linguistici.

Come hanno sottolineato George Lakoff e Eve Sweetser nella prefazione di [52], la differenza tra le teorie logiche e quelle cognitive risiede nei loro diversi oggetti di studio; ecco come descrivono i modelli della logica:

> These are objectivist models, models of the actual world, or of a possibile world, or an actual or possibile situation. Possible worlds and situations are not models of the human mind, but models of the world as it is assumed to be or might be[8].
>
> [*Mental Spaces: Aspects of Meaning Construction in Natural Language*, p.xi]

3.2.1 Gli spazi mentali

L'approccio di Fauconnier è incentrato sul concetto di *spazio mentale*, che è stato presentato la prima volta all'Accademia della Crusca a Firenze nel 1978.

[8]Questi sono modelli oggettivisti, modelli del mondo reale, o di un mondo possibile, o di una situazione reale o possibile. I mondi possibili e le situazioni non sono modelli della mente umana, ma modelli del mondo come assumiamo che sia o come potrebbe essere. [*traduzione mia*]

Gli spazi mentali sono domini nei quali la conoscenza di un agente sarebbe strutturata. Il ragionamento consisterebbe dunque nella "manipolazione" di questi spazi e nelle operazioni che potrebbero essere compiute su di essi.

La costruzione di questi spazi prenderebbe le mosse da dei cosiddetti "costruttori di spazio" (*space builders*), costrutti linguistici o grammaticali e fattori pragmatici o retorici, che innescherebbero, nello stato cognitivo dell'agente, questo processo di costruzione.

Così Fauconnier e Sweetser in [155]:

> The basic idea is that, as we think and talk, mental spaces are set up, structured, and linked under pressure from grammar, context, and culture. The effect is to create a network of spaces through which we move as discourse unfolds[9].
>
> [Cognitive Links and Domains, p. 11]

Il punto di partenza di Fauconnier è quindi una rappresentazione dello stato cognitivo e della base di conoscenza degli agenti come ripartiti in spazi mentali (*mental spaces*), all'interno dei quali avvengono i processi di ragionamento.

La possiblità di poter effettuare operazioni tra elementi che si trovano collocati in spazi mentali diversi è garantita dal principio di identificazione, che stabilisce la connessione tra spazi all'interno della configurazione totale.

Esistono dei meccanismi cognitivi per passare da uno spazio a un altro e per creare, mediante semplici operazioni, nuovi spazi. Questi meccanismi sono le cosiddette *cross-domain functions* (funzioni trans-dominio) e possono essere il riferimento, l'inferenza, la proiezione di struttura, che può assumere diverse forme. Una di queste operazioni che riveste un particolare interesse in relazione ai controfattuali è il *blending* (che potremmo tradurre con "miscelamento"), processo attraverso il quale è possibile creare un nuovo spazio mentale a partire da due spazi di input.

Ecco come Fauconnier lo definisce in [53]:

---

[9] L'idea basilare è che, quando pensiamo e parliamo, gli spazi mentali vengono fissati, strutturati e connessi sotto la pressione della grammatica, del contesto e della cultura. L'effetto è la creazione di una rete di spazi attraverso la quale ci muoviamo man mano che si svolge il discorso. [*traduzione mia*]

> It operates on two Input mental spaces to yield a third space, the *blend*. The blend *inherits partial structure* from the input spaces and *has emergent structure* of its own[10].
>
> [*Mappings in Thought and Language*, p.149]

Il risultato di tutti questi processi è il carattere fluido, dinamico, creativo del discorso, in cui connessioni e passaggi sono temporanei, in continua evoluzione e il significato dei costrutti linguistici viene costantemente rinegoziato.

3.2.2 Controfattuali analogici

Ma veniamo ora al modo in cui Fauconnier affronta il problema dei controfattuali. Secondo Fauconnier, è riduttivo pensare (come è stato fatto nella tradizione filosofica) che un controfattuale rappresenti una situazione immaginaria che differisce dalla realtà esattamente rispetto a ciò che è espresso nell'antecedente del controfattuale.

A suo avviso, la struttura del controfattuale non è vero-funzionale, ma analogica, essendo questa il risultato della proiezione di una struttura da un dominio su un altro [54].

Per spiegare meglio come si costruisce uno spazio controfattuale, Fauconnier parte da un esempio: "In Francia, il Watergate non avrebbe fatto alcun male a Nixon".

Secondo Fauconnier, quando una persona afferma un enunciato di questo genere, il suo stato cognitivo si configura secondo dei *frames*: per prima cosa, un *frame* generico, che prevede dei ruoli e delle relazioni tra questi ruoli, nell'esempio il *frame* (chiamato F) sarà quello di una democrazia occidentale; all'interno di questo *frame*, un Paese ha un presidente eletto dai cittadini, il presidente è il capo di un partito politico che compete con gli altri per il governo del Paese, le azioni del presidente sono vincolate dalle leggi, dall'opinione pubblica ecc. e un'azione danneggia il presidente se scatena una

[10]Opera su due spazi mentali di Input per produrre un terzo spazio, il *blend*. Il blend *eredita una struttura parziale* dagli spazi di input e *ha una struttura emergente* sua propria. [*traduzione mia*]

reazione negativa nell'opinione pubblica oppure se è illegale, nel qual caso il presidente viene punito per averla commessa.

Il secondo passo è quello di costruire uno spazio B, dove i ruoli presenti nel *frame* F generico assumono un valore preciso: il presidente in questione è Nixon, il Paese gli Stati Uniti, i cittadini gli americani ecc.

Viene poi costruito un secondo *frame* F', che ha la maggior parte della struttura in comune con F e i cui valori vengono fissati dando luogo allo spazio G (che rappresenta la situazione analoga relativa alla Francia).

L'ultimo passo è quindi la costruzione dello spazio controfattuale, C, basato anch'esso sul *frame* F', che importa informazione sia da B che da G e che porta con sé della nuova informazione, che non era presente nei due spazi di partenza.

L'operazione che permette la creazione dello spazio controfattuale a partire da altri spazi e *frames* è proprio il *blending*.

Ecco come Fauconnier in [53] descrive il processo di formazione dello spazio controfattuale:

> The counterfactual space that the sentence prompts us to build is a *blend* of the two inputs. It inherits the generic frame from both inputs, and the specific additional political and social properties of France from Input 2, by virtue of the space builder *in France*[11].
>
> [*Mappings in Thought and Language*, p.159]

La costruzione dello spazio controfattuale sarà dunque un processo analogico di proiezione di struttura dallo spazio dal quale è generato e la sua funzione è quella di fornire informazioni, per via indiretta, sulla relazione che questo nuovo spazio intrattiene con lo spazio di partenza, essendo sottoposto alle condizioni sugli spazi ipotetici (*matching conditions*).

Altri studi più recenti, come per esempio [164], si sono posti l'obiettivo di fornire evidenza sperimentale del fatto che i soggetti ragionano controfattualmente sulla base di una corrispondenza analogica con la conoscenza di sfondo che possiedono rispetto a un determinato ambito.

[11] Lo spazio controfattuale che l'enunciato ci porta a costruire è un *blend* dei due input. Eredita il frame generico da entrambi gli input e le proprietà politiche e sociali aggiuntive della Francia dall'Input 2, grazie al costruttore di spazio *in Francia*. [*traduzione mia*]

3.2.3 Considerazioni conclusive

La posizione espressa da Fauconnier è molto interessante per il lavoro che stiamo per intraprendere, poiché gli assunti di partenza, che la conoscenza sia ripartita e che il ragionamento abbia luogo in domini parziali ci paiono appropriati e funzionali alla spiegazione di alcuni fenomeni relativi al ragionamento osservati empiricamente, come per esempio la ricorrenza di determinati errori durante l'esecuzione di alcune procedure di ragionamento.

Altre due intuizioni molto significative per il lavoro che ci accingiamo a svolgere nel capitolo 4 sono:

- l'importanza della relazione che intercorre tra lo spazio base e lo spazio controfattuale, ovvero il criterio di scelta delle proprietà da importare nel passaggio da uno stato all'altro;
- l'importanza della possibilità di inferire qualcosa nello spazio base a partire da qualcos'altro che è stato dedotto nello spazio controfattuale.

La prima intuizione è così espressa in [53]:

> [...] many other blends are compatible with the original sentence. [...] To understand the sentence in context is to have some idea of the kind of blend intended. But it may take a lot of elaboration for speaker and hearer to converge on sufficiently similar constructions[12].
>
> [*Mappings in Thought and Language*, pp. 160–161]

La seconda proprietà è invece condensata nell'affermazione [53]:

> In this case also, inferences will be made in the blend and exported to the inputs[13].

[12] [...] molti altri blend sono compatibili con l'enunciato originario. [...] Comprendere l'enunciato nel contesto equivale ad avere qualche idea del tipo di blend inteso. Ma può essere necessaria molta elaborazione da parte del parlante e dell'ascoltatore per convergere su costruzioni sufficientemente simili. [*traduzione mia*]

[13] Anche in questo caso, le inferenze saranno tratte nel blend ed esportate negli input. [*traduzione mia*]

[*Mappings in Thought and Language*, p.163]

Riteniamo la teoria di Fauconnier estremamente valida da un punto di vista intuitivo, poiché esprime in maniera molto appropriata la dinamica di certe forme di ragionamento, ma essa non è accompagnata da un'adeguata formalizzazione (né probabilmente era nelle intenzioni di Fauconnier fornirne una). Nel capitolo 4 tenteremo di dimostrare che una tale formalizzazione è desiderabile e possibile.

3.3 Le rappresentazioni ripartite di Dinsmore

Una posizione molto simile a quella di Fauconnier o, meglio, a essa ispirata, è quella di John Dinsmore, che si propone di fondare un paradigma per la rappresentazione della conoscenza in intelligenza artificiale sulle intuizioni contenute nei lavori di Fauconnier.

3.3.1 Spazi e contesti

Dinsmore, in [45] e [46], in analogia con la nozione di spazio mentale descritta da Fauconnier, presenta la base di conoscenza come ripartita in quelle che chiama *partitioned representations* (rappresentazioni ripartite) o, più precisamente, spazi. La ripartizione della rappresentazione mentale in spazi consente una più facile risoluzione di molti problemi legati alla logica o al ragionamento.

Prima conseguenza di questo approccio è che l'oggetto dell'analisi linguistica non sono più gli enunciati presi in isolamento, ma *enunciati in uno spazio*, detti più semplicemente *asserzioni*. Quindi, se **S** è uno spazio e **P** un enunciato, allora **S** | **P** è l'asserzione dell'enunciato **P** nello spazio **S**.

Altre nozioni importanti introdotte da Dinsmore in [46] sono quelle di *contesto primario* e *contesto secondario*. Il contesto primario di uno spazio è un'asserzione che specifica come interpretare gli enunciati in uno spazio; per esempio, per indicare che tutto ciò che viene asserito nello spazio S_1 è una credenza dell'agente A, Dinsmore userebbe una notazione di questo genere:

$$S_0 \mid \textit{Nelle credenze dell'agente } A \; [[S_1]]$$

Da notare la doppia parentesi quadra di $[[S_1]]$, che differenzia il contesto primario da quello secondario, espresso mediante una parentesi quadra singola, per esempio $[S_2]$; essa serve a Dinsmore per suggerire l'esistenza di un certo tipo di vincolo, ossia l'eredità di contenuti dei due spazi sottesi dai contesti, come è facile evincere dalla definizione fornita da Dinsmore in [46]:

> A secondary context provides a kind of mapping from the contents of one space to the contents of another that is a consequence of the semantics of the primary contexts involved[14].
>
> [*Partitioned Representations*, p.67]

Ma vediamo ora quali tipi di ragionamenti abbiano luogo dentro e attraverso questi spazi.

3.3.2 Ragionamento parrocchiale e ripartito

Il ragionamento si svolgerebbe dunque all'interno di questi spazi e il loro contenuto avrebbe lo scopo di *simulare* una realtà possibile o, ancor meglio, una porzione di una realtà possibile; per tale ragione Dinsmore lo ha definito *simulative reasoning*, ragionamento simulativo [45]:

> Simulative reasoning is a highly efficient inference technique insofar as it treats difficult inferences over a potentially large set of complex propositions as relatively easy inferences over a small set of simple propositions[15].
>
> [Mental Spaces from a Functional Perspective, p.4]

[14] Un contesto secondario fornisce una specie di corrispondenza dai contenuti di uno spazio ai contenuti di un altro che è una conseguenza della semantica dei relativi contesti primari. [*traduzione mia*]

[15] Il ragionamento simulativo è una tecnica inferenziale altamente efficiente poiché tratta inferenze difficili sopra un insieme potenzialmente grande di proposizioni complesse come inferenze relativamente facili su piccoli insiemi di proposizioni. [*traduzione mia*]

Il passo denota anche la funzionalità di un tale tipo di ragionamento, che si applica a domini ristretti piuttosto che all'intera base di conoscenza.

Il ragionamento simulativo si divide poi in due diversi tipi di ragionamento, quello che avviene tutto all'interno di uno spazio e quello che ha luogo tra spazi diversi. Al primo Dinsmore ha dato in [46] il nome di *ragionamento parrocchiale* (*parochial reasoning*) e lo ha così definito:

> A space consolidates information that belongs together in one place, to model a coherent possible "reality" or situation. [...] Because of the license to ignore information external to a space in reasoning within a space, I call this localized reasoning *parochial*[16].
>
> [*Partitioned Representations*, pp. 47 e 53]

Il ragionamento parrocchiale è circoscritto a un singolo spazio e ha quindi il vantaggio di essere più "focalizzato" e dunque efficace. Le regole di inferenza operanti all'interno di un singolo spazio vengono dette *regole standard*.

Il ragionamento che viene condotto attraverso contesti diversi è basato sulle cosiddette *regole della ripartizione* e potremmo quindi definirlo *ragionamento ripartito*.

Le regole della ripartizione permettono di effettuare operazioni tra e su spazi diversi. Tali regole sono:

- *context climbing* o *decontestualizzazione*, che permette di decontestualizzare parzialmente un enunciato asserito in uno spazio, esplicitando qualcosa del suo contesto;
- *space initialization* o *inizializzazione dello spazio*, che permette, sotto certe condizioni, di creare nuovi spazi e il loro contesto primario;
- *space augmentation* o *espansione dello spazio*, che permette di aggiungere informazione a uno spazio;

[16] Uno spazio consolida informazione che sta tutta insieme in un posto, per modellare una "realtà" o situazione possibile coerente. [...] Grazie alla licenza di ignorare informazione esterna allo spazio nel ragionare all'interno di uno spazio, chiamo questo ragionamento localizzato *parrocchiale*. [*traduzione mia*]

- *space identity* o *identità di spazio*, che permette di porre in corrispondenza un enunciato con se stesso.

3.3.3 Ragionamento ripartito e controfattuali

Per quanto riguarda il controfattuale, Dinsmore parte dal riconoscimento del fatto che esso instauri un certo tipo di relazione tra due spazi, i contenuti dei quali sono fortemente interconnessi e, in particolare, uno dei due dipende in maniera cruciale dall'altro.

Tuttavia non si può certo dire che il nuovo spazio (quello controfattuale) erediti *in toto* il contenuto del vecchio, si tratta piuttosto di un'eredità di default (o *default inheritance*, come la definisce lo stesso Dinsmore).

Ecco come presenta in [45] il fenomeno dei controfattuali:

> The content of one space can depend crucially on the content of another as a function of the semantics of the respective contexts and yet not exhibit absolute inheritance. This is the case for counterfactual or "if S *were* true, then ..." spaces, as opposed to "$S \rightarrow$ ..." or simple "if S is true, then ..." spaces. [...] The kind of inheritance involved in this case cannot be absolute [...]. Such cases require a weaker form of inheritance, *default inheritance*[17].
>
> [Mental Spaces from a Functional Perspective, pp.11-12]

Questo passaggio riporta due importanti intuizioni, ossia che il controfattuale istituisce un nuovo spazio di ragionamento e che questo spazio deve avere delle regole di formazione specifiche, diverse da quelle di un normale condizionale.

---

[17] Il contenuto di uno spazio può dipendere in maniera cruciale dal contenuto di un altro come funzione della semantica dei rispettivi contesti e tuttavia non esibire ereditarietà assoluta. Questo è ciò che accade con gli spazi controfattuali o del tipo "se S *fosse* vero, allora ...", contrariamente agli spazi "$S \rightarrow$ ..." o ai semplici "se S è vero, allora ...". [...] Il tipo di ereditarietà caratteristico di questo caso non può essere assoluto [...]. Questi casi richiedono una forma di ereditarietà più debole, l'*ereditarietà di default*. [*traduzione mia*]

Quale sia l'informazione che viene ereditata da uno spazio a un altro e quale invece debba restare confinata nello spazio di partenza non viene precisato da Dinsmore che, attraverso un esempio, mostra che certe informazioni, incompatibili con la proposizione antecedente che definisce il nuovo spazio, non vengono trasmesse. L'impressione che si ha è che, come nel caso di Stalnaker e Lewis, le informazioni preservate dall'ereditarietà di default vengano decise di volta in volta sulla base di considerazioni pragmatiche.

3.3.4 Considerazioni conclusive

La teoria di Dinsmore, oltre che essere ancora un po' insoddisfacente per i nostri fini dal punto di vista formale, è per noi ancora troppo ancorata alla dimensione oggettiva. Più che la simulazione, vorremmo che nel ragionamento prevalesse l'elemento dell'interpretazione, nella quale è il soggetto cognitivo a giocare un ruolo più centrale.

Inoltre, Dinsmore insiste molto sul tipo di relazione che deve sussistere tra lo spazio controfattuale e lo spazio rappresentante il mondo, ma non si sofferma molto ad analizzare il tipo di inferenza che viene condotto all'interno dello spazio controfattuale. Questo disinteresse per la forma "locale" di inferenza non è limitato al caso dei controfattuali, ma è caratteristico di tutta la sua analisi, che è piuttosto incentrata su quali siano gli effetti su uno spazio di ciò che è vero in un altro spazio.

Infine, la teoria di Dinsmore sembra essere sottesa dall'idea che il contesto abbia una sorta di ruolo di "traduzione" delle proposizioni dagli spazi specifici a quello rappresentante il mondo (che Dinsmore chiama *spazio base*). La conseguenza di questa assunzione è che lo spazio base è gerarchicamente superiore a tutti gli altri e l'operazione consistente nel trasferire le proposizioni in questo spazio potrebbe essere equiparata a una sorta di "decontestualizzazione assoluta".

Questa assunzione è ancora troppo forte dal nostro punto di vista, poiché, sebbene possa essere vero che essa cattura bene la posizione del realismo, trascura il fatto che, se si vuole assumere la prospettiva di un agente cognitivo dotato di conoscenza limitata sul mondo e situato in uno specifico ambiente, non è affatto ovvio (anzi, è alquanto improbabile) che si possa mai essere

in grado di esplicitare completamente le dipendenze contestuali dell'informazione ricevuta. Dunque, se lo scopo finale è quello di rappresentare la conoscenza di un agente, questa "assunzione di decontestualizzazione" costituisce una limitazione.

In conclusione, a nostro avviso l'idea di partenza di Dinsmore di rappresentare la conoscenza dell'agente cognitivo è corretta, ma nei fatti essa non può essere completamente realizzata nel suo sistema, poiché in esso si assume la possibilità di decontestualizzare completamente l'informazione. Sarebbe a nostro parere più adeguato un sistema che potesse catturare anche una concezione più relativista della conoscenza. In esso il cosiddetto "spazio del mondo reale" sarebbe comunque rappresentabile, ma come sottocaso nel quale venisse esplicitamente assunta l'esistenza di un tale "contesto privilegiato". Presenteremo nel capitolo 4 un sistema che a noi pare adeguato per descrivere questa più ampia prospettiva.

Capitolo 4

Un modello formale per il ragionamento controfattuale

Tutto quel che inventi è vero, di questo puoi star certo
[Efraim Medina Reyes, *C'era una volta l'amore ma ho dovuto ammazzarlo*, p.159]

In questo capitolo presentiamo il sistema formale che abbiamo deciso di adottare per la trattazione del ragionamento controfattuale, la Semantica a Modelli Locali (SML). Per prima cosa, quindi, presentiamo il sistema stesso, con le sue caratteristiche distintive e la filosofia che lo sottende.

In un secondo momento vengono fornite le definizioni formali più importanti che vengono in seguito riprese, allo scopo di esprimere, attraverso di esse, le principali nozioni che caratterizzano il ragionamento controfattuale. Infine, verranno poste esplicitamente a confronto, attraverso un esempio, le soluzioni offerte dai sistemi più tradizionali con quelle proposte dalla SML, per metterne in luce le differenze più salienti e l'approccio di conseguenza differente alla trattazione di determinati problemi.

4.1 Che cos'è la semantica a modelli locali

Tra la fine degli anni Ottanta e l'inizio dei Novanta emerge, nell'ambito degli studi sull'intelligenza artificiale, l'esigenza di rielaborare i sistemi formali della logica verso prospettive più cognitive.

Tra i problemi che si presentano in quegli anni a chi si occupa di intelligenza artificiale, molti possono essere fatti risalire alla difficile conciliabilità di due fattori:

- da una parte, il paradigma teorico dominante, che vede la conoscenza come un'enorme banca-dati, accessibile a diversi livelli dai vari agenti, ma comunque comune a tutti, così come il corrispettivo linguaggio, mutuato dalla logica classica;

- dall'altra, la realtà dei fatti, ossia agenti che, essendo stati progettati da *designers* diversi, usano una terminologia diversa per esprimere gli stessi concetti (o operazioni), hanno modi diversi di raccogliere e catalogare la conoscenza, e a volte usano perfino regole inferenziali diverse.

Una soluzione a questi e a rompicapo simili è stata offerta, a partire dai primi anni Novanta, da Fausto Giunchiglia e dal gruppo di ricerca da lui fondato, il *Mechanized Reasoning Group* (MRG), con sede a Trento e a Genova.

Gli sforzi di Giunchiglia e del suo gruppo si sono concentrati inizialmente su una soluzione di tipo *sintattico*, concretizzatasi nell'elaborazione dei *MultiLanguage Systems* – MLS – (Sistemi MultiLinguaggio, vedi [69], [35], [75], [74], [142]), che in seguito si sono evoluti in *MultiContext Systems* – MCS – (Sistemi MultiContestuali, presentati in [68], ma anche attraverso esposizioni meno tecniche e più discorsive, come [70], [71], [72]). Nel corso degli anni il gruppo di ricerca ha poi fornito una vasta produzione di articoli su molte possibili applicazioni e sugli svariati problemi (per esempio il ragionamento su azioni, o metateorico, o in contesti modali o con modalità epistemiche) ai quali è possibile offrire una soluzione in linea con le idee contenute nei MCS, come nel caso di [7], [8], [9], [20], [18], [65], [76], [77].

Una volta consolidata la parte sintattica, il gruppo si è poi rivolto verso la costruzione della semantica (la *Local Models Semantics* o Semantica a Modelli Locali – SML), a partire dal 1994, con [64], e poi con lavori quali [11], [21], [73], [141], per finire con l'articolo che può essere considerato la sistematizzazione più completa al momento disponibile, ovvero [66].

Questo interesse per la sfera più "cognitiva" si è imposto perché gli agenti artificiali, nel riprodurre le forme di ragionamento tipiche del senso comune,

hanno bisogno al tempo stesso e in eguale misura sia del rigore delle formalizzazioni logiche, in vista del raggiungimento di risultati sempre più precisi, sia di una prospettiva più "situata", dal momento che questi agenti si ritrovano poi a operare in un ambiente e in concomitanza con altri agenti.

Nel momento stesso in cui si tenta di coniugare questi due approcci, sorgono immediatamente una serie di problematiche la cui risoluzione è fondamentale per riuscire a realizzare agenti artificiali dotati di qualcosa che possa essere definito sotto qualche rispetto intelligenza.

In particolare, su una di queste tematiche, quella della località, si è concentrata l'attenzione del gruppo di ricerca di Giunchiglia. Ma in che cosa consiste esattamente questo problema della località? In poche parole, esso è riassumibile nell'idea secondo la quale, quando ragiona, un agente non utilizza mai tutto l'insieme di conoscenza a lui disponibile, ma solamente una piccola parte di essa. Non solo, quando ragiona su un particolare problema utilizza delle regole, dei presupposti e delle forme di ragionamento che vengono lasciate completamente in secondo piano quando affronta un altro specifico problema, perché inutili o addirittura in conflitto con quelli che usa in questo secondo caso.

Ecco come Giunchiglia ha espresso questa idea in uno dei suoi primi articoli sull'argomento, *Contextual Reasoning*:

> Our basic intuition is that reasoning is usually performed on a subset of the global knowledge base: we never consider *all we know* but only a very small subset of it[1].
>
> [Contextual Reasoning, p.2]

Queste "porzioni" della conoscenza globale degli agenti sono ciò che Giunchiglia chiama *contesti*. Questa nozione presenta molte affinità con i concetti di spazio mentale e rappresentazione ripartita presenti nei lavori di Fauconnier e Dinsmore, ma se ne differenzia in quanto il contesto non è un generico spazio cognitivo entro cui si sviluppa il ragionamento, bensì una *teoria*.

[1]La nostra intuizione di base è che normalmente il ragionamento è condotto su un sottoinsieme della conoscenza di base globale: non consideriamo mai *tutto ciò che sappiamo* ma solo un sottoinsieme molto piccolo della nostra conoscenza. [*traduzione mia*]

Dovendo questa teoria rappresentare la prospettiva dell'agente ragionante, essa sarà limitata in due direzioni: in ampiezza – e sarà quindi *parziale*, cioè lascerà indeterminati tutti quegli elementi che esulano dalla conoscenza dell'agente e da ciò che l'agente ritiene rilevante per il ragionamento in corso – e in profondità – e sarà quindi *approssimata*, nel senso che i ragionamenti all'interno di essa avverranno sempre a un determinato livello di dettaglio, commisurato al grado di conoscenza dell'agente sull'oggetto e all'obiettivo del ragionamento. Sarà infine *prospettica*, nel senso che rappresenterà la prospettiva dell'agente ragionante su quel problema.

Il contesto risulterà quindi essere una teoria che è al tempo stesso parziale, approssimata e prospettica, come è spiegato in dettaglio in [10].

> **Partiality** We say that *a representation is partial when it describes only a subset of a more comprehensive state of affairs* [...]
>
> **Approximation** We say that *a representation is approximate when it abstracts away some aspects of a given state of affairs* [...]
>
> **Perspective** We say that *a representation is perspectival when it encodes a spatio-temporal, logical, and cognitive point of view on a state of affairs*[2].
>
> [Contextual Reasoning Distilled, pp.9–11]

Un altro elemento che vale la pena rimarcare è la doppia valenza della parzialità dei contesti, i quali sono quindi parziali in due sensi: nel senso che essi descrivono solo una porzione limitata della realtà, e quindi ci saranno degli elementi di essa che non saranno presi nemmeno in considerazione e nel senso che, essendo la conoscenza di ogni agente imperfetta, ci saranno

[2] **Parzialità** Diciamo che *una rappresentazione è parziale quando descrive solo una sottoparte di uno stato di fatto più comprensivo* [...] **Approssimazione** Diciamo che *una rappresentazione è approssimata quando astrae da alcuni aspetti di un dato stato di fatto* [...] **Prospettiva** [...] Diciamo che *una rappresentazione è prospettica quando codifica un punto di vista spazio-temporale, logico e cognitivo su uno stato di fatto.* [*traduzione mia*]

certe asserzioni relative alla porzione in esame alle quali l'agente non sarà in grado né di dare né di rifiutare l'assenso. Sia questi asserti che le loro negazioni saranno quindi compatibili con la conoscenza di base dell'agente e formalmente questo si tradurrà nel fatto che un contesto rappresentante la conoscenza dell'agente su quel determinato argomento conterrà dei modelli locali[3] in cui un dato enunciato di questi sarà vero e dei modelli locali in cui esso sarà falso.

Una volta che la conoscenza degli agenti è stata ripartita in contesti, segue in maniera abbastanza consequenziale che ogni processo di ragionamento è di natura contestuale; può tuttavia essere di due tipi: intracontestuale o intercontestuale. Intracontestuale è il ragionamento che si svolge tutto all'interno di un unico contesto, in osservanza del cosiddetto *principio di località*; intercontestuale è il ragionamento che si svolge attraverso più contesti legati da un qualche tipo di relazione, cioè in osservanza del *principio di compatibilità*.

A questo punto è indispensabile enunciare i due principi fondamentali del ragionamento contestuale, il *principio di località* e il *principio di compatibilità* [66].

> 1. **Principle (of Locality):** reasoning uses only part of what is potentially available (e.g., what is known, the available inference procedures). The part being used while reasoning is what we call *context* (of reasoning);
> 2. **Principle (of Compatibility):** there is *compatibility* among the reasoning performed in different contexts[4].
>
> [Local Model Semantics, or Contextual Reasoning = Locality + Compatibility, p.2]

Ciò che ci preme soprattutto mettere in evidenza è che la semantica a modelli

[3]Per una definizione della nozione di *modello locale* si rimanda alla sezione 4.2.

[4]**Principio di Località**: il ragionamento usa solo parte di ciò che è potenzialmente disponibile (cioè ciò che è conosciuto, le procedure di inferenza disponibili). La parte che viene usata è ciò che chiamiamo *contesto* di ragionamento.

Principio di Compatibilità: esiste *compatibilità* tra il ragionamento condotto in differenti contesti. [*traduzione mia*]

locali non è una teoria del contesto in senso stretto, essa è piuttosto una teoria del ragionamento contestuale.

Questo significa che ciò di cui la teoria veramente si occupa è il processo dinamico di ragionamento, non solo la ripartizione statica della conoscenza nello spazio cognitivo.

Le caratteristiche per noi più interessanti di questa semantica sono due, la prima legata al principio di località e la seconda legata al principio di compatibilità.

Il principio di località afferma che ogni contesto è una teoria a sé stante; questo formalmente significa che ogni contesto ha un suo proprio linguaggio (che può essere differente dai linguaggi degli altri contesti), un suo proprio insieme di assiomi (cioè i principi immutabilmente veri possono variare da contesto a contesto) e delle proprie regole di inferenza (e quindi le procedure di ragionamento accettate come buone possono non essere sempre le stesse).

Il principio di compatibilità afferma che esistono determinati vincoli tra i contesti che fanno sì che le conclusioni raggiunte in un contesto ragionando solo localmente possono essere modificate in seguito all'applicazione dei vincoli di compatibilità indotta dalla relazione esistente con altri contesti.

Mostreremo ora due esempi che hanno lo scopo di chiarire un po' meglio la funzione e l'uso dei principi appena presentati.

Il primo esempio è relativo al ragionamento su punti di vista. Immaginiamo di avere due agenti, *Mr.1* e *Mr.2*, che stanno osservando una "scatola magica" come quella nella figura 4.1, che ha determinate curiose proprietà, quali l'essere divisa in sei settori, completamente trasparente, ma tale per cui non sia possibile distinguere a che livello di profondità sono posizionati gli oggetti al suo interno.

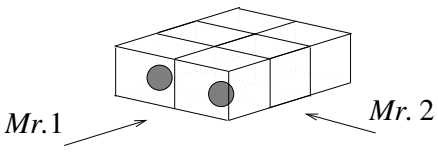

Figura 4.1: La scatola magica

All'interno della scatola vengono poste delle sfere, ma non potendo distin-

guere (né *Mr.1* né *Mr.2*) a che profondità queste si trovino, tutto ciò che *Mr.1* può legittimamente dire è se veda o meno delle sfere nella sezione di destra o in quella di sinistra, mentre ciò che *Mr.2* può affermare è se veda o meno sfere nella sezione di destra, in quella centrale, o in quella di sinistra. Utilizziamo le lettere predicative D, C e S per indicare che gli agenti vedono una sfera rispettivamente nella sezione di destra, in quella di centro o in quella di sinistra.

Nella fattispecie, se le sfere sono posizionate nella scatola come nella figura 4.1, la situazione viene descritta in due modi diversi dai due agenti: *Mr.1* utilizza la formula $(D \wedge S)$, mentre *Mr.2* esprime ciò che vede con la formula $(S \wedge \neg C \wedge \neg D)$.

Vediamo ora in che modo i due principi enunciati sopra si manifestano. Il principio di località, per esempio, si manifesta nelle seguenti osservazioni:

- *Mr.1* vede due sfere nella scatola, mentre *Mr.2* ne vede solo una;
- Essendo la scatola, vista dalla prospettiva di *Mr.1*, composta da due soli settori, non ha senso per quest'ultimo il concetto di "centro" relativamente alla situazione osservata. Di conseguenza, il suo linguaggio conterrà le lettere predicative D e S, ma non C. C sarà invece contenuta, insieme a D e S, nel linguaggio di *Mr.2*;
- *Mr.1* e *Mr.2* possono affermare, entrambi correttamente e contemporaneamente, D e $\neg D$ rispettivamente, poiché D assume un diverso significato nei due casi;
- la posizione della "stessa" sfera (quella che entrambi gli agenti vedono) è correttamente descritta con la formula D da *Mr.1* e con la formula S da *Mr.2*.

Il principio di località mostra che la conoscenza che gli agenti ricavano dall'osservazione della situazione non è la stessa e in questo caso è incompleta per entrambi; inoltre i due linguaggi sono distinti, come mostrano in particolar modo gli ultimi tre punti sopra elencati.

Il principio di compatibilità, invece, fa sì che gli agenti possano ragionare sulle relazioni che intercorrono tra i loro diversi punti di vista; tali relazioni

discendono ovviamente dal fatto che i due agenti si trovano, in effetti, di fronte alla stessa situazione. Infatti, in questo caso,

- se *Mr.1* non vede nessuna sfera, allora neanche *Mr.2* ne può vedere alcuna (e viceversa);
- se *Mr.1* vede una sfera sulla sinistra o sulla destra, allora anche *Mr.2* vede una sfera o sulla sinistra, o al centro o sulla destra (per le proprietà magiche della scatola);
- allo stesso modo, se *Mr.2* vede una sfera sulla sinistra, al centro o sulla destra, allora anche *Mr.1* vede una sfera sulla sinistra o sulla destra.

Queste correlazioni tra i due punti di vista, che possiamo chiamare più appropriatamente *relazioni di compatibilità* sono rappresentate nella figura 4.2.

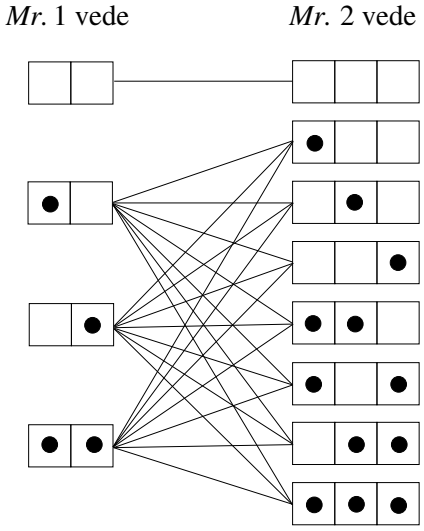

Figura 4.2: Compatibilità tra punti di vista

Il secondo esempio, in qualche modo collegato al primo, è basato sui contesti di credenza e si prefigge di presentare i contesti delle credenze di un terzo agente, ϵ, e delle credenze che questi attribuisce a *Mr.1* e *Mr.2*.

Se ϵ attribuisce a *Mr.1* la credenza che ci sia una sfera sulla destra e a *Mr.2* la credenza che ci sia una sfera sulla sinistra, il contesto che rappresenta

le credenze di ϵ conterrà le due seguenti credenze: "*Mr.1* crede che ci sia una sfera sulla destra" e "*Mr.2* crede che ci sia una sfera sulla sinistra"; il contesto delle credenze che ϵ attribuisce a *Mr.1* conterrà la credenza "C'è una sfera sulla destra" e il contesto delle credenze che ϵ attribuisce a *Mr.2* conterrà la credenza "C'è una sfera sulla sinistra". I tre contesti sono rappresentati nella figura 4.3.

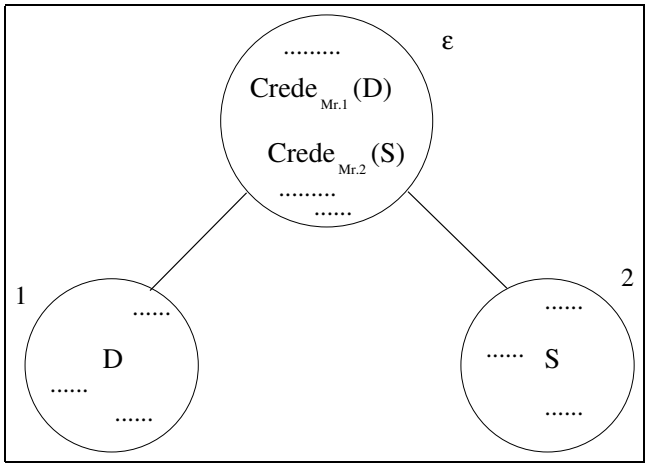

Figura 4.3: Contesti di credenza (SMC)

La spiegazione della figura 4.3 è molto semplice: il cerchio contrassegnato con ϵ rappresenta il contesto delle credenze di ϵ. Le formule $Crede_{Mr.1}(D)$ e $Crede_{Mr.2}(S)$ significano rispettivamente che ϵ crede che *Mr.1* creda D (cioè che ci sia una sfera sulla destra) e che ϵ crede che *Mr.2* creda S (ossia che ci sia una sfera a sinistra). Il cerchio contrassegnato da 1 rappresenta il contesto delle credenze che ϵ attribuisce a *Mr.1* e quindi la presenza di D al suo interno sta a significare che ϵ attribuisce a *Mr.1* la credenza che ci sia una sfera a destra. Il cerchio contrassegnato con 2 rappresenta il contesto delle credenze che ϵ attribuisce a *Mr.2* e quindi la presenza di S in esso sta a significare che ϵ attribuisce a *Mr.2* la credenza che ci sia una sfera sulla sinistra.

Da notare, sempre rispetto alla figura, che il contesto indicato da 1 *non* rappresenta ciò che *Mr.1* crede, ma ciò che ϵ crede che *Mr.1* creda; analogamente per *Mr.2*.

Anche in questo esempio è manifesta l'azione dei due principi: per cominciare, quello di località permette di mantenere distinte le credenze di ϵ da quelle che egli attribuisce agli altri due agenti:

- la credenza $Crede_{Mr.1}(D)$, ossia la credenza che *Mr.1* creda che ci sia una sfera a destra non appartiene a 2, cioè all'insieme delle credenze che ϵ attribuisce a *Mr.2*, ma nemmeno a 1, l'insieme delle credenze che ϵ attribuisce a *Mr.1*;

- ϵ può attribuire a *Mr.1* la credenza D e contemporaneamente attribuire a *Mr.2* la credenza $\neg D$;

- nel contesto 1 (delle credenze che ϵ attribuisce a *Mr.1*) non possono esserci né C né $\neg C$ poiché postuliamo che ϵ sappia che *Mr.1* non ha una nozione di centro.

Anche in questo caso, il principio di compatibilità serve a mettere in luce le interconnessioni tra contesti:

- la relazione che sussiste tra il contesto ϵ e il contesto 1 determina il fatto che poiché in ϵ si ha la formula $Crede_{Mr.1}(D)$, in 1 si avrà la formula D;

- la relazione che sussiste tra il contesto ϵ e il contesto 2 determina il fatto che poiché in ϵ si ha la formula $Crede_{Mr.2}(S)$, in 2 si avrà la formula S;

- ...

Naturalmente la struttura che abbiamo qui rappresentato è una versione molto semplificata, poiché presumibilmente ϵ si rappresenterà in qualche modo anche ciò che *Mr.1* crede di credere, o ciò che *Mr.1* crede che *Mr.2* creda, ciò che *Mr.2* crede che *Mr.1* creda e così via. La struttura può quindi diventare anche molto complessa e ramificata.

In conclusione, un contesto è quindi immerso in una struttura di relazioni con altri contesti, la quale influenza il ragionamento.

Riteniamo queste caratteristiche molto importanti sia da un punto di vista formale, sia da un punto di vista intuitivo e crediamo che esse possano rendere conto in maniera soddisfacente dei processi di ragionamento di senso comune e, come vedremo meglio nelle sezioni 4.3 e 4.5, in particolare dei processi di ragionamento di carattere controfattuale.

4.2 Qualche definizione nella semantica a modelli locali

In questa sezione elenchiamo una serie di definizioni, che sono state fornite in [66], relative a tutte le nozioni principali della SML e che verranno riutilizzate in seguito nella sezione 4.3, opportunamente riadattate, per il ragionamento controfattuale.

Partendo dall'idea intuitiva di contesto appena definita, ossia una teoria avente un linguaggio, degli assiomi e delle regole di inferenza propri, le prime definizioni da fornire saranno quelle di linguaggio e modello.

Definizione 4.2.1 (Famiglia di linguaggi) *$\{L_i\}_{i\in I}$ è una famiglia di linguaggi definiti su un insieme di indici I e ogni L_i è un linguaggio formale usato per dire che cosa è vero in un contesto*

Definizione 4.2.2 (Classe di interpretazioni – per L_i) *M_i è la classe di tutti i possibili modelli (interpretazioni) per L_i*

Definizione 4.2.3 (Modello locale) *Ogni*

$$m \in M_i$$

è un modello locale di L_i, dove ciascun modello locale è un classico modello à la Tarski.

Definizione 4.2.4 (Sequenza di compatibilità) *Una sequenza di compatibilità $\boldsymbol{s}$ (per l'insieme di linguaggi $\{L_i\}$) è una sequenza*

$$\boldsymbol{s} = \langle \boldsymbol{c}_0, \boldsymbol{c}_1, \ldots, \boldsymbol{c}_i \rangle$$

dove, per ogni $i \in I$, $\mathbf{c}_i$ è un sottoinsieme di M_i. Chiamiamo $\mathbf{c}_i$ l'i-esimo elemento di $\mathbf{s}$. Ogni $\mathbf{c}_i$ è dunque formato da un insieme di modelli locali per L_i.

La figura 4.4 mostra le relazioni esistenti tra linguaggi, modelli e sequenze di compatibilità.

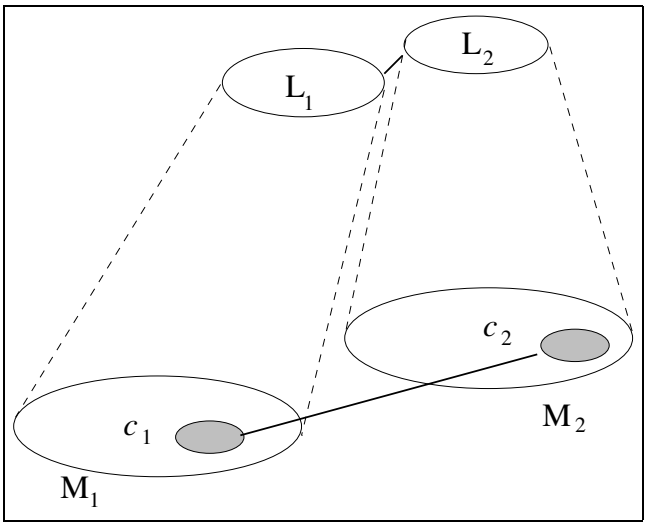

Figura 4.4: Relazione di compatibilità

Definizione 4.2.5 (Relazione di compatibilità) *Una relazione di compatibilità* $\boldsymbol{C}$ *(per* $\{L_i\}$*) è un insieme* $\boldsymbol{C} = \{\boldsymbol{s}\}$ *di sequenze di compatibilità* $\boldsymbol{s}$*. La relazione di compatibilità* $\boldsymbol{C}$ *è una relazione del tipo:*

$$\boldsymbol{C} \subseteq \prod_{i \in I} 2^{M_i}$$

dove $\prod_{i \in I} 2^{M_i}$ è il prodotto cartesiano della collezione $\{2^{M_i} : i \in I\}$; $\mathbf{C}$ è quindi un sottoinsieme di tutte le possibili combinazioni di modelli locali.

Definizione 4.2.6 (Modello) *Un modello (per* $\{L_i\}$*) è una relazione di compatibilità* $\boldsymbol{C}$ *tale che:*

- $\boldsymbol{C} \neq \emptyset$
- $\langle \emptyset, \emptyset, \ldots, \emptyset, \ldots \rangle \notin \boldsymbol{C}$

Definizione 4.2.7 (Contesto) *Dato un modello* $\boldsymbol{C} = \{\langle \boldsymbol{c}_0, \boldsymbol{c}_1, \ldots, \boldsymbol{c}_i, \ldots \rangle\}$ *definiamo formalmente un contesto come ogni* $\boldsymbol{c}_i$*, cioè l'insieme dei modelli locali* $m \in M_i$ *permessi da* $\boldsymbol{C}$ *sotto una particolare sequenza di compatibilità.*

Forniamo ora altre tre importanti definizioni, quelle di soddisfacibilità, validità e conseguenza logica, che hanno una valenza molto generale e possono essere utilizzate anche in domini specifici (come quello del ragionamento controfattuale) semplicemente sostituendo le relazioni di compatibilità del caso alla generica **C**.

Definizione 4.2.8 (Soddisfacibilità)
Sia $\boldsymbol{C} = \{\boldsymbol{s}\}$ *con* $\boldsymbol{s} = \langle\ \boldsymbol{c}_0,\ \boldsymbol{c}_1,\ \ldots,\ \boldsymbol{c}_i,\ \ldots\ \rangle$ *un modello e* $i : \Phi$ *una formula.* $\boldsymbol{C}$ *soddisfa* $i : \Phi$*, in simboli* $\boldsymbol{C} \models i : \Phi$ *se, per tutte le* $\boldsymbol{s}$ *in* $\boldsymbol{C}$

$$\boldsymbol{s} \models \Phi$$

dove $\boldsymbol{s} \models \Phi$ *se, per tutti gli* $m \in \boldsymbol{c}_i$*,* $m \models \Phi$

Un modello soddisfa una formula se tutte le sequenze di compatibilità che lo compongono la soddisfano e una sequenza di compatibilità, a sua volta, soddisfa una formula se tutti i modelli locali di tutti i contesti che la compongono soddisfano tale formula.

Definizione 4.2.9 (Validità) *Una formula* $i : \Phi$ *è valida, in simboli* $\models i : \Phi$ *se tutti i modelli soddisfano* $i : \Phi$

Definizione 4.2.10 (Conseguenza logica rispetto a un modello)
Una formula $i : \Phi$ *è una conseguenza logica di un insieme di formule* Γ *rispetto a un modello* $\boldsymbol{C}$*, in simboli* $\Gamma \models_{\boldsymbol{C}} i : \Phi$ *se ogni sequenza* $\boldsymbol{s} \in \boldsymbol{C}$ *soddisfa:*

$$\forall j \in I, j \neq i, \boldsymbol{c}_j \models \Gamma_j \Longrightarrow (\forall m \in \boldsymbol{c}_i, m \models \Gamma_i \Longrightarrow m \models \Phi)$$

Intuitivamente, la parte sinistra della definizione risponde al principio di compatibilità e seleziona le sequenze di compatibilità che soddisfano Γ_j, mentre la parte destra risponde al principio di località e seleziona, all'interno di $\mathbf{c}_i$, i modelli locali che soddisfano Φ.

Come accennato in precedenza, la semantica a modelli locali ha un corrispettivo sintattico, i cosiddetti *Sistemi MultiContesto*, formalizzati in [142] e sui quali molto lavoro è stato fatto[5].

Per una disamina più approfondita si rimanda direttamente agli articoli segnalati, ma vale la pena almeno elencare le principali definizioni.

Per cominciare, un Sistema MultiContesto (la controparte sintattica di un modello) è definito come un insieme di contesti (dove ogni contesto è un sistema assiomatico standard) e un insieme di regole, dette *Regole Ponte* (o *Bridge Rules*) che permettono di compiere operazioni attraverso contesti, ossia processi in cui premesse e conclusioni si trovino in contesti diversi.

Definizione 4.2.11 (Sistema MultiContesto (SMC)) *Sia I una famiglia di indici. Un* Sistema MultiContesto *è definito come una coppia:*

$$\langle \{C_i\}_{i \in I}, BR \rangle$$

dove $\{C_i\}_{i \in I}$ è un insieme di contesti e BR un insieme di regole ponte.

A sua volta un contesto è definito come un sistema formale assiomatico:

Definizione 4.2.12 (Contesto) *Sia L un linguaggio formale, $\Omega \subseteq L$ un insieme di assiomi in L e Δ un insieme di regole di inferenza definite su L. Un* contesto *c è definito come la tripla:*

$$\langle L, \Omega, \Delta \rangle$$

dove L è detto il linguaggio di c, Ω è detto l'insieme di assiomi di c e Δ è l'apparato deduttivo di c.

Le regole ponte (corrispondenti alle relazioni di compatibilità) sono invece così definite:

[5] Cfr., per esempio, [68],[70], [71], [72].

Definizione 4.2.13 (Regola ponte) *Una* regola ponte *è una regola della forma:*

$$\frac{c_{n+1} : \Phi_{n+1}}{c_1 : \Phi_1, \ldots, c_n : \Phi_n}$$

dove $c_1, \ldots, c_{n+1}$ *sono contesti,* $c_{n+1} \neq c_i$ $(i = 1, \ldots, n)$ *e* $\Phi_1, \ldots, \Phi_{n+1}$ *sono formule appartenenti ai linguaggi di* $c_1, \ldots, c_{n+1}$ *rispettivamente.*

È stato dimostrato in [142] che questi sistemi formali sono corretti e completi.

4.3 Una semantica a modelli locali per il ragionamento controfattuale

Analogamente a quanto fatto prima per le definizioni generali, poniamo che L_a sia il linguaggio della teoria c_a che contiene le credenze dell'agente a, (ossia, contiene i termini del linguaggio e gli assiomi, cioè i fatti e le leggi che l'agente considera valere *sempre*). $L_F \subseteq L_a$ è il linguaggio di c_F, ossia del contesto che l'agente utilizza per ragionare su un problema specifico ed esprime i fatti che l'agente ritiene essere veri nella situazione sulla quale sta ragionando; $L_{CF} \subseteq L_a$ è il linguaggio di c_{CF}, ossia del contesto ipotetico che l'agente costruisce per ragionare a partire dall'ipotesi controfattuale. Per semplicità, assumiamo che L_a, L_F e L_{CF} siano tutti proposizionali, con l'unica aggiunta, per L_a e per L_F, di tre operatori modali: $\otimes(A, C)$, che traduce il legame controfattuale tra A e C – "Se fosse successo A sarebbe successo C", $\odot(A, C)$, che traduce la possibilità controfattuale di C dato A – "Se fosse successo A avrebbe potuto succedere C" e $\oslash(A, C)$, che traduce il legame semifattuale tra A e C – "Se anche fosse successo A comunque non sarebbe successo C", quindi $L_{CF} \subset L_F \subseteq L_a$.

Definiamo M_a come la classe di tutti i possibili modelli (interpretazioni) per L_a. Gli elementi sono tutti gli $m_a \in M_a$ sono detti anche *modelli locali* (per L_a). Allo stesso modo, M_F contiene tutti gli $m_F \in M_F$, possibili interpretazioni di L_F e M_{CF} contiene tutti gli $m_{CF} \in M_{CF}$, possibili interpretazioni di L_{CF}.

Ora, il processo che si sta tentando di caratterizzare è quello di un agente che, a partire da due fatti che conosce (o crede di conoscere) sulla realtà relativa a un dato problema, ipotizza che per uno di questi fatti (la premessa) si inverta il valore di verità, ragiona a partire da questa premessa e decide se, controfattualmente, la conclusione continua a mantenere il valore di verità originale anche sotto la nuova ipotesi o se lo cambia anch'essa.

Per comodità, assumiamo la prospettiva in cui l'agente parte da fatti che non si sono verificati (e quindi $\neg A$ e $\neg C$) e ipotizza che il primo, A, si verifichi. In altri termini, siamo nella situazione in cui l'agente si chiede: "Se fosse successo A sarebbe successo C?".

Questo processo, rappresentato nella figura 4.5 è reso formalmente possibile dalla relazione di controfattualità che, attraverso l'imposizione di alcuni vincoli che essi devono soddisfare, definisce, all'interno di M_F, il contesto fattuale (c_F) e all'interno di M_{CF} il contesto controfattuale (c_{CF})[6], in altri termini, la relazione di controfattualità fissa le condizioni che due contesti devono soddisfare perché possano essere considerati uno controfattuale dell'altro. A questo punto è sufficiente verificare qual è il valore di verità del conseguente in tutti i modelli locali di c_{CF}.

Diamo ora le principali definizioni della semantica a modelli locali relative al ragionamento controfattuale.

Definizione 4.3.1 (Coppia di controfattualità) *Una coppia di controfattualità $s_{(A,C)}$ rispetto a due fatti A e C è una coppia (ossia una sequenza con due elementi) di compatibilità*

$$s_{(A,C)} = \langle c_F, c_{CF} \rangle$$

dove c_F e c_{CF} sono sottoinsiemi rispettivamente di M_F e di M_{CF} che soddisfano il seguente vincolo:

$$\textit{Se } c_F \models \neg A \wedge \neg C, \textit{ allora } c_{CF} \models A$$

Definizione 4.3.2 (Relazione di controfattualità) *Una relazione di controfattualità $\Re_{(A,C)}$ rispetto a due fatti A e C è un insieme $\{s_{(A,C)}\}$ di*

[6]Da non dimenticare che la definizione di fattualità e controfattualità è, nel nostro sistema, completamente epistemica.

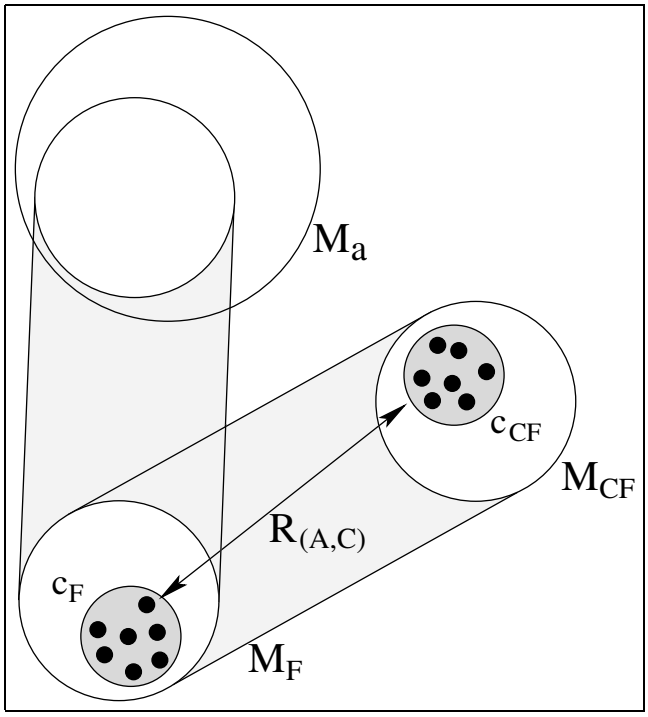

Figura 4.5: Coppia di controfattualità

coppie di controfattualità per A e C, come mostrato nella figura 4.6. È quindi una relazione del tipo:

$$\Re_{(A,C)} \subseteq 2^{M_F} \times 2^{M_{CF}}$$

Anche in questo caso, quindi, $\Re_{(A,C)}$ è sottoinsieme di tutte le possibili combinazioni di modelli locali.

Definizione 4.3.3 (Modello controfattuale) *Un modello controfattuale è una relazione di controfattualità $\Re_{(A,C)}$ tale che:*

- $\Re_{(A,C)} \neq \emptyset$
- $\langle \emptyset, \emptyset, \ldots, \emptyset, \ldots \rangle \notin \Re_{(A,C)}$

Definizione 4.3.4 (Contesto fattuale) *Data una relazione di controfattualità $\Re_{(A,C)}$, definiamo contesto fattuale ogni c_F, cioè l'insieme dei modelli locali $m_F \in c_F$ permessi da $\Re_{(A,C)}$ all'interno di ogni singola coppia di controfattualità.*

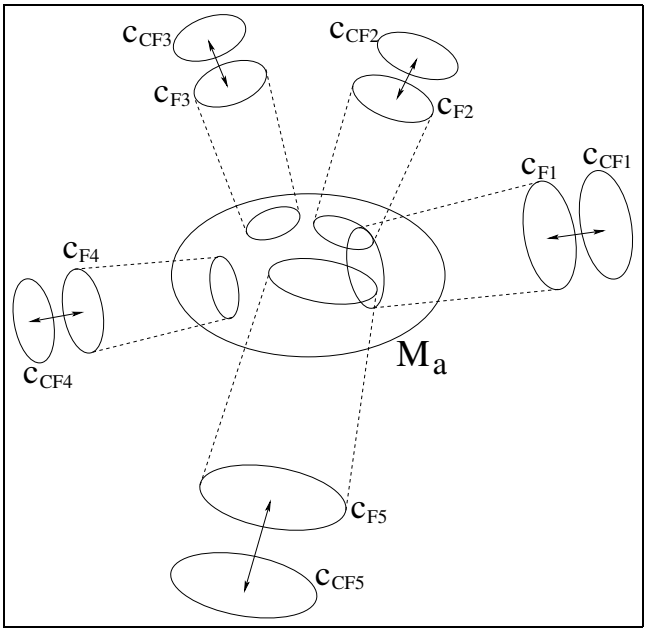

Figura 4.6: Relazione di controfattualità

Definizione 4.3.5 (Contesto controfattuale) *Data una relazione di controfattualità $\Re_{(A,C)}$, definiamo contesto controfattuale ogni c_{CF}, cioè l'insieme dei modelli locali $m_{CF} \in c_{CF}$ permessi da $\Re_{(A,C)}$ all'interno di ogni singola coppia di controfattualità.*

Nel seguito, qualora questo non dia adito ad ambiguità, useremo le notazioni c_F e c_{CF} indifferentemente per indicare i contesti fattuale e controfattuale oppure i modelli locali facenti parte dei rispettivi contesti.

La definizione forse più interessante che vorremmo fornire è quella di *conseguenza controfattuale*. Questa definizione muove dall'assunto che esistano processi di ragionamento nei quali premesse e conclusione appartengono a contesti diversi. Nel caso dei controfattuali questa peculiarità si traduce nel fatto che, per poter inferire qualcosa a livello ipotetico (nel contesto controfattuale) è necessario partire da fatti contenuti nel contesto fattuale. Il risultato di un ragionamento controfattuale (cioè l'assegnazione di un valore di verità a un enunciato controfattuale nella teoria specifica nella quale si sta ragionando) è ottenuto da un'operazione *su teorie* e precisamente dal-

la selezione di alcuni modelli locali all'interno del modello della teoria che soddisfano determinati vincoli.

Una volta fornita la definizione di conseguenza controfattuale, non è difficile ottenere le nozioni correlate di *possibilità controfattuale* (ossia, l'analogo di ciò che Lewis definisce *might counterfactuals*, controfattuali aventi la forma: "Se fosse successo A, avrebbe potuto succedere C") e di *conseguenza semifattuale* ("Se anche fosse successo A, non sarebbe comunque successo C").

Definizione 4.3.6 (Conseguenza controfattuale) *C segue controfattualmente da A in una particolare situazione, descritta da un contesto fattuale c_F, ovvero*

$$c_F \models \otimes(A, C)$$

sse per ogni c_{CF} tale che c_F e c_{CF} sono una coppia di controfattualità per (A, C), $\forall m \in c_{CF}, m \models C$

Definizione 4.3.7 (Possibilità controfattuale) *C potrebbe seguire controfattualmente da A in una particolare situazione, descritta da un contesto fattuale c_F, ovvero*

$$c_F \models \odot(A, C)$$

sse esiste un c_{CF} tale che c_F e c_{CF} sono una coppia di controfattualità per (A, C), $\forall m \in c_{CF}, m \models C$

Definizione 4.3.8 (Conseguenza semifattuale) *C segue semifattualmente da A in una particolare situazione, descritta da un contesto fattuale c_F, ovvero*

$$c_F \models \oslash(A, C)$$

sse per ogni c_{CF} tale che c_F e c_{CF} sono una coppia di controfattualità per (A, C), $\forall m \in c_{CF}, m \models \neg C$

Definizione 4.3.9 (Conseguenza controfattuale analitica) *C segue controfattualmente da A analiticamente, ovvero*

$$c_a \models \otimes(A, C)$$

sse per ogni coppia di controfattualità per (A, C), $c_F \models \otimes(A, C)$

Da notare che, secondo queste definizioni, un enunciato controfattuale non vero non equivale necessariamente a un enunciato controfattuale falso. Infatti, tale enunciato potrebbe:

1. semplicemente non essere controfattuale, nel senso di non rispettare i vincoli imposti ai due contesti, fattuale e controfattuale;

2. essere indeterminato (nel caso in cui il conseguente risultasse vero in alcuni modelli locali del contesto controfattuale e falso in altri);

3. essere falso e in tal caso sarebbe vero l'enunciato semifattuale avente lo stesso antecedente.

Possono inoltre essere enunciate delle relazioni che legano i tre operatori modali:

- $\otimes(A, C) \rightarrow \odot(A, C)$: "Se fosse successo A sarebbe successo C" implica "Se fosse successo A avrebbe potuto succedere C". Ovvero, se tutti i modelli locali in c_{CF} soddisfano C, allora almeno un modello locale in c_{CF} soddisfa C;

- $\otimes(A, C) \rightarrow \neg\oslash(A, C)$: "Se fosse successo A sarebbe successo C" implica che non è vero che "Se anche fosse successo A non sarebbe comunque successo C". Ovvero se tutti i modelli locali in c_{CF} soddisfano C, allora non è vero che nemmeno un modello locale in c_{CF} soddisfa C;

- $\neg \otimes (A, C) \rightarrow \odot(A, C) \vee \oslash(A, C)$: "Non è vero che se fosse successo A sarebbe successo C" implica che "Se fosse successo A avrebbe potuto succedere C" oppure che "Se anche fosse successo A non sarebbe comunque successo C". Ovvero, se non è vero che tutti i modelli locali in c_{CF} soddisfano C, allora o almeno uno di essi soddisfa C, oppure nessuno di essi soddisfa C;

- $\otimes(A, C) \leftrightarrow \oslash(A, \neg C)$: "Se fosse successo A sarebbe successo C" è equivalente a "Se anche fosse successo A non sarebbe comunque successo $\neg C$". Ovvero, se tutti i modelli locali in c_{CF} soddisfano C, allora nessun modello locale in c_{CF} soddisfa $\neg C$ e viceversa;

- $\neg \odot (A, C) \leftrightarrow \oslash(A, C)$: "Non è vero che se fosse successo A avrebbe potuto succedere C" è equivalente a "Se anche fosse successo A non sarebbe comunque successo C". Ovvero, se non è vero che almeno un modello locale in c_{CF} soddisfa C, allora nessun modello locale in c_{CF} soddisfa C e viceversa;

- $\neg \oslash (A, C) \leftrightarrow \odot(A, C)$: "Non è vero che se anche successo A sarebbe successo C" è equivalente a "Se fosse successo A avrebbe potuto succedere C". Ovvero, se non è vero che nessun modello locale in c_{CF} soddisfa C, allora almeno un modello locale in c_{CF} soddisfa C e viceversa.

Per quanto riguarda invece le nozioni di soddisfacibilità, validità e conseguenza logica, queste restano le stesse enunciate in [66] per il caso generico, con l'unico accorgimento di sostituire la relazione di controfattualità $\Re_{(A,C)}$ alla generica relazione di compatibilità $\mathbf{C}$.

4.4 Semantica a modelli locali per i controfattuali: un esempio analizzato con la SML

Ci serviremo ora di un esempio classico, proposto da Kit Fine, quello di "Nixon e l'olocausto", per mostrare come, nella nostra trattazione, sia semplice rappresentare il fatto che lo stesso enunciato controfattuale, a seconda di quali elementi l'agente decida di importare nel contesto di ragionamento, possa risultare vero, falso o indecidibile. L'enunciato dell'esempio recita così: "Se Nixon avesse premuto il bottone, ci sarebbe stato l'olocausto nucleare", dove il bottone in questione, collocato nella famigerata stanza dei bottoni della Casa Bianca, sarebbe collegato a un sistema di lancio di testate nucleari.

Esempio 4.4.1 *Se Nixon avesse premuto il bottone, ci sarebbe stato l'olocausto?*

In questo esempio molto semplice, l'agente a sa che Nixon non ha premuto il bottone e che non c'è stato l'olocausto e si domanda se ci sarebbe stato l'olocausto se solo Nixon avesse premuto il bottone e tutta la sua teoria sul problema ammonta a un solo assioma che dice cha la pressione del bottone causa l'olocausto.

La teoria contiene dunque un solo assioma: $B \rightarrow O$ e tutti i suoi possibili modelli locali sono dunque:

- $m_{a_1} = B \wedge O$
- $m_{a_2} = \neg B \wedge O$
- $m_{a_3} = \neg B \wedge \neg O$

Tra questi modelli, la relazione di controfattualità individua quelli che faranno parte del contesto fattuale c_F come tutti i modelli in cui non sono vere né B né O. In questo caso ne risulterà un solo modello:

- $m_{F_1} = \neg B \wedge \neg O$

All'interno di M_{CF} andranno ora individuati i modelli locali che fanno parte del contesto controfattuale c_{CF} e saranno precisamente i modelli in cui è vera B, e quindi,

- $m_{CF1} = B \wedge O$

Poiché in questo unico modello O è vera, il controfattuale risulta vero.

Esempio 4.4.2 *Sappiamo che un guasto al circuito elettrico ha il potere di impedire l'olocausto, ma non facciamo alcuna ipotesi sul guasto. Se Nixon avesse premuto il bottone, ci sarebbe stato l'olocausto?*

In questo caso l'agente sa che la pressione di un bottone, unita al buon funzionamento del circuito di trasmissione (cioè l'assenza di guasti) porta all'olocausto, sa che non ci sono stati né pressione del bottone né olocausto,

si chiede se, alla pressione del bottone da parte di Nixon sarebbe seguito l'olocausto, senza fare alcuna ipotesi aggiuntiva su eventuali guasti.

La teoria contiene l'assioma $B \wedge \neg G \rightarrow O$.

I suoi possibili modelli sono dunque i seguenti:

- $m_{a_1} = B \wedge G \wedge O$
- $m_{a_2} = B \wedge G \wedge \neg O$
- $m_{a_3} = B \wedge \neg G \wedge O$
- $m_{a_4} = \neg B \wedge G \wedge O$
- $m_{a_5} = \neg B \wedge G \wedge \neg O$
- $m_{a_6} = \neg B \wedge \neg G \wedge O$
- $m_{a_7} = \neg B \wedge \neg G \wedge \neg O$

Di questi, i modelli che fanno parte del contesto fattuale sono solo quelli dove B e O sono entrambe false (nessuna informazione su G). E dunque:

- $m_{F_1} = \neg B \wedge G \wedge \neg O$
- $m_{F_2} = \neg B \wedge \neg G \wedge \neg O$

E quelli che fanno parte del contesto controfattuale solo quelli in cui B è vera, ma in cui nessuna ipotesi viene fatta su G:

- $m_{CF_1} = B \wedge G \wedge O$
- $m_{CF_2} = B \wedge G \wedge \neg O$
- $m_{CF_3} = B \wedge \neg G \wedge O$

Poiché in due dei tre modelli locali O è vera, ma nel terzo modello è falsa, il valore di verità del controfattuale resterà indeterminato in questo caso, sarà tuttavia possibile affermare la possibilità controfattuale, ossia "Se Nixon avesse premuto il bottone, avrebbe potuto esserci l'olocausto".

Esempio 4.4.3 *Sappiamo che un guasto impedisce l'olocausto. Se Nixon avesse premuto il bottone e si fosse verificato un guasto nel circuito, ci sarebbe stato l'olocausto?*

Questo caso è molto simile al precedente: l'agente sa che un guasto nel circuito elettrico può impedire l'olocausto, non sa se nella realtà questo guasto si sia verificato o meno, ma sa che Nixon non ha premuto il bottone e non c'è stato l'olocausto. Si chiede cosa sarebbe successo se Nixon avesse premuto il bottone e non si fosse verificato nessun guasto. La domanda è: "Ci sarebbe stato l'olocausto?"

La teoria contiene (come nell'esempio precedente) l'assioma $B \wedge \neg G \rightarrow O$.

I suoi possibili modelli sono dunque ancora:

- $m_{a_1} = B \wedge G \wedge O$
- $m_{a_2} = B \wedge G \wedge \neg O$
- $m_{a_3} = B \wedge \neg G \wedge O$
- $m_{a_4} = \neg B \wedge G \wedge O$
- $m_{a_5} = \neg B \wedge G \wedge \neg O$
- $m_{a_6} = \neg B \wedge \neg G \wedge O$
- $m_{a_7} = \neg B \wedge \neg G \wedge \neg O$

Ugualmente, i modelli che fanno parte del contesto fattuale sono solo quelli dove B e O sono entrambe false (nessuna informazione su G). E dunque:

- $m_{F_1} = \neg B \wedge G \wedge \neg O$
- $m_{F_2} = \neg B \wedge \neg G \wedge \neg O$

Questa volta, però, le assunzioni del contesto controfattuale sono cambiate: si prendono quindi i modelli in cui B è vera, ma G è falsa:

- $m_{CF_1} = B \wedge \neg G \wedge O$

Poiché l'unico modello del contesto controfattuale catturato da questa specifica relazione di controfattualità verifica O, allora in questo caso il controfattuale risulta vero.

Esempio 4.4.4 *Sappiamo che, qualunque cosa decida di fare o non fare Nixon, un guasto impedisce l'olocausto. Se Nixon avesse premuto il bottone e si fosse verificato un guasto nel circuito, ci sarebbe stato l'olocausto?*

L'unico assioma della teoria stavolta non prende in considerazione la pressione del bottone, ma solo il fatto che un guasto impedisce in ogni caso l'olocausto, l'assioma è: $G \rightarrow \neg O$. I modelli soddisfacibili di M_a sono quindi:

- $m_{a_1} = B \wedge G \wedge \neg O$
- $m_{a_2} = B \wedge \neg G \wedge O$
- $m_{a_3} = B \wedge \neg G \wedge \neg O$
- $m_{a_4} = \neg B \wedge G \wedge \neg O$
- $m_{a_5} = \neg B \wedge \neg G \wedge O$
- $m_{a_6} = \neg B \wedge \neg G \wedge \neg O$

I modelli locali del contesto fattuale hanno il vincolo di soddisfare $\neg B$ e $\neg O$, sono quindi:

- $m_{F_1} = \neg B \wedge G \wedge \neg O$
- $m_{F_2} = \neg B \wedge \neg G \wedge \neg O$

Mentre i modelli del contesto controfattuale devono soddisfare $B \wedge G$, l'unico modello "superstite" è dunque:

- $m_{CF_1} = B \wedge G \wedge \neg O$

E, dal momento che in tale modello (quindi in tutti i modelli del contesto controfattuale) è soddisfatta $\neg O$, allora il controfattuale sarà falso e sarà invece vero il semifattuale: "Se anche Nixon avesse premuto il bottone, non ci sarebbe comunque stato l'olocausto".

Infine diamo un suggerimento ancora molto preliminare di come il nostro modello potrebbe affrontare il problema dell'iterazione.

Esempio 4.4.5 *Se fosse vero che, se Nixon avesse premuto il bottone ci sarebbe stato l'olocausto, allora sarebbe vero che non ci sono stati guasti?*

Possiamo partire da una teoria che contenga solamente l'assioma $G \rightarrow \neg O$, cioè che quando ci sono guasti non si verifica l'olocausto, e costruire l'insieme dei modelli soddisfacibili:

- $m_{a_1} = B \wedge G \wedge \neg O$
- $m_{a_2} = B \wedge \neg G \wedge O$
- $m_{a_3} = B \wedge \neg G \wedge \neg O$
- $m_{a_4} = \neg B \wedge G \wedge \neg O$
- $m_{a_5} = \neg B \wedge \neg G \wedge O$
- $m_{a_6} = \neg B \wedge \neg G \wedge \neg O$

Di questi, i modelli fattuali sono quelli in cui valgono $\neg B$ e $\neg O$, ossia i modelli nei quali Nixon non preme il bottone e non si ha l'olocausto.

Tali modelli sono:

- $m_{F_1} = \neg B \wedge G \wedge \neg O$
- $m_{F_2} = \neg B \wedge \neg G \wedge \neg O$

Il primo passo è quello di costruire un contesto controfattuale in cui sia vero l'antecedente del controfattuale globale, ossia in cui sia vero il controfattuale "Se Nixon avesse premuto il bottone, ci sarebbe stato l'olocausto nucleare" e nel quale si verifichi se sia vero o meno che in quel caso ci sarebbe stato un guasto. Stipulare questo enunciato controfattuale come vero nel contesto controfattuale ha una serie di conseguenze. La prima è quella che, anche in questo primo contesto controfattuale, sia falso che Nixon abbia premuto il bottone e che ci sia stato l'olocausto nucleare. Quindi, almeno in partenza, i modelli locali di questo contesto sono gli stessi del contesto fattuale:

- $m_{CF_1} = \neg B \wedge G \wedge \neg O$

- $m_{CF_2} = \neg B \wedge \neg G \wedge \neg O$

Ma a questo punto, per poter verificare del guasto, è necessario costruire un secondo contesto controfattuale, chiamiamolo $c_{CF'}$, nel quale postulare che Nixon abbia premuto il bottone. I modelli di $c_{CF'}$ sono dunque:

- $m_{CF'_1} = B \wedge G \wedge \neg O$
- $m_{CF'_2} = B \wedge \neg G \wedge O$
- $m_{CF'_3} = B \wedge \neg G \wedge \neg O$

Tuttavia, poiché la nostra relazione di controfattualità ci impone che il controfattuale "Se Nixon avesse premuto il bottone, ci sarebbe stato l'olocausto nucleare" sia vero nel contesto c_{CF}, dobbiamo togliere da $c_{CF'}$ i modelli che non siano compatibili con la verità di tale enunciato in c_{CF}. Questa operazione individua il solo modello:

- $m_{CF'_2} = B \wedge \neg G \wedge O$

nel quale G è falsa e il controfattuale "globale" è quindi verificato.

4.5 In che modo la semantica a modelli locali è adatta a rappresentare il ragionamento controfattuale

La prima riflessione che ci ha condotto nella direzione della semantica a modelli locali è relativa al fatto che, quando si formulano ipotesi controfattuali, in realtà, data anche la complessità del tipo di ragionamento, si tende a isolare solo alcuni aspetti del problema che si sta esaminando e a inquadrare tale problema in una determinata prospettiva.

A questo punto crediamo di poter affermare che le teorie basate sui mondi possibili (e quelle a esse equivalenti) portino con sé una conseguenza controintuitiva, cioè che un agente, ragionando su un'ipotesi controfattuale, debba considerare tutta la conoscenza che ha a disposizione relativamente al mondo

reale e, prima di poter valutare un enunciato controfattuale, debba considerare il valore di verità di tutti gli enunciati relativi al mondo (o a quella porzione di mondo) formulabili nel suo linguaggio.

La nostra ipotesi è invece che il soggetto ragionante costruisca appositamente uno spazio di ragionamento parziale (quello racchiuso dal contesto controfattuale appunto), il quale è costruito a partire da un contesto di lavoro (quello contenente solo quelle credenze dell'agente, relative a quella limitata porzione della realtà sulla quale si sta ragionando, che questi ha esplicitamente presenti e che impiega per formare la teoria particolare che utilizza per quel ragionamento).

Va precisato che formalmente un contesto è un insieme di modelli (quelli che vengono definiti modelli locali), ognuno dei quali può essere visto come un modello classico (*à la* Tarski). Proprio questa caratteristica è all'origine della parzialità tipica del contesto, poiché in esso le formule del linguaggio non devono necessariamente acquisire un valore di verità definito, dal momento che esse potrebbero essere vere in alcuni modelli del contesto e false in altri.

Determinante nella scelta di assumere la prospettiva della semantica a modelli locali nella nostra analisi dei controfattuali è stata la constatazione del fatto che tutti i suoi concetti basilari sono fortemente adatti a descrivere il fenomeno del ragionamento controfattuale.

In primo luogo, ci pare appropriato affermare che, quando ragioniamo controfattualmente, in qualche modo creiamo una nuova teoria, con dei presupposti diversi rispetto a quelli che utilizziamo quando ragioniamo su porzioni di realtà; questi presupposti possono essere, banalmente, una diversa assegnazione di valori di verità ad alcune formule del linguaggio (quando poniamo un'ipotesi controfattuale, il più delle volte quello che facciamo consiste semplicemente nell'affermare la verità di un enunciato falso in quella che chiamiamo la "situazione reale" o viceversa); oppure, più raramente, un diverso insieme di assiomi (è possibile violare ipoteticamente dei principi che consideriamo inviolabili allorché ragioniamo sulla realtà).

Inoltre, appare chiaro come i principi di località e di compatibilità siano validi nel caso del ragionamento controfattuale. Per cominciare, parte del ragionamento controfattuale si svolge evidentemente su un terreno diverso rispetto a quello dei fatti reali, in uno spazio dotato di regole proprie (que-

sta è la parte intracontestuale del ragionamento, che avviene all'interno del contesto controfattuale). Tuttavia, l'ipotesi controfattuale non è qualcosa che nasca improvvisamente e sia avulsa da tutto, ha dei vincoli ben precisi che discendono dal fatto che l'ipotesi generalmente è basata su un'osservazione reale e di conseguenza il contesto controfattuale mantiene delle relazioni con il contesto dal quale viene originato. Queste relazioni costituiscono la dimensione intercontestuale.

Infine, la relazione che lega due contesti in maniera tale da renderli uno controfattuale rispetto all'altro incorpora molto spesso uno spostamento di focus relativo alla regione *parziale* che si sta analizzando (se facciamo rotolare una sfera su un piano, prima o poi questa si ferma, ma se fossimo in assenza di attrito, essa continuerebbe a muoversi di moto uniforme – ossia, la legge vale solo se la porzione in esame è limitata a una zona sottoposta alle leggi d'attrito), o un cambiamento del livello di *approssimazione* dell'indagine (non è vero che in un dato posto la temperatura è sempre minore di 35 gradi, ma sarebbe vero se effettuassimo la misurazione tutti i giorni a mezzanotte – ovvero, le cose cambiano con l'aggiunta di un parametro di valutazione), o uno slittamento di *prospettiva* (io posso affermare che non è vero che in questo momento la sedia si trova dietro al tavolo, ma lo sarebbe se fossi tu a pronunciare questa frase – poiché la sedia e il tavolo sono posizionati uno dietro l'altra tra me e te), oppure una combinazione di questi processi [10].

Questo approccio presenta una serie di notevoli vantaggi: innanzitutto, un singolo enunciato controfattuale può dare luogo alla formazione di una molteplicità di coppie di contesti fattuale/controfattuale, all'interno dei quali vengono condotti dei ragionamenti che portano a conclusioni molte volte diverse. La diretta conseguenza di ciò è il fatto che per non tutti gli enunciati controfattuali sia possibile determinare univocamente un valore di verità, ma la maggior parte di essi sarà vero o falso a seconda del contesto controfattuale (compatibile con il contesto di partenza) all'interno del quale viene condotto il ragionamento.

Questo rispecchia il fatto che agenti diversi possono assentire o dissentire rispetto allo stesso enunciato e il fatto che anche lo stesso agente può negare l'assenso a un enunciato al quale in un primo momento l'aveva dato, in seguito a piccole precisazioni fornite da altri agenti; pensiamo a un agente che

affermi: "Se avessi saputo che c'era la coda in autostrada, ieri avrei preso il treno", al quale un secondo agente replichi: "Ieri c'era lo sciopero dei treni", facendo concludere al primo: "Allora anche se avessi saputo della coda, non avrei preso comunque il treno". Questo esempio mostra il carattere non monotono del ragionamento controfattuale. I differenti criteri di valutazione dell'enunciato controfattuale sono schematizzati nella figura 4.7, che rappresenta il modello classico di Lewis che effettua la valutazione su una porzione di una sfera di mondi e nella figura 4.7 che, parallelamente, mostra come i vari contesti fattuali e, conseguentemente, controfattuali, siano costruiti a partire da una porzione del mondo; si noti inoltre che ciascun contesto può avere un linguaggio suo proprio che normalmente *non* si identifica con tutto il linguaggio che un agente utilizzerebbe per parlare del mondo, ma solo con una parte di esso.

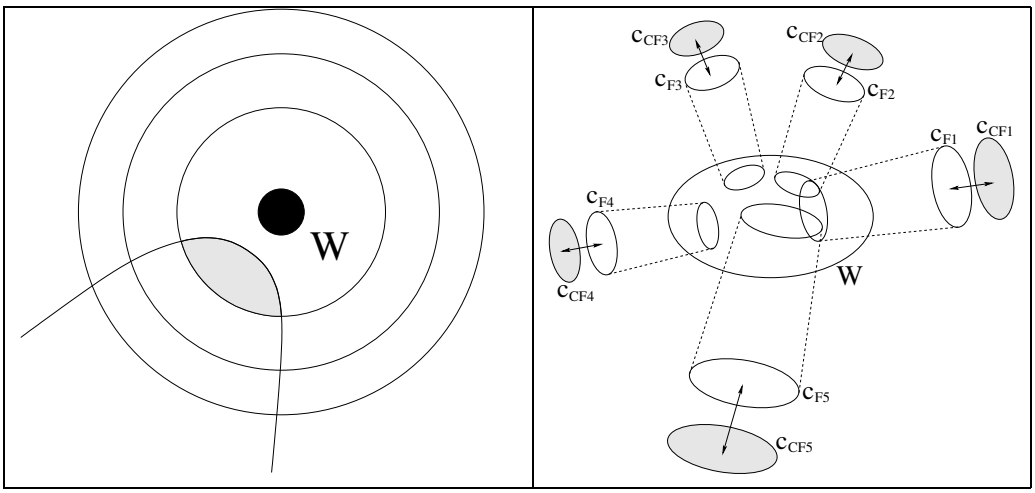

Figura 4.7: Valutazione di un controfattuale *à la* Lewis vs. SML

Nel caso del ragionamento di senso comune, i contesti possono essere molto articolati (nel senso di composti da molti modelli locali) e tanto più strutturati quanto più approfondita è la conoscenza dell'oggetto di ragionamento da parte dell'agente e quanto più è importante per l'agente dare una risposta corretta al problema. Di conseguenza, a partire da essi sarà possibile costruire una miriade di contesti controfattuali costituiti da un alto numero di modelli locali; in questi casi sarà quindi molto più facile pervenire alla

situazione in cui un enunciato non è né assolutamente vero né assolutamente falso.

Non è da escludere, tuttavia, che in casi più semplici (si pensi a un problema statico a informazione perfetta nella teoria dei giochi, con un albero di alternative definito o a dei quesiti logici elementari, in cui vengono forniti a priori tutti gli elementi sui quali ragionare) – oppure quando si richiede una risposta in tempi molto brevi, questi contesti possono essere anche molto semplici, al limite composti di un solo modello locale.

Risulta a questo punto evidente che, seguendo questa linea di pensiero, l'elemento fondamentale attorno al quale tutte le nozioni in gioco prendono forma è la relazione che sussiste tra due contesti, tale per cui uno venga considerato controfattuale rispetto all'altro.

La nostra analisi sposta il nucleo del fenomeno del controfattuale dall'enunciato e dalla relazione tra proposizioni alla relazione tra teorie.

Il vantaggio maggiore è però forse quello offerto dal fatto che, facendo variare i vincoli sulla relazione di controfattualità, è possibile rappresentare un'ampia gamma di fenomeni.

Questo significa che esisteranno un insieme di vincoli-base che determineranno quando un contesto è controfattuale rispetto a un altro ma, a seconda del valore di verità che assume il conseguente del controfattuale all'interno dei modelli locali individuati dalla relazione di controfattualità, sarà possibile individuare anche altre tipologie di condizionali dell'irrealtà, come le possibilità controfattuali (o *might counterfactuals*) e i semifattuali.

Un altro obiettivo è quello di verificare se, modificando opportunamente certi vincoli, sia possibile rappresentare anche altri tipi di ragionamento supposizionale, come quello sulla possibilità e se, aggiungendone degli altri, sia possibile rappresentare i meccanismi di selezione (delle condizioni rilevanti o dei mondi possibili) che vengono utilizzati in altri approcci, come per esempio quelli basati sul minimo cambiamento (vedi Lewis, Stalnaker), per rappresentare diverse modalità di ragionamento controfattuale.

4.6 Prospettiva cognitiva e "metafisica" a confronto sui controfattuali

Presentiamo di seguito un esempio[7] che mostra il motivo per cui, a nostro avviso, la prospettiva cognitiva che interpreta il ragionamento controfattuale come un'operazione su teorie sia più appropriata della prospettiva classica che stabilisce una relazione tra una supposta realtà e delle alternative meramente possibili che le si oppongono.

Immaginiamo di avere un uomo d'affari, Gino, che ha un appuntamento di lavoro a Londra un certo giorno alle 10 del mattino (ora locale).

Essendo Gino un personaggio un po' pigro e distratto, delega ogni tipo di organizzazione "logistica" alla sua segretaria, la quale prenota il volo.

Gino sale sul'aereo all'aeroporto di Milano alle 8.30 (ora locale) e, durante il volo, si addormenta e non sente dunque l'annuncio della hostess, che avverte i passeggeri di spostare le lancette dell'orologio per via della differenza oraria tra l'Italia e la Gran Bretagna.

Atterra a Londra alle 9 ora locale ma, dal momento che il suo orologio segna le 10 – ed essendo così distratto da non accorgersi che a ogni angolo dell'aeroporto è posizionato un orologio che segna l'ora esatta –, deduce di essere in ritardo per l'appuntamento e formula il seguente pensiero controfattuale: "Se fossi arrivato puntuale, avrei comprato all'aeroporto un regalo per la persona con la quale ho appuntamento" ma, poiché è convinto di essere in ritardo, non compra il regalo.

Questo esempio ha due importanti conseguenze: la prima è che viene a cadere un teorema fondamentale della teoria di Lewis, secondo il quale l'implicazione controfattuale si ridurrebbe a implicazione materiale ogniqualvolta l'antecedente del controfattuale risultasse vero, poiché nel mondo abitato da Gino è vero che l'aereo è atterrato puntuale e tuttavia è falso che Gino compri il regalo; l'enunciato controfattuale risulterebbe dunque falso, tuttavia Gino lo giudica vero e si comporta di conseguenza. Questo mostra che gli agenti

[7] Questo esempio è stato suggerito, anche se non esattamente in questa forma, dal Prof. John Perry, durante una conversazione avvenuta un paio di anni fa a Trento.

valutano ciò che è fattuale e ciò che è controfattuale unicamente sulla base delle loro credenze e in base a questa valutazione agiscono.

Poco importa quindi di ciò che succede nel mondo abitato da Gino, egli agisce in base alle ipotesi che formula sulla scia di ciò che crede della realtà.

La seconda conseguenza è la possibilità, per la SML di rappresentare, in questo caso, due punti di vista: quello dell'agente ragionante, che, sulla scorta del *suo* ragionamento controfattuale, dà l'assenso all'enunciato e quello di un osservatore esterno che, sulla base dell'osservazione del comportamento di Gino può negare l'assenso poiché, rispetto a Gino, il suo contesto di ragionamento possiede in più l'informazione che l'orologio di Gino non indica l'ora locale esatta.

Un approccio epistemico pare dunque più adeguato di uno "metafisico" per spiegare le conseguenze pratiche del ragionamento controfattuale.

Parte II

Il ragionamento controfattuale su azioni razionali

Capitolo 5

Nozioni fondamentali per una teoria del ragionamento pratico

Il racconto non sarà piacevole quanto il fatto: ma non sarebbe giusto che, mentre voi vi siete limitato a ragionar bene o male su quest'affare, ve ne derivasse un piacere pari a quello che spetta a me, che ci ho messo tempo e fatica.

[Choderlos de Laclos, *I legami pericolosi*, Lettera LXXXV]

In questo capitolo ci occuperemo di delineare le nozioni fondamentali sulle quali si impernia la seconda parte della tesi; ciò ha un duplice scopo: quello di tracciarne il percorso e al tempo stesso i limiti.

Infatti, poiché la finalità di tale parte è quella di mostrare come due diverse forme di razionalità possano entrambe assumere una valenza controfattuale e al tempo stesso di suggerire che la controfattualità può avere un ruolo nelle azioni razionali, è necessario innanzitutto spiegare che cosa si intende in questa sede per agente razionale, razionalità strumentale, razionalità *ex-post*.

5.1 Chi o che cos'è un agente razionale

John Pollock, in [132] afferma:

> Practical reasoning aims at deciding what actions to perform in light of the goals a rational agent possesses.[1]
>
> [New Foundations for Practical Reasoning, p.113]

Sembra quindi necessario fornire delle definizioni di partenza dei concetti di base impiegati nel ragionamento pratico.

Cominciamo con le definizioni che ci vengono fornite dal dizionario della lingua italiana:

> **Agente:** Chi, ciò che, agisce.
>
> **Agire:** Compiere un'azione.
>
> **Azione:** Atto dell'agire, dell'operare; atto del funzionare, del produrre determinati effetti, anche con riferimento a oggetti inanimati, concetti astratti o altro; operato individuale che implica una valutazione morale; manifestazione di un'energia, di una forza fisica o spirituale.
>
> **Razionale:** Che ha la ragione, che è provvisto di ragione; che procede dalla ragione pura o astratta; fondato sulla scienza o su un procedimento scientifico; studiato rigorosamente e realizzato con studio e metodo, così da adempiere nel modo migliore al suo scopo; che si sviluppa per deduzione logica da principi.
>
> *Sinonimi:* positivo, ragionato, reale, realizzabile, vero, coerente, congruo, giusto, logico, sensato, concreto, corporeo, fisico, palpabile, pragmatico, pratico, realistico, visibile, convincente, giudizioso, scientifico, valido. *Contrari:* irrazionale, cervellotico, chimerico, fantomatico, favoloso, fiabesco, illusorio, immaginario, ipotetico, irreale, straordinario, utopistico, fantastico, irragionevole, contraddittorio, delirante, illogico, incoerente, incongruente, ingiusto, empirico, pazzesco.

---

[1] Il ragionamento pratico ha lo scopo di decidere quali azioni eseguire alla luce degli obiettivi che un agente razionale possiede. [*traduzione mia*]

> **Ragione:** La facoltà di pensare stabilendo rapporti e legami tra i concetti, di giudicare bene discernendo il vero dal falso, il giusto dall'ingiusto; discorso, conversazione, ragionamento; argomentazione, prova, dimostrazione.
>
> **Obiettivo:** Scopo o fine che si vuole raggiungere.
>
> [*Il Nuovo Zingarelli*, Zanichelli]

Quello che si può dedurre da queste definizioni è che, secondo il senso comune, un agente è un individuo che non si limita a esistere nel mondo, ma esercita la sua influenza sul mondo attraverso l'azione, che è a sua volta espressione di qualcosa che è insito nell'agente e che lo muove a tale azione.

Lo stesso Pollock in [130] dà queste definizioni di agente e agente razionale:

> As I will use the term, an *agent* is any system capable of acting on its environment to render it more congenial to its continued survival.
>
> [...]
>
> As I have described it, a rational agent has beliefs and likes or dislikes. [...] A rational agent must have some internal "doxastic" states that are at least fairly well correlated with some states of its environment, and some *conative disposition* to "like or dislike" its situation[2].
>
> [The Phylogeny of Rationality, p.240]

L'azione razionale appare come un'azione mossa dalla ragione, ossia da quella facoltà che permette di gestire i concetti e i loro legami reciproci sulla base di un criterio, di un metodo. Proprio questo metodo è ciò che rende un'azione

[2]Nel senso in cui userò il termine, un *agente* è un sistema in grado di agire sul suo ambiente per renderlo più congeniale alla sua sopravvivenza continuata. [...] Per come l'ho descritto, un agente razionale ha credenze e gusti positivi o negativi. [...] Un agente razionale deve avere degli stati interni "doxastici" che siano perlomeno abbastanza correlati con alcuni stati del suo ambiente e alcune *disposizioni conative* ad "apprezzare o disprezzare" la sua situazione. [*traduzione mia*].

razionale, cioè le conferisce le qualità di coerenza, validità e pragmaticità che la caratterizzano in quanto razionale[3].

Ma in che cosa consiste questo metodo? Davidson in [44] ha asserito che un'azione è qualcosa che un agente fa che è "intenzionale secondo una qualche descrizione", quindi il metodo in un certo qual modo deve collegare l'intenzione all'azione, costituire la ragione normativa dell'azione e renderla intellegibile in primo luogo agli occhi dell'agente stesso che la compie.

Pur condividendo questa prospettiva di Davidson sull'azione, la caratterizzazione che egli dà della nozione di intenzione è ancora molto debole, essendo il risultato dell'abbinamento di desideri e credenze.

Un filosofo che ha studiato approfonditamente i rapporti tra credenze, desideri, intenzioni, pianificazione e azioni è Michael Bratman[4] ed è precisamente la connotazione che egli dà all'intenzione che vorremmo importare nella nostra prospettiva: l'intenzione è quello stato mentale in cui un agente si trova quando sceglie, tra le varie alternative disponibili, un obiettivo e si impegna a perseguirlo per il futuro. L'intenzione svolge quindi una funzione fondamentale rispetto alla pianificazione.

In [23] si legge:

> A theory of future intentions needs to explain why we ever *bother* to form them. Why do we not just cross our bridges when we come to them? One answer is that we want to avoid the need for deliberation at the time of action. But, more importantly, we form future intentions as parts of larger plans whose role is to aid *co-ordination* of our activities over time. Further, we do not adopt these plans, in all their detail, all at once. Rather, as time goes on we add to and adjust our plans. As elements in

[3]In quanto segue, ogni volta che si parla di azione si intende ogni genere di azione, ivi compresi gli atti linguistici che, secondo quanto indicato da John Austin in [1], sono da considerarsi azioni a pieno titolo.

[4]Dalle idee di Bratman si è sviluppata una corrente di pensiero che ricopre attualmente una certa importanza nel dominio dell'intelligenza artificiale, della quale parleremo più diffusamente nella sezione 8.2.1, che ha ideato i cosiddetti modelli BDI (*Belief Desire Intention*), e di cui due importanti contributi sono indubbiamente [135] e [36].

> these plans, future intentions force the formation of yet further intentions and constrain the formation of other intentions and plans. For example, they force the formation of intentions concerning means, and constrain later plans to be consistent with prior plans.[5]
>
> [Davidson's Theory of Intention, p.223]

Questa citazione è particolarmente rilevante per la nostra trattazione, in quanto mette in luce in maniera molto concisa e sistematica quali sono i motivi principali che giustificano il processo di pianificazione negli agenti razionali. Questi sono:

- spesso non è conveniente trovarsi a dover deliberare nel momento dell'azione perché non se ne ha il tempo; la pianificazione permette di arrivare al momento di agire sapendo già cosa si vuole fare;
- la pianificazione aiuta a coordinare le azioni che si compiono perseguendo obiettivi diversi contemporaneamente;
- l'interazione con un ambiente in costante evoluzione rende desiderabile se non necessaria la capacità di riaggiustare progressivamente i piani e ciò avviene attraverso la formazione di intenzioni che si riferiscono sia al raggiungimento di obiettivi sia alla ricerca dei mezzi necessari.

Quando agisce, l'agente si impegna in un'attività che è diretta a un obiettivo, che l'agente stesso ha adottato sulla base di considerazioni sulle opzioni che

[5]Una teoria delle intenzioni future deve spiegare perché mai ci *preoccupiamo* di formarcele. Perché semplicemente non attraversiamo i nostri ponti quando giungiamo a essi? Una risposta è che vogliamo evitare di aver bisogno di deliberare al momento dell'azione. Ma, più importante, formiamo intenzioni future come parte di piani più ampi il cui ruolo è di aiutare il *coordinamento* delle nostre attività nel tempo. Inoltre, non adottiamo questi piani, in tutto il loro dettaglio, tutti in una volta. Piuttosto, col passare del tempo, facciamo aggiunte e aggiustiamo i nostri piani. Come elementi di questi piani, le intenzioni future inducono la formazione di ulteriori intenzioni future e vincolano la formazione di altre intenzioni e altri piani. Per esempio, inducono la formazione di intenzioni relative ai mezzi e vincolano i piani futuri alla consistenza con i piani precedenti. [*traduzione mia*]

gli sono disponibili. Inoltre, l'agente è consapevole sia del fatto di essere impegnato nell'azione, sia del fatto che l'azione è per lui finalizzata a un certo obiettivo.

Il metodo sottostante e caratterizzante un'azione razionale è dunque una procedura normativa che funge da collegamento tra l'azione stessa e l'obiettivo che con essa l'agente intende realizzare, o alla cui realizzazione l'azione dovrebbe concorrere.

La nozione di agente e azione che verranno utilizzate nel prosieguo sono la risultante di quanto detto finora e sono molto vicine alle definizioni fornite da Castelfranchi in [32]:

> At a very basic level, an agent is any entity *able to act*, i.e., to produce some causal effect and some change in its environment.
>
> [...]
>
> In other terms, the agent's behavior is aimed at producing some result: thus we are talking of a *goal-oriented* action and of a goal-oriented agent. [...] Among goal-oriented systems I will consider in particular *goal-directed* systems. In these systems not only action is based on perception, but the latter is also the perception of the action's effects and results, and the agent regulates and controls its actions on such a basis. *The agent is endowed with goals*, i.e., internal anticipatory and regulatory representations of action results.[6].
>
> [Modelling social action for AI agents, p.160]

[6]A un livello basilare, un agente è un'entità *in grado di agire*, cioè di produrre effetti causali e cambiamenti nel suo ambiente. [...] In altre parole, il comportamento dell'agente è finalizzato a produrre qualche risultato: quindi stiamo parlando di un'azione *orientata a un obiettivo* e di un agente orientato a un obiettivo. [...] Tra i sistemi orientati a un obiettivo considererò in particolare i sistemi *diretti da un obiettivo*. In questi sistemi non solo un'azione è basata sulla percezione, ma quest'ultima è anche la percezione degli effetti e dei risultati dell'azione e l'agente regola e controlla le sue azioni su tale base. *L'agente è dotato di obiettivi*, cioè, di rappresentazioni interne anticipatorie e regolative dei risultati dell'azione. [*traduzione mia*]

Sono due quindi le proprietà che emergono da queste definizioni come differenzianti un agente razionale da agenti che razionali non sono. In primo luogo, se un agente in generale è per definizione orientato verso un obiettivo, un agente razionale è un agente che mette in atto una procedura metodologica rigorosa in vista del raggiungimento di quell'obiettivo. Questa metodologia è diretta all'obiettivo in due sensi: quello più generale definito dal fatto di terminare la sua azione al raggiungimento di quell'obiettivo, ma anche quello più specifico di costruirsi una rappresentazione mentale dell'obiettivo, dell'azione e delle conseguenze risultanti da questa.

Quella che abbiamo definito, da un punto di vista prescrittivo, come procedura metodologica, in una prospettiva descrittiva può essere interpretata come il processo di pianificazione diretto a un obiettivo. La prima proprietà in questione sarà dunque la capacità di formare piani.

Così Pollock [130]:

> A rational agent directs its activity on the basis of its beliefs about the expected values of combinations of features, trying always to better its situation, i.e., render it more to its likings. It does this by choosing goals whose achievement will have that effect, and then selecting and executing courses of action that aim at the achievement of those goals.
>
> [...]
>
> An agent tries to achieve goals by designing and executing courses of action aimed at realizing them. Designing such a course of action is *planning*[7].
>
> [Phylogeny of Rationality, p.276]

[7]Un agente razionale dirige la sua attività sulla base delle sue credenze sui valori attesi di combinazioni di caratteristiche, cercando sempre di migliorare la propria situazione, cioè, di renderla più vicina alle sue preferenze. Fa questo scegliendo obiettivi il cui raggiungimento avrà quell'effetto e in seguito selezionando ed eseguendo corsi di azione che abbiano come scopo il raggiungimento di quegli obiettivi. [...] Un agente cerca di raggiungere obiettivi teorizzando ed eseguendo corsi di azione finalizzati alla loro realizzazione. Teorizzare tale corso d'azione è *pianificazione*. [*traduzione mia*]

La seconda proprietà è la capacità di scegliere l'obiettivo da perseguire sulla base delle proprie preferenze, di modo che lo stato nel quale l'agente si troverebbe una volta conseguito l'obiettivo sia per lui preferibile per qualche rispetto allo stato in cui si trova al momento in cui sviluppa l'intenzione di perseguire proprio quell'obiettivo.

Un agente razionale deve quindi essere in grado di percepire gli aspetti rilevanti di una situazione, valutare la loro desiderabilità e determinare dei piani per trasformare la situazione corrente in una più desiderabile.

Vedremo nel seguito come la procedura metodologica che abbiamo sostenuto rendere un agente razionale non è una sola, ma ne esistono almeno due, fondate su due opposti modi di procedere. La descrizione di queste diverse forme di razionalità sarà oggetto della sezione 5.2.

5.2 Diversi tipi di razionalità

In questa sezione verranno descritte due diverse forme di razionalità: nel paragrafo 5.2.1 ci si occuperà della forma classicamente utilizzata in filosofia, economia e scienze cognitive, caratteristica della cosiddetta *Teoria delle decisioni* e che, mantenendo fissi gli obiettivi che l'agente si è prefissato, permette di rivedere i mezzi necessari a raggiungerli; in 5.2.2, invece, si analizzerà quella che James March, ribaltando la lettura avanzata dai teorici della razionalità classica, che la interpretavano come una manifestazione di irrazionalità, ha definito come *razionalizzazione ex-post*, che procede "a ritroso" dai fini ai mezzi, cioè mantiene fissi i mezzi da utilizzare alterando l'obiettivo da raggiungere, attraverso una modifica delle preferenze dell'agente che si trova di fronte alla decisione da intraprendere.

5.2.1 La razionalità strumentale

L'analisi delle forme di razionalità prende le mosse dunque dalla forma che è stata analizzata diffusamente dai modelli tradizionali di teorie della razionalità in filosofia (*Decision Theory*, teoria delle decisioni), in economia (*Rational Choice Model*, teoria della scelta razionale) e nelle scienze cognitive, ossia la razionalità strumentale.

Secondo quanto indicato da Bouvier [22], le prime versioni di teorie microeconomiche partivano da assunzioni molto rigide e piuttosto irreali e facevano riferimento al cosiddetto *homo oeconomicus*, concetto coniato da John Stuart Mill [117] e definito come:

> un individuo che, posto di fronte a diverse alternative, ha un insieme completo di preferenze, può assumere informazioni perfette senza costo, ha autonomia decisionale, e tende a massimizzare-ottimizzare il proprio interesse o la propria utilità. In questo modello di riferimento in ambito economico, la razionalità coincide con un insieme di assiomi che assicurano una coerenza logica alla scelta individuale.
>
> [*Azioni, Razionalità e decisioni*]

L'*homo oeconomicus* è un individuo che agisce esclusivamente sulla base di considerazioni tese a massimizzare il proprio benessere, senza subire l'influenza di situazioni emotive. I valori di base che guidano le sue azioni sono rivolti al massimo soddisfacimento della propria utilità e in questo senso il soggetto viene considerato come un decisore guidato esclusivamente dal proprio interesse. In sostanza, è un essere egoista, in quanto pensa solo ed esclusivamente in termini di massimizzazione delle proprie preferenze. Il decisore è definito quindi come un soggetto razionale nel senso che adotta sempre la scelta che gli permette di massimizzare la propria utilità sulla base di un insieme di preferenze dato. Shotter in [140] afferma che "la caratteristica peculiare del decisore è l'assoluto rispetto dei principi della razionalità", dove, evidentemente, il riferimento è diretto a questa idea di razionalità "perfetta".

Razionalità sostanziale

Le assunzioni su cui erano basati questi primi modelli, che Herbert Simon ha sottoposto a critica per poi proporre dei modelli focalizzati su una caratterizzazione più realistica e meno ideale del decisore, possono essere riassunte nei seguenti punti:

- il decisore viene visto come *homo oeconomicus*;

- il decisore è guidato nella scelta esclusivamente da una funzione di utilità;
- il decisore è dotato di un insieme di preferenze completo;
- sono disponibili tutte le informazioni necessarie e a costo nullo.

Per caratterizzare questo tipo di concezione, con tutto il suo portato di assunzioni, Simon ha coniato il termine *razionalità sostanziale*; le assunzioni sopraelencate sono necessarie affinché quei modelli abbiano validità formale e logica (come mostrato in [143], [144] e [146]).

Ecco come viene definita la razionalità sostanziale in [144]:

> Il comportamento è razionale in senso sostanziale quando è appropriato al raggiungimento di dati obiettivi all'interno di limiti imposti da date condizioni e vincoli. Da notare che, per definizione, la razionalità del comportamento dipende dall'agente per un solo aspetto: i suoi obiettivi. Dati questi obiettivi, il comportamento razionale è completamente determinato dalle caratteristiche dell'ambiente in cui ha luogo.
>
> [*Causalità, razionalità, organizzazione*, p.293]

In sostanza, Simon afferma che, all'interno della teoria della scelta razionale, il decisore sceglie l'azione che meglio soddisfa il suo obiettivo secondo le preferenze date. Questo comportamento non può che essere quello preferibile in assoluto tra tutti i corsi d'azione possibili. Il processo di formulazione dell'obiettivo non è direttamente preso in considerazione dalla teoria classica, in quanto il problema viene aggirato attraverso l'idea della massimizzazione dell'utilità, secondo la quale il decisore agisce solo ed esclusivamente in relazione al miglior soddisfacimento della propria utilità.

La conseguenza diretta di questa concezione del decisore razionale è la caratterizzazione delle preferenze, affinché esse siano coerenti, secondo queste linee sostanziali:

1. **completezza**: le preferenze sono complete, nel senso che il decisore è sempre in grado di scegliere quale alternativa ha più valore tra due o

più oggetto della scelta; il decisore quindi non ha vuoti di preferenze, nel senso che sa ordinare tutte le opzioni che gli vengono presentate;

2. **riflessività**: con questa proprietà si afferma che una determinata alternativa, posta a confronto con se stessa, "è buona almeno quanto" se stessa. Tale relazione è necessaria per evitare che il decisore cada in contraddizione;

3. **transitività**: questa proprietà è riassumibile attraverso questo semplice principio: se un'alternativa A è ritenuta migliore dell'alternativa B, e l'alternativa B è ritenuta migliore dell'alternativa C, allora A è ritenuta migliore di C;

4. **invarianza**: l'ordinamento delle alternative non viene influenzato dal modo in cui le alternative vengono presentate;

5. **dominanza**: se esistono diverse dimensioni che devono essere considerate nella scelta, tra due alternative simili verrà scelta l'alternativa che ha almeno una dimensione dominante.

Come sottolineato da Simon, il difetto più evidente di questi modelli della teoria economica classica è che non pone alcun limite alla razionalità dei soggetti, le uniche restrizioni che teorizza sono di natura "strutturale" o ambientale e quindi ininfluenti ai fini della decisione in quanto tale, poiché non soggette alla volontà o all'intervento del decisore, che quindi dovrebbe avere una conoscenza perfetta dei vincoli ambientali e un'elevatissima capacità di calcolo.

Al fine di enunciare una teoria della razionalità più conforme alla realtà, il primo passo che Simon compie è quello di criticare le assunzioni della teoria classica della razionalità che giudica inesatte, poiché danno luogo a un modello irreale che definisce *olimpico* e di mostrare quali siano invece i limiti da mettere in luce in una teoria della razionalità più rispondente ai fatti.

Razionalità procedurale

Quello che ne risulta è quella che Simon definisce *razionalità procedurale*, ossia una razionalità che si definisce sulla base delle procedure di risoluzione adottate piuttosto che sulle soluzioni finali ottenute.

Le assunzioni implicite nella teoria della razionalità sostanziale sono:

- capacità computazionale perfetta;
- conoscenza perfetta;
- indipendenza degli attori.

La capacità computazionale perfetta a sua volta si compone di due sottoparti:

- la capacità di processare *tutti* i dati necessari per poter prendere una decisione, cioè l'assenza di limiti soggettivi di elaborazione dei dati e quindi la capacità di trovare sempre l'azione che otterrà il risultato migliore;
- l'assenza di problemi di sequenzialità, ovvero la capacità, quando si trova a dover affrontare più problemi contemporaneamente, di risolverli senza dover adottare una scaletta di priorità; in altri termini, il soggetto non ha problemi di attenzione.

Per quanto riguarda il primo di questi due punti, Simon mostra, attraverso esempi tratti dal gioco degli scacchi, come, essendo gli scacchi un gioco a dominio chiuso, teoricamente un agente *olimpico* dovrebbe essere in grado di computare tutte le 10^{120} mosse e quindi per lui dovrebbe essere indifferente scegliere la strategia da adottare tutta insieme all'inizio della partita, oppure decidere volta per volta cosa rispondere alle mosse dell'avversario.

Nella realtà dei fatti, anche per il calcolatore più potente risulta alquanto difficile e laborioso considerare tutte queste alternative ed essere quindi, oltreché efficace, anche efficiente, dal momento che i tempi di una tale computazione sarebbero piuttosto lunghi.

Sembrerebbe che il modo di procedere dei campioni di scacchi sia invece alquanto diverso e sia basato sulla considerazione di non più di un centinaio di alternative nella scelta di una mossa o una strategia:

> La realtà in pratica è che è di solito meglio generare solo alcune mosse dell'intero insieme delle mosse possibili, valutando queste piuttosto approfonditamente, piuttosto che generarle tutte, valutandole superficialmente.
>
> [*Causalità, razionalità, organizzazione*, p.268]

Un'altra limitazione caratteristica dell'uomo (ma in parte anche della macchina) che non deve essere trascurata è quella imposta dalla ridotta capacità di attenzione, che determina l'obbligo di risolvere i problemi sequenzialmente piuttosto che contemporaneamente; ciò determina un maggior dispendio di tempo e la necessità di sviluppare la capacità di assegnare un ordine di priorità ai problemi che si presentano.

Anche all'interno dell'assunzione di conoscenza perfetta Simon distingue due idee [144] [146]:

- l'informazione completa sull'insieme delle alternative disponibili, alternative che sono definite dalla situazione e che sono conosciute in modo non ambiguo;
- la conoscenza, almeno in modo probabilistico, delle conseguenze che deriverebbero da ogni alternativa possibile.

La conoscenza di tutte le conseguenze e la mancanza di incertezza futura rispecchiano una concezione del mondo ingenuamente deterministica, poiché lo si intende come un'entità oggettiva governata dal principio di causalità. In altre parole, ad ogni possibile azione corrisponde una sola reazione; il mondo è quindi una lunga catena di eventi-causa ed eventi-risultato [110]. Se si considera questo aspetto assieme all'ipotesi presentata sopra, ovvero che le preferenze del decisore sono esogene e non influenzate dal susseguirsi degli eventi, la razionalità strumentale si riduce ad una modalità di scelta sempre predeterminata, in quanto il soggetto sceglie sempre l'alternativa che ha come risultato la conseguenza ottimale relativamente alla propria utilità.

L'ultima idealizzazione della teoria classica che Simon prende in considerazione è la presunta indipendenza del risultato rispetto a eventuali azioni di altri agenti. In base alla teoria, ogni individuo prende le decisioni solo sulla

scorta delle proprie preferenze e senza fare alcuna assunzione rispetto alle azioni degli altri agenti.

Questa ipotesi è alquanto irrealistica[8], poiché, in un contesto di socialità, il comportamento (effettivo e atteso) degli altri influenza in maniera determinante le mosse di ogni singolo agente.

Per esempio, in [159], Thomas Ulen elenca una serie di circostanze nelle quali ci si attende che gli agenti si comportino in maniera egoistica per massimizzare i propri interessi, invece, probabilmente influenzati dal contesto sociale, inaspettatamente cooperano:

> These experimental results present a puzzle for rational choice theory: why do people cooperate when there appears to be a rational basis for *not* cooperating? One possibility is that people start any given interaction from the presumption that it is better to cooperate than not; they continue to cooperate until when evidence shows this to be ill-advised; and then they quit cooperating[9].
>
> [Rational Choice Theory in Law and Economics, p.803]

Le caratteristiche sopraelencate, nelle parole di James March [110], delineano "una razionalità a priori che prescinde dal decisore e dal contesto della decisione".

In sostanza, il punto che autori come Simon e March vogliono evidenziare è che, essendo le teorie della razionalità sostanziale lontane dall'avere

[8]Thomas Ulen elenca in [159] una serie di evidenze empiriche che mostrano come, in determinate situazioni, gli agenti si comportino in maniera "erronea" (o perlomeno inattesa) rispetto ai criteri della razionalità classica e gli esempi più interessanti riguardano appunto situazioni di socialità nelle quali gli agenti si comportano in maniera inaspettatamente altruistica o, all'opposto, danneggiano in parte se stessi mossi da pulsioni di invidia o vendetta.

[9]Questi risultati sperimentali presentano un rompicapo per la teoria della scelta razionale: perché la gente coopera quando sembra esserci una base razionale per *non* cooperare? Una possibilità è che la gente comincia ogni data interazione dall'assunzione che è meglio cooperare piuttosto che non cooperare; continuano a cooperare finché l'evidenza non mostra che questo sia sconsigliabile; allora smettono di cooperare. [*traduzione mia*]

un effettivo riscontro nelle situazioni reali, è preferibile elaborare una nuova concezione di razionalità che tenga in considerazione le limitazioni alle quali sono sottoposti i decisori e non si limiti a bollare come irrazionali comportamenti che, pur non aspirando all'ottimalità, posseggano comunque un certo grado di sensatezza.

Così Simon [147]:

> The point was not that people are consciously and deliberately irrational, although they sometimes are, but that neither their knowledge nor their powers of calculation allow them to achieve the high level of optimal adaptations of means to ends that is posited in economics[10].
>
> [*Economics, bounded rationality and the Cognitive Revolution*]

March [111] aggiunge:

> As decision makers struggle with these limitations, they develop procedures that maintain the basic framework of rational choice but modify it to accomodate the difficulties. Those procedures form the core of theories of limited rationality[11].
>
> [*A Primer on Decision Making: how Decisions Happen*, p.11]

La soluzione additata da entrambi gli autori consiste nello slittamento da una concezione della razionalità come capacità di trovare la soluzione ottimale a una in cui la razionalità è da identificarsi con la capacità di individuare una procedura che consenta di raggiungere dei risultati "buoni" in tempi ragionevoli, questo anche perché la ricerca di soluzioni consuma risorse oltreché tempo e ha quindi un costo.

---

[10] La questione non era se le persone siano coscientemente e deliberatamente irrazionali, sebbene a volte lo siano, ma che né la loro conoscenza né il loro potere di calcolo permettono di raggiungere l'alto livello di adattamenti ottimali dei mezzi ai fini che è supposto dall'economia. [*traduzione mia*]

[11] Scontrandosi con questi limiti, i decisori sviluppano procedure che mantengono la griglia interpretativa fondamentale della scelta razionale, ma la modificano per ridurne le difficoltà. Queste procedure costituiscono il nucleo delle teorie della razionalità limitata. [*tr. it. di Stefano Micelli in*: [112]]

Ecco come March [111] illustra la differenza tra i due modi di intendere la razionalità:

> Rationality is defined as a particular and very familiar class of procedures for making choices. In this procedural meaning of "rational", a rational procedure may or may not lead to good outcomes. The possibility of a link between the rationality of a process (sometimes called "procedural rationality") and the intelligence of its outcomes (sometimes called "substantive rationality") is treated as a result to be demonstrated rather than an axiom[12].
>
> [*A Primer on Decision Making: how Decisions Happen*, p.2]

La razionalità procedurale, a differenza della razionalità sostanziale, non ha come obiettivo quello di ottenere il miglior esito possibile da una decisione, quanto piuttosto quello di individuare una procedura decisionale che permetta all'agente di condurre una ricerca con criterio.

Tale criterio consiste nel fissare le condizioni affinché una scelta venga considerata soddisfacente e permetta, di fronte a un albero decisionale molto ampio e/o profondo, di arrestare la scelta prima di aver vagliato tutte le possibili alternative e di aver percorso tutto l'albero. Questa procedura è ciò che definiamo come *euristica*.

Simon in [145] afferma:

> In tutte queste situazioni per esplorare un piccolo numero di alternative promettenti si usano euristiche selettive ed analisi a lunghezza finita, per terminare la ricerca quando una alternativa soddisfacente è stata trovata.
>
> [Dalla razionalità sostanziale alla razionalità procedurale, p.301]

[12]La razionalità è definita come una particolare classe di procedure per compiere scelte; non è detto tuttavia che tali procedure razionali conducano necessariamente a esiti positivi. La possibilità di un legame fra razionalità di un processo (ciò che è chiamato razionalità procedurale) e la bontà dei suoi esiti (chiamata a volte razionalità sostanziale) è considerata un risultato da dimostrare piuttosto che un assioma. [*tr. it. di Stefano Micelli in*: [112]]

E, poco oltre:

> Lo spostamento dalle teorie della razionalità sostanziale alle teorie della razionalità procedurale richiede un cambiamento nello stile scientifico, con il passaggio da un'enfasi sul ragionamento deduttivo, all'interno di un ristretto insieme di assiomi, ad un'enfasi sulla esplorazione dettagliata dei complessi algoritmi del pensiero.
>
> [Dalla razionalità sostanziale alla razionalità procedurale, pp.315–316]

In accordo con quanto già affermato nella sezione 5.1, in questa trattazione ci si rifa a una concezione procedurale della razionalità e, basandosi ancora una volta sul lavoro di March, si tenta di allargare ancor più il dominio delle azioni e strategie razionali, includendo in esso anche quei procedimenti di ragionamento nei quali l'elemento costante è rappresentato dalle risorse (ovvero i mezzi) già acquisite dal soggetto e la variabile sottoposta a revisione sono le preferenze e, di conseguenza, gli obiettivi. Questo argomento sarà oggetto della prossima sezione.

5.2.2 La razionalità *ex-post*

In questo paragrafo si procederà ad analizzare una seconda forma di razionalità, le cui caratteristiche generali sono state messe in luce da James March e che si cercherà di specificare ulteriormente in relazione ad alcuni aspetti.

Secondo molti autori, una caratteristica distintiva molto importante degli agenti razionali è che questi sanno persistere nell'intenzione di raggiungere un obiettivo anche se, nel periodo che intercorre tra il momento iniziale in cui viene formata l'intenzione e il termine del piano, i loro desideri possono subire delle 'fluttuazioni" che tenderebbero a far abbandonare l'impresa e a dirigerli verso altri obiettivi. Jon Elster, per esempio, in [50] paragona l'agente razionale a Ulisse che si fa legare all'albero della nave per poter resistere alla tentazione rappresentata dal canto delle sirene.

Quello che noi, seguendo March[13], vorremmo sottolineare è che, pur essendo vero che la perseveranza è una componente molto importante per la

[13]March ha un predecessore – in questo senso – in Florian Znaniecki, che nel 1936, in

razionalità, la flessibilità lo è altrettanto in situazioni nelle quali perseverare diventerebbe autolesionistico.

La razionalizzazione *ex-post* secondo March

L'analisi di March parte dall'ampliamento dello spettro delle limitazioni alla razionalità che gli agenti reali si trovano a dover fronteggiare, alcune delle quali, come abbiamo visto in 5.2.1, erano già state evidenziate da Simon.

Le difficoltà di cui parla March, che rendono difficile (se non impossibile) agli agenti razionali di mantenere coerenti le preferenze sono le seguenti[14]:

- **Complessità decisionale**: questa difficoltà è una conseguenza dell'ambiguità dell'ambiente, ovvero la situazione ambientale spesso si presenta alquanto complessa e di difficile comprensione per l'agente, che è costretto ad abbandonare qualsiasi pretesa di servirsi di computazioni esatte e deve rifugiarsi in euristiche.

- **Conflittualità degli obiettivi**: l'agente non è "monolitico" nelle sue intenzioni, ovvero intrattiene più desideri contemporaneamente e può non esistere un unico obiettivo che li soddisfi tutti; al contrario, molto spesso il raggiungimento di un obiettivo impedisce all'agente di conseguirne un altro.

- **Incertezza sulle preferenze future**: l'agente non è incerto solo in relazione alle conseguenze future delle sue azioni presenti, ma anche relativamente alle preferenze che egli stesso intratterrà nel momento in cui gli effetti dell'opzione da lui adottata saranno stati ottenuti.

---

[165], ha definito l'azione sociale come adattamento progressivo di progetti, anticipazioni di possibilità, verso fini spesso predeterminati dai mezzi disponibili, secondo criteri di efficienza (bilancio dello scopo raggiungibile con risorse date) più che di efficacia (mobilitazione delle risorse indispensabili a un fine dato).

[14]March rileva anche come il modello classico di razionalità trascuri altri elementi che influenzano pesantemente le scelte, quali l'intuizione, la tradizione, la fede. In questa sede essi non verranno presi in considerazione poiché, seppur influenti, essi sono periferici rispetto alla razionalità per come è qui intesa, cioè come un processo di ricerca di un metodo rigoroso per raggiungere un obiettivo.

Altri problemi sono stati messi in luce anche da Jon Elster in [50], dove si afferma:

> Un soggetto razionale nell'usuale definizione è semplicemente chiunque abbia delle preferenze coerenti e complete *in qualsiasi istante di tempo dato*. Io credo che la nozione di uomo razionale dovrebbe essere estesa così da includere considerazioni di tipo temporale. Per essere precisi, alcune condizioni di coerenza dovrebbero venir imposte sia alla *scelta delle successioni* del soggetto sia alla sua *successione di scelte*. Un caso in cui la prima condizione non è soddisfatta è quello delle preferenze temporalmente incoerenti, ed uno in cui non è soddisfatta la seconda è quello del cambiamento endogeno delle preferenze.
>
> [tr.it *Ulisse e le sirene*, p.127]

Questa citazione fornisce lo spunto per avviarsi verso la visione che March propone: esiste una diversa forma di razionalità che parte dal presupposto che ogni agente razionale sia costantemente impegnato nell'interpretazione delle proprie azioni, di ciò che accade e della situazione circostante in generale.

Un cambiamento nell'interpretazione di ciò che è accaduto o sta accadendo può determinare una revisione nell'ordinamento delle preferenze, ossia ciò che prima era considerato massimamente preferibile dall'agente potrebbe non esserlo più e viceversa. Ma un riordino delle preferenze spesso porta con sé come diretta conseguenza la maggior desiderabilità di un obiettivo diverso rispetto a quello che si stava perseguendo.

Per tornare alla prospettiva dell'agente che esercita un ragionamento pratico, è da rilevare il fatto che, mentre nei modelli strumentali le preferenze determinano quale sia lo stato da perseguire come obiettivo, in questo diverso modello l'interpretazione di uno stato può determinare nuove o diverse preferenze. Così March [109]:

> Although preferences are used to choose among actions, it is also often true that actions and experience with their consequences affect preferences concurrently.
>
> [...]

> One of the primary ways in which individuals and organizations develop goals is by interpreting the actions they take and one feature of good action is that it leads to the development of new preferences[15].
>
> [How Decisions Happen in Organizations, pp.99-100]

Secondo la lettura di March, spesso i decisori, piuttosto che *raggiungere* un certo stato, *ci si trovano*; a quel punto cercano di dare un'interpretazione sensata dello stato in cui si trovano e, sulla base di essa, rivedono l'ordinamento delle loro preferenze.

In [109] si legge:

> Recent studies of organizations indicate that decisions often stem from a logic of appropriateness rather than a logic of consequentiality and that decision-making may often be better understood in terms of other consequences than their outcomes. To say that decisions "happen" instead of "are made" is to suggest that the organizational processes that result in decisions may be poorly comprehended by a conception of intentional, future-oriented choice[16].
>
> [How decisions happen in organizations, pp.96-97]

Anche John Pollock [130] considera la possibilità di una forma di razionalità che abbia un impatto maggiore sulle preferenze piuttosto che sulle azioni

[15]Sebbene le preferenze siano usate per scegliere tra le azioni, è spesso anche vero che le azioni e l'esperienza con le loro conseguenze influenzano al tempo stesso le preferenze. [...] Uno dei modi primari attraverso i quali gli individui e le organizzazioni sviluppano gli obiettivi è interpretando le azioni che compiono e una caratteristica di una buona azione è che porta a sviluppare nuove preferenze. [*traduzione mia*]

[16]Studi recenti delle organizzazioni indicano che le decisioni spesso derivano da una logica di appropriatezza piuttosto che da una logica di consequenzialità e che il processo decisionale può spesso essere meglio compreso nei termini di conseguenze diverse dai suoi effetti. Dire che le decisioni "accadono" piuttosto che "sono prese" significa suggerire che i processi organizzativi che sfociano nelle decisioni possono essere poco compresi da una concezione intenzionale e orientata al futuro della scelta. [*traduzione mia*]

da intraprendere in vista del raggiungimento di un fine, anche se quest'ultimo la vede come *extrema ratio* piuttosto che come regolare procedura di ragionamento:

> A different kind of case occurs when we cannot change our situation but can change our conative structure so that we like our situation better[17].
>
> [Phylogeny of Rationality, p.584]

Una citazione da [110] può aiutare a riassumere la posizione di March relativamente alla successione azione>interpretazione>revisione delle preferenze:

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

Laddove, in [109] spiega da dove prende le mosse la reinterpretazione della situazione corrente e qual è il criterio per selezionare le nuove preferenze:

> Search is stimulated by a failure to achieve a goal, and it continues until it reveals an alternative that is good enough to satisfy

[17]Un caso diverso è quando non possiamo cambiare la nostra situazione ma possiamo cambiare la nostra struttura conativa in modo che la nostra situazione ci piaccia di più. [*traduzione mia*]

> existing, evoked goals. New alternatives are sought in the neighborhood of old ones. Failure focuses search on the problem of attaining goals that have been violated; success allows search resources to move to other domains.
>
> [...]
>
> This classic-control system does two things to keep performance and goals close. First, it adapts goals to performance: Decision makers learn what they should expect. At the same time, it adapts performance to goals [...][18].
>
> [How Decisions Happen in Organizations, p.98]

Se la nostra interpretazione è corretta, il processo di razionalizzazione *ex-post* avanzato da March può essere così schematizzato:

- Esecuzione di un'azione
- Fallimento nel conseguimento dell'obiettivo prefissato
- Reinterpretazione dello stato ottenuto (al posto dell'obiettivo originario)
- Identificazione dello stato attuale con un obiettivo *desiderabile* (e dunque *desiderato*)
- Riordinamento delle preferenze

Senza voler trascurare l'importanza di questa innovativa trattazione della razionalità fornita da March, vorremmo aggiungere che riteniamo che essa e

[18] La ricerca è stimolata dal fallimento nel raggiungimento di un obiettivo e continua finché non trova un'alternativa che sia abbastanza buona per soddisfare gli obiettivi esistenti ed evocati. Le nuove alternative sono ricercate nelle vicinanze delle vecchie. Il fallimento dirige la ricerca sul problema degli obiettivi da conseguire che sono stati violati; il successo determina la ricerca di risorse per muoversi verso altri domini. [...] Questo sistema classico di controllo fa due cose per mantenere uniti la performance e gli obiettivi. Per prima cosa, adatta gli obiettivi alla performance: i decisori imparano che cosa devono aspettarsi. Al tempo stesso, adatta la performance agli obiettivi [...]. [*traduzione mia*]

la razionalità strumentale non siano alternative esclusive, quanto piuttosto modalità complementari della razionalità.

Questa complementarietà aiuta almeno in parte a completare la trattazione di March nei punti nei quali essa rimane aperta. Questi punti a nostro avviso sono:

- Non si capisce da dove provenga l'azione il cui fallimento genera la nuova interpretazione, poiché, se March vuole evitare il regresso all'infinito, deve ipotizzare che a monte di tutto ci sia un'azione non razionale.

- Non viene mai spiegato il motivo per il quale la reinterpretazione debba necessariamente prendere le mosse da un fallimento e non possa mai essere originata da un successo.

- Non si spiega perché la risposta a un fallimento debba necessariamente essere la revisione delle preferenze e non possa essere banalmente la revisione di un mezzo (un'azione da compiere) all'interno del piano.

- Non si capisce come l'agente procede alla revisione della sua interpretazione (attraverso quale processo di ragionamento).

- Infine, la vicinanza è un criterio un poco vago e debole per stabilire quale nuovo obiettivo decidere di perseguire.

Cercheremo di proporre nella sezione 7.1 delle soluzioni che integrino la teoria di March nelle lacune che abbiamo segnalato.

Capitolo 6

Il ragionamento controfattuale come un tipo di ragionamento sui mezzi

Truth comes out of error more easily than out of confusion.[1]
[Francis Bacon]

Nel capitolo 5 sono stati presentati due tipi di razionalità che gli agenti cognitivi possono mettere in atto quando si trovano a elaborare un piano finalizzato all'azione. Lo scopo di questo capitolo e del seguente è di evidenziare come il ragionamento controfattuale sia uno strumento fondamentale per la valutazione dei piani già portati a termine (mettendoli a confronto con i piani alternativi che erano percorribili al momento della scelta) e come esso sia applicabile indifferentemente ad ambedue le forme di razionalità.

La prima a essere analizzata (in questo capitolo) sarà la teoria della razionalità cosiddetta strumentale, la più ampiamente considerata e impiegata, soprattutto negli studi di intelligenza artificiale e di economia; secondo quanto affermato dai suoi fautori, un agente, una volta che si è prefissato un obiettivo secondo una certa scala di preferenze, considera tutti i mezzi e le capacità che può procurarsi per mettere in atto il piano migliore che possa condurlo al raggiungimento dell'obiettivo.

[1] La verità viene più facilmente dall'errore che dalla confusione. [*traduzione mia*]

In quest'ottica, considerare le alternative a un piano dato significa andare alla ricerca di diversi mezzi e capacità da impiegare allo scopo. È quindi alla riconsiderazione di ciò che genericamente potremmo chiamare "mezzi" che si applica il ragionamento controfattuale in una prospettiva di razionalità classica o strumentale.

Ma, prima di accingerci all'analisi dettagliata del ragionamento controfattuale nell'ottica strumentale, procediamo a mostrare come il ragionamento controfattuale possa tradursi in un processo di revisione dei piani finalizzato all'apprendimento per il futuro.

6.1 Il ragionamento controfattuale come strumento di apprendimento

Riprendendo le definizioni del paragrafo 5.1, ricordiamo che un agente cognitivo [32] è un agente che è diretto, nelle sue azioni, verso un obiettivo. Inoltre, il modo che ha per dimostrare la propria razionalità è quello di costruire, in base ai mezzi a sua disposizione (ossia le risorse che può impiegare e le capacità che possiede) e in base alle sue preferenze (ossia a ciò che giudica desiderabile e al tempo stesso perseguibile), un piano efficace per raggiungere l'obiettivo.

Così, prima di mettersi in azione, un agente osserva l'ambiente circostante e la propria situazione e valuta, tra i vari elementi a sua disposizione, quali scegliere perché entrino a far parte del suo piano per ottenere l'obiettivo e quindi in sostanza sottoscrive un piano a discapito di altri piani alternativi e formula delle previsioni sull'esito del piano prescelto.

Una volta realizzato il piano e messe in pratica le azioni pianificate, l'agente confronta le aspettative con il risultato effettivamente ottenuto.

I due possibili scenari risultanti sono dunque il successo o il fallimento del piano; in caso di successo, la riuscita del piano potrebbe essere dovuta alla lungimiranza con la quale l'agente lo ha ideato, oppure semplicemente alla fortunata confluenza di circostanze favorevoli, oppure l'agente potrebbe aver conseguito l'obiettivo, ma a un costo troppo elevato. D'altro canto, il

fallimento può essere il giusto esito di un piano mal congegnato, oppure la sfortunata conseguenza del verificarsi di eventi sfavorevoli non prevedibili.

Il processo di ragionamento che permette di discernere tra la lungimiranza e la fortuna o lo spreco di risorse da una parte e tra l'inettitudine e la sfortuna dall'altra è proprio la riconsiderazione dei piani attraverso il ragionamento controfattuale.

Il ragionamento controfattuale permette infatti, conoscendo l'esito del piano effettivamente intrapreso, di ipotizzare delle modifiche a tale piano e di immaginarne l'esito (alternativo a quello constatato nella realtà).

Partendo dal caso del successo, se il piano va ipoteticamente a buon fine anche in altre versioni alterate controfattuali, ciò significa probabilmente che il motivo della sua riuscita è da ricercarsi nelle circostanze esterne piuttosto che nella bontà del piano stesso; se invece le alternative controfattuali sembrano destinate a fallire, allora il piano è da considerarsi corretto.

In entrambi i casi, affinché il piano sia idoneo per essere riproposto, è necessario che il rapporto tra costi e benefici sia inferiore nel caso realizzato rispetto alle alternative controfattuali.

Viceversa, in caso di fallimento, se anche i piani alternativi controfattuali non vanno a buon fine, non è ai difetti del piano che può essere imputato l'insuccesso e il piano può al limite essere migliorato solo riducendo i costi; se invece i piani alternativi raggiungono ipoteticamente l'obiettivo, probabilmente nel piano esisteva qualcosa che ha impedito il successo.

Ovviamente la questione è normalmente un po' più complessa, però questo resoconto semplicistico dovrebbe servire a dare una misura della pervasività del ragionamento controfattuale nell'ambito pratico.

Ma in che cosa consiste, più concretamente, la funzione del ragionamento controfattuale? Esso serve a fornire un'euristica per le situazioni future che presentino una rassomiglianza di un certo tipo con quella che è stata affrontata col piano presente. Ciò che l'agente ricava è un'indicazione sul comportamento più appropriato da assumere in circostanze analoghe, alla luce di ciò che si è verificato. L'agente impara dal ragionamento controfattuale, congiunto all'osservazione dell'effettivo svolgimento degli eventi, se il piano che ha appena portato a termine è, a conti fatti, adeguato al tipo di situazione che stava affrontando o se va in qualche modo modificato. In

quest'ultimo caso, spesso il ragionamento controfattuale fornisce anche dei suggerimenti sul tipo di cambiamento da apportare al piano (sulla base dei piani alternativi che hanno controfattualmente un esito più soddisfacente del piano scelto).

L'apprendimento conseguente al ragionamento controfattuale avrà dunque come esito in alcuni casi la revisione del piano di partenza, in altri una conferma dello stesso e dell'opportunità di riutilizzarlo in situazioni analoghe. Ma vediamo un po' più nel dettaglio:

1. **Casi di successo**: il piano è andato a buon fine.

   Ho fatto x e ho ottenuto l'obiettivo y. (6.1)

   L'agente prova comunque a riconsiderare il piano per vedere se il successo non sia fortuito e se il piano non sia migliorabile riducendone i costi.

   (a) *Scenari alternativi ugualmente vincenti*: pur modificando in parte il piano, l'obiettivo sembra comunque raggiungibile.

   Anche se non avessi fatto x (o se avessi fatto al suo posto x'), avrei ottenuto comunque l'obiettivo y. (6.2)

   Se questo avviene a un costo minore nello scenario controfattuale, l'agente cerca di migliorare il suo piano eliminando dei passaggi superflui o sostituendo quelli non ottimali, se invece il rapporto costi-benefici del piano di partenza è ancora quello ottimale, il piano viene confermato.

   (b) *Scenari alternativi perdenti*: modificando anche leggermente il piano, l'obiettivo non sembra più raggiungibile.

   Se non avessi fatto x (o se avessi fatto al suo posto x'), non avrei ottenuto l'obiettivo y. (6.3)

   L'agente decide che quel piano è sufficientemente buono per essere adottato anche in futuro, poiché ne ha ricevuto conferma dalla riconsiderazione controfattuale.

2. **Casi di fallimento**: il piano è fallito.

   Ho fatto x e non ho ottenuto y. (6.4)

   L'agente prova a riconsiderare il piano per vedere se riesce, attraverso qualche modifica, a renderlo più efficace.

   (a) *Scenari alternativi vincenti*: modificando il piano, l'obiettivo sembra raggiungibile.

   Se non avessi fatto x (o se avessi fatto al suo posto x'), avrei ottenuto l'obiettivo y. (6.5)

   L'agente decide che il piano va rivisto nel senso di eliminarne una parte che ha impedito l'esito positivo, oppure di arricchirlo con qualche elemento non precedentemente preso in considerazione.

   (b) *Scenari alternativi ugualmente perdenti*: pur modificando il piano, l'obiettivo non sembra comunque raggiungibile.

   Anche se non avessi fatto x (o se avessi fatto al suo posto x'), non avrei ottenuto comunque l'obiettivo y. (6.6)

   L'agente decide che la disfatta non è dovuta a difetti del piano e decide quindi di confermarlo e adottarlo comunque nel futuro, magari facendo più attenzione ai fattori esterni al piano stesso, oppure di adattarlo a nuovi obiettivi, utilizzando il tipo di razionalità che sarà oggetto del capitolo 7.

Questo è dunque a nostro avviso lo schema generale che si applica sia alla razionalità strumentale che alla razionalità retrospettiva (che sarà trattata nel prossimo capitolo); nel primo caso, l'elemento del piano sottoposto a revisione sono i mezzi "messi in campo" per il raggiungimento del fine – risorse materiali e capacità – nel secondo caso, ciò che viene sottoposto a revisione sono le preferenze dell'agente, che possono variare nel tempo per via di sopravvenuti cambiamenti nell'ambiente o nella sua prospettiva cognitiva stessa. Tale modifica delle preferenze può portare poi con sé, a sua volta, la decisione di perseguire nuovi e diversi obiettivi.

6.2 Il ragionamento controfattuale come processo di revisione o conferma dei piani

Nonostante esistano dei lavori sul pensiero controfattuale nell'ambito della teoria delle decisioni e della scelta razionale (si vedano il seminale lavoro di Kahneman e Tversky, [88], il più recente [89] e [56] in italiano), essi sono perlopiù dedicati alla spiegazione di risultanze sperimentali in disaccordo con le previsioni alla luce di distorsioni indotte dal pensiero controfattuale. La nostra proposta è invece orientata piuttosto a mostrare come il ripensamento del passato sia istruttivo per il futuro e come questo possa essere di fatto impiegato.

Il punto di partenza più naturale per questo tipo di indagine è la forma di razionalità presa in esame dalla letteratura sulla teoria della decisione "classica", la razionalità strumentale o mezzi-fini.

Riprendendo quanto già esposto nel paragrafo 6.1, abbozziamo qui uno schema preliminare per il processo di riconsiderazione controfattuale, finalizzata al cambiamento, che segue l'esecuzione di un piano:

- esecuzione di un'azione;
- fallimento o successo insoddisfacente (costi troppo alti) nel conseguimento dell'obiettivo prefissato;
- riconsiderazione (attraverso il ragionamento controfattuale) dello stato ottenuto (corrispondente o meno all'obiettivo originario) e delle alternative disponibili al momento della scelta;
- confronto (per tutte le alternative) tra il beneficio netto (ottenuto nel caso dell'alternativa scelta e atteso negli altri casi) e i costi netti (sostenuti o da sostenersi per procurarsi i mezzi idonei al raggiungimento del fine);
  - se l'obiettivo è stato raggiunto ma i costi superano i benefici, ricerca di mezzi alternativi più "a buon mercato";

  - se l'obiettivo non viene raggiunto ma i benefici di un piano alternativo ne superano i costi, ricerca di mezzi più idonei (più costosi ma non tanto da superare i benefici);

- nuovo piano.

Riconsiderazione in casi di fallimento

Per quanto riguarda i piani già portati a termine, la situazione che più naturalmente induce gli agenti a riconsiderare l'accaduto è quella in cui il piano non porta al conseguimento dell'obiettivo; partiamo dunque da questo caso.

Normalmente, se un agente ha progettato un piano per ottenere un certo scopo, questo significa che l'agente reputa il raggiungimento di quell'obiettivo come qualcosa di auspicabile che possa migliorare il suo stato e che ritiene che tale obiettivo sia alla sua portata.

Date queste premesse, di fronte a un fallimento la sua prima reazione dovrebbe ragionevolmente essere quella di adoperarsi per vedere se non sia possibile procurarsi dei mezzi più idonei per quel fine. Quindi l'agente, mantenendo invariato il suo impegno verso il conseguimento di quel fine, ipotizza (controfattualmente) di variare i mezzi da impiegare allo scopo. Vediamo qual è il processo sottostante.

Per illustrare i vari possibili meccanismi di revisione dei mezzi che possono entrare in gioco, possiamo partire da un semplice esempio di un ipotetico viaggiatore che decida, partendo da Milano in aereo, di trascorrere una settimana di vacanza a Strasburgo. Poniamo anche che, una volta sul posto, si renda conto che il budget che ha a disposizione non è sufficiente per pagare l'albergo per tutti i giorni di vacanza previsti ed è costretto così a tornare a casa prima del tempo, fallendo l'obiettivo che si era posto. L'agente potrebbe allora esprimere la seguente riflessione:

> Se avessi preso il treno, invece dell'aereo, avrei potuto permettermi una settimana a Strasburgo (6.7)

Ciò che può indurre l'agente al ripensamento può essere, da un lato, un cambiamento avvenuto nell'ambiente che non era stato previsto al momento

della stesura del piano, oppure, dall'altro lato, la presa di coscienza, da parte dell'agente, di un elemento rilevante per il piano, che non aveva considerato o aveva sottovalutato. Vedremo ora due esempi del genere.

Cambiamento nell'ambiente

Come esempio di cambiamento nell'ambiente, immaginiamo che il viaggiatore a un certo punto si fosse interessato delle tariffe aeree e alberghiere e avesse calcolato che viaggiando in aereo e pernottando in un albergo di categoria medio-alta, il suo budget sarebbe stato sufficiente per permettersi una settimana a Strasburgo.

Tuttavia, la settimana prescelta dal nostro viaggiatore coincide con la settimana di apertura del famosissimo mercato di Natale e quindi le tariffe sono inesorabilmente più alte di quelle inizialmente preventivate, ma il viaggiatore ritiene, erroneamente, di mantenere quel piano, che quindi fallisce.

Il fallimento è dunque in questo caso da imputare a un mutamento dell'ambiente, nella fattispecie all'aumento delle tariffe; ciò dovrebbe indurre l'agente a rivedere il suo piano per il futuro, cercando dei mezzi meno dispendiosi (viaggiando in treno o pernottando in un albergo di una categoria inferiore) per trascorrere la sua settimana a Strasburgo.

Se le tariffe non fossero aumentate, avrei potuto permettermi una settimana a Strasburgo (6.8)

Cambiamento nella prospettiva cognitiva dell'agente

A volte, però, il fallimento può avere luogo anche se le circostanze non sono sostanzialmente cambiate rispetto alla situazione che l'agente ha osservato al momento della pianificazione, solamente la rilevanza di qualche elemento già presente era sfuggita all'attenzione dell'agente.

Ritorniamo al caso del viaggiatore: possiamo immaginare che egli abbia deciso di fare il viaggio in aereo, rischiando poi di non avere sufficiente denaro per la vacanza, perché riteneva che il treno impiegasse troppo tempo a percorrere la distanza Milano-Strasburgo.

Dopo la vacanza terminata anticipatamente per via del piano fallimentare, l'agente può acquisire l'informazione che in realtà arrivare in centro a

Strasburgo viaggiando da Milano in treno o in aereo non fa molta differenza, considerato il tempo d'attesa negli aeroporti e gli spostamenti dall'aeroporto alla città.

Allora l'agente formula la seguente riflessione:

Se avessi preso il treno avrei impiegato lo stesso tempo che in aereo e avrei potuto permettermi una settimana a Strasburgo (6.9)

Riconsiderazione in casi di successo

Anche se più rara e meno "vitale", esiste per l'agente la possibilità di riconsiderare un piano che sia andato a buon fine con l'intento di migliorarlo. L'agente potrebbe rendersi conto che, pur avendo raggiunto l'obiettivo, ha sprecato delle risorse troppo costose, che avrebbe potuto risparmiare e impiegare diversamente, poiché il fine sarebbe stato comunque conseguito e alcuni mezzi impiegati erano dunque superflui.

Torniamo all'esempio che ci è ormai familiare: il viaggiatore stavolta aveva un budget un po' più alto ed è riuscito comunque a trascorrere la sua settimana di vacanza a Strasburgo come aveva progettato, ma si rende conto che avrebbe potuto risparmiare un bel po' di soldi senza eccessive perdite di tempo.

Se avessi preso il treno, invece dell'aereo, avrei potuto risparmiare dei soldi sulla vacanza a Strasburgo (6.10)

Cambiamento nell'ambiente

Gli esempi qui sono analoghi a quelli del fallimento: l'agente ha trascorso la sua settimana a Strasburgo, ma ha speso più del dovuto perché le tariffe erano state aumentate a causa del mercato di Natale. L'insegnamento che dovrebbe trarre per il futuro è che un evento non calcolato come una manifestazione particolare può cambiare il costo effettivo di uno dei mezzi che erano stati approntati e questo dovrebbe portare a pensare a dei mezzi alternativi come, in questo caso, il treno.

Se non fossero aumentate le tariffe, non avrei speso così tanto (6.11)

Cambiamento nella prospettiva cognitiva dell'agente

Di nuovo, pur rimanendo inalterata la situazione circostante, l'agente può in un secondo momento, ripensando al piano, scoprire di non aver considerato un elemento che avrebbe potuto fargli comunque raggiungere il fine, ma con un dispendio minore.

Ritornando all'esempio del viaggiatore, la vacanza è andata bene, ma riflettendo su tutto il tempo che ha perso in spostamenti, realizza di aver impiegato più o meno lo stesso tempo che sarebbe stato richiesto da un viaggio in treno, spendendo però molti più soldi.

Può quindi formulare il suo ragionamento in questi termini:

> Se avessi preso il treno, avrei impiegato lo stesso tempo ad arrivare a Strasburgo, risparmiando (6.12)

Un agente ciecamente fedele al suo obiettivo avrebbe a questo punto esaurito tutte le sue possibilità di riconsiderazione dei piani passati. Non sempre però gli agenti – almeno quelli umani – sono così costanti; spesso si arrendono e passano ad altro, oppure possono incontrare sul loro cammino qualcosa che risulti loro più appetibile e li distolga quindi dall'obiettivo inizialmente prefissato. La razionalità che guida gli agenti in questi casi è quella che potremmo definire razionalità *ex-post* o retrospettiva e sarà oggetto del prossimo capitolo.

Capitolo 7

L'atteggiamento controfattuale e la razionalità retrospettiva

Il n'y a rien de si conforme à la raison que ce désaveu de la raison[1]
[Blaise Pascal, *Pensée*, p.272]

Questo capitolo ha lo scopo di mostrare come l'atteggiamento controfattuale si applichi anche al tipo di razionalità meno classico, messo in luce da autori come James March e Herbert Simon, che agisce sulle preferenze dell'agente individuando nuovi obiettivi.

Una precisazione importante a livello preliminare riguarda l'uso dei termini "*ex post*" e "retrospettivo" che, se intesi nella loro accezione temporale abituale, risulterebbero banalmente sempre applicabili al ragionamento controfattuale che, per definizione, si riferisce, nella quasi totalità dei casi, a eventi passati che vengono considerati da una prospettiva posteriore. Si è deciso in questa sede di recuperare la terminologia secondo l'uso che ne fa March.

La versione di razionalità retrospettiva qui presentata differisce da quella di March e per certi versi tenta di ampliarla, proponendone un'applicazione diretta non solo alla giustificazione dello stato acquisito dall'agente, ma anche all'elaborazione di nuovi obiettivi precedentemente non considerati appetibili.

[1]Nulla è così conforme alla ragione come questa sconfessione della ragione. [tr.it. in *Pensieri*, a cura di P. Serini, Einaudi, Torino 1962, p.57]

7.1 La teoria di March rivisitata

Prima di presentare la nostra analisi della razionalità *ex-post* è opportuno fare una premessa: il modello di March porta alle estreme conseguenze la critica alla razionalità classica, tanto che a tratti sembra sottintendere che il ragionamento strumentale stesso sia una sorta di "sovrastruttura" di cui gli economisti e i filosofi si servono per giustificare certi comportamenti osservati, che in realtà sono più spesso l'esito di ricostruzioni a posteriori, secondo la massima enunciata da Mark Twain: "Nella vita reale la cosa giusta non succede mai al posto giusto nel momento giusto: è compito dello storico rimediarvi.", dove allo storico può essere sostituito un più generico "studioso".

Sebbene riteniamo che il fenomeno della giustificazione a posteriori di azioni che non erano parte di un piano del tipo strumentale sia tutt'altro che trascurabile, tuttavia non è questa componente che vorremmo mettere in luce, quanto piuttosto la capacità degli agenti di formare nuovi piani *a partire* da una reinterpretazione dei fatti.

In quest'ottica, quindi, gli agenti, *quando agiscono razionalmente*, tendono a servirsi della razionalità strumentale o della razionalità *ex-post* a seconda delle circostanze; presumibilmente saranno più propensi a ragionare secondo il paradigma strumentale di fronte a situazioni sufficientemente trasparenti, per assumere un atteggiamento più retrospettivo in situazioni caratterizzate da forte ambiguità. Fatte queste premesse, non ci resta che riconsiderare i punti lasciati aperti da March.

- Per cominciare, nel nostro caso la razionalizzazione *ex-post* normalmente parte dall'esito di un'azione che è il risultato di un precedente piano strumentale, anche se non sono esclusi i casi di razionalizzazioni a posteriori di azioni non razionali, ma riflesse. Non esistono quindi problemi di regresso all'infinito.

- Noi consideriamo anche i casi in cui la riconsiderazione parte da un successo: è possibile infatti immaginare che, una volta raggiunto l'obiettivo che si era prefissato, un agente si renda conto che in realtà questo non è così soddisfacente come se lo era figurato (si ricordi il problema delle

preferenze temporalmente incoerenti messo in luce da Elster [50]), oppure che i costi che ha affrontato per raggiungerlo sono stati eccessivi rispetto al beneficio ricavato. Nulla vieta che, anche in questi casi, un agente razionale riconsideri l'accaduto, produca una nuova interpretazione e con essa una revisione endogena delle sue preferenze e infine si ponga un obiettivo diverso per il futuro.

- Sempre in riferimento a quanto affermato prima, un'azione che fallisce un obiettivo non deve *necessariamente* (come sembrerebbe lasciar intuire March) determinare una revisione delle preferenze o una reinterpretazione, anzi, nella maggior parte dei casi gli agenti razionali tendono a modificare semplicemente il piano, includendovi azioni aggiuntive o procurandosi nuove risorse.

- Riteniamo opportuno inoltre precisare qual è lo strumento che un agente razionale ha a disposizione per rivedere la propria interpretazione (ma anche per rinconsiderarla al fine di confermarla e per riconsiderare il piano stesso). Questo compito è assolto nel nostro modello dal ragionamento controfattuale.

- Il criterio per selezionare il nuovo obiettivo (oppure anche, nel nostro caso, il nuovo mezzo da procurarsi), che March identifica con una generica *neighborhood* (vicinanza) è da noi individuato dal beneficio netto atteso (dal raggiungimento dell'ipotetico obiettivo), ridotto dell'effetto costi affondati [17], che sarà analizzato in maggior dettaglio tra breve.

Se, dunque, come fa lo stesso March, ipotizziamo che la razionalizzazione *ex-post* prenda le mosse da un'azione già portata a termine e che ha già fornito un certo risultato (sia esso l'obiettivo inizialmente individuato o uno stato diverso), possiamo dedurre che il processo di riconsiderazione con conseguente (eventuale) revisione delle preferenze sia compiuto attraverso un ragionamento controfattuale. Nella sezione 7.2 affronteremo in dettaglio la spiegazione di come attraverso un ragionamento controfattuale vengano messi a confronto (sia in caso di fallimento che in caso di successo) i corsi d'azione che l'agente ha scelto di perseguire con quelli che erano contemporaneamente disponibili.

Inoltre, nella sezione 6.2, è stato mostrato come il medesimo meccanismo di revisione o conferma possa essere applicato anche ai piani stessi, definendo quali siano le risorse e le capacità che ci si deve procurare e che si devono mettere in opera affinché il piano abbia l'esito più desiderabile.

L'ultimo punto sarà invece oggetto della prossima sezione.

Il principio dei costi affondati e la formulazione di nuovi piani[2]

Secondo gli approcci classici, quando un agente si trova di fronte a una scelta e deve vagliare una serie di alternative, deve calcolare l'utilità (risultante da una combinazione di costi e benefici) di ciascuna alternativa. Nel fare questo, al beneficio atteso totale deve sottrarre i costi sostenuti per acquistare i mezzi necessari al raggiungimento del fine sperato; questi costi vanno a loro volta *ammortizzati* a seconda dell'uso che si prevede di fare di tali mezzi. In altre parole, il ripetuto utilizzo di un mezzo (o una capacità) diminuisce il costo unitario di ogni singola utilizzazione; questo è esattamente quello che in economia viene chiamato *effetto delle economie di riuso*[17].

Tuttavia, gli approcci classici considerano, nel calcolo della funzione di utilità di una determinata alternativa, solamente i costi relativi legati all'uso di quella stessa alternativa, disinteressandosi totalmente dei costi derivanti dal non utilizzo delle altre risorse che erano disponibili all'agente al momento della scelta e che non hanno potuto essere impiegate nel tentativo di conseguire l'obiettivo; questo perché li considerano completamente legati al passato e immodificabili.

Per esempio, si consideri la seguente definizione tratta dall'*Economic Analysis Handbook*:

> **Sunk Cost** – A cost incurred in the past that will not be affected

[2]Le idee contenute in questa sezione e molte delle loro applicazioni analizzate nel capitolo 7 sono frutto del lavoro iniziato negli scorsi mesi in collaborazione con Matteo Bonifacio, Paolo Bouquet e Diego Ponte; una prima esposizione di alcuni dei concetti qui utilizzati è contenuta in [17].

> by any present or future decision. Sunk costs should be ignored in determining whether a new investment is worthwhile[3].
>
> [*Economic Analysis Handbook*]

La posizione inaugurata da March, alla quale noi ci riallacciamo, sostiene che il non utilizzo di una certa risorsa porta con sé una perdita di valore determinata dal mancato ammortamento di quella risorsa, quindi, quando l'agente decide di usare un mezzo e non un altro sfrutta l'effetto delle economie di riuso del mezzo scelto e al tempo stesso perde l'effetto delle economie di riuso legato al mezzo "lasciato da parte"[4].

La teoria classica trascura volutamente questo doppio effetto delle economie di riuso perché in essa si considera che gli investimenti siano sempre almeno in parte reversibili, ad esempio si ritiene che le risorse possono essere vendute per acquistarne delle altre, oppure delle nuove capacità; questo non è sempre vero, poiché certe risorse non sono appetibili per il mercato o sono difficilmente convertibili in qualcos'altro e mantengono quindi un certo tasso di irreversibilità.

L'influenza negativa dell'effetto delle economie di riuso generato dalle risorse inutilizzate, quando è relativo a investimenti irreversibili, viene definito *effetto dei costi affondati*.

Quello che si sta sostenendo è che, finché l'effetto dei costi affondati, sommato agli altri costi, si mantiene al di sotto del beneficio netto atteso dell'obiettivo che si sta perseguendo, ha senso per l'agente continuare a perseguirlo e quindi ricercare nuovi mezzi per raggiungerlo qualora quelli a disposizione

---

[3]**Costo Affondato** – Un costo affrontato nel passato che non sarà influenzato da nessuna decisione presente o futura. I costi affondati dovrebbero essere ignorati nella determinazione dell'opportunità di un nuovo investimento. [*traduzione mia*]

[4]Altri autori, come Brockner e Rubin ([25] e [24]), sottolineano l'importanza di meccanismi psicologici come il "salvare la faccia" a sostegno dell'effetto dei costi affondati. Tuttavia, la perdita di reputazione potrebbe essere considerata un caso particolare di costo affondato, dal momento che questa sembra essere determinata unicamente da azioni compiute nel passato e non da decisioni presenti o future, ma in realtà questa può portare un agente a scegliere di comportarsi in un determinato modo solamente per essere coerente con le proprie scelte del passato.

dovessero rivelarsi insufficienti; quando, all'opposto, l'impatto dell'effetto dei costi affondati diventa tale per cui il conseguimento dell'obiettivo non giustifichi più la spesa, è negli interessi dell'agente (e quindi razionale da parte sua) abbandonare l'obiettivo, rivedere le preferenze e fissarne uno nuovo che permetta un miglior utilizzo delle risorse già disponibili che devono essere "recuperate" ammortizzandone il costo.

Si potrebbe a questo punto avere l'erronea impressione che l'effetto dei costi affondati abbia un ruolo solo nella razionalità *ex-post*, ma non è così, poiché esso non agisce specificatamente sui mezzi o sulle preferenze, ma sugli interi piani. Non è dunque uno strumento caratteristico di un tipo particolare di razionalità, ma è piuttosto utilizzato per orientare la scelta dell'agente verso uno o l'altro tipo di razionalità (classica o retrospettiva) in presenza di un problema specifico.

Riassumiamo quindi brevemente in uno schema (che è una rielaborazione arricchita dello schema precedente tracciato nel paragrafo 6.2) il nostro modello della razionalità:

- esecuzione di un'azione (riflessa o parte di un ragionamento "standard" strumentale);
- fallimento o successo insoddisfacente (costi troppo alti) nel conseguimento dell'obiettivo prefissato;
- riconsiderazione (attraverso il ragionamento controfattuale) dello stato ottenuto (corrispondente o meno all'obiettivo originario) e delle alternative disponibili al momento della scelta;
- confronto (per tutte le alternative) tra il beneficio netto (ottenuto nel caso dell'alternativa scelta e atteso negli altri casi) e i costi netti (comprendenti costi fissi ed effetto dei costi affondati):
  - se i costi superano i benefici, riordinamento delle preferenze e conseguente selezione di un nuovo obiettivo;
  - se i benefici superano i costi, perseveranza nel conseguimento di quell'obiettivo e ricerca di mezzi alternativi;
- nuovo piano.

7.2 Il ragionamento controfattuale sui fini

Abbiamo visto nel capitolo 6 come il ragionamento controfattuale sia un fondamentale strumento per riconsiderare e sottoporre ad analisi un piano che è già stato portato a termine, con esito positivo o negativo. Tale processo di riconsiderazione è necessariamente controfattuale, poiché la scelta dei mezzi idonei al conseguimento del particolare obiettivo è già stata compiuta ed è ormai un dato di fatto.

Ugualmente controfattuale dovrà essere allora anche il ragionamento rivolto alla riconsiderazione di obiettivi che l'agente si sia posto o abbia ricevuto dall'esterno e che consideri come dei "dati di fatto".

La situazione non è molto dissimile da quella discussa in precedenza: proviamo a pensare a un agente che si trovi nella situazione di aver già portato a termine un piano (con successo o meno); nulla vieta a questo agente di riconsiderare, al posto dei mezzi impiegati in vista del raggiungimento del fine, il fine stesso e la sua adeguatezza ai mezzi a disposizione dell'agente.

In altre parole, di fronte a un problema di inadeguatezza tra mezzi e fini, una delle opzioni a disposizione dell'agente (quella considerata nel capitolo 6) è di mantenere invariato il fine che si era prefissato di raggiungere e riconsiderare i mezzi necessari a raggiungerlo; l'altra opzione, che analizzeremo in questo paragrafo, è quella di considerare immutabili i mezzi (per svariati motivi, riassumibili nell'ipotesi dell'effetto dei costi affondati[5] descritta nella sezione 7.1) e "lavorare" sulle preferenze, scegliendo per il futuro un obiettivo più appropriato per le risorse e capacità che l'agente ha a disposizione.

Per essere ancora più precisi, ciò che l'agente sottopone a revisione, in questo caso, non sono né gli obiettivi in sé (l'assunzione di un nuovo obiettivo è piuttosto una conseguenza derivata della revisione in esame), né le preferenze (queste possono essere rivedute per vari motivi, ma la loro revisione non ha necessariamente dei risvolti pratici, poiché un obiettivo può essere massimamente preferibile, ma fuori dalla portata dell'agente), ma le *intenzioni*, cioè, in altri termini, l'agente di volta in volta decide qual è l'alternativa massimamente preferibile tra quelle che gli sono accessibili dati i mezzi che

[5]In poche parole, la perdita derivante dal non utilizzo di altri mezzi che si avevano a disposizione.

ha attualmente a disposizione, oppure quelli che potrebbe ragionevolmente pensare di procurarsi. Tuttavia, per semplicità, continueremo a parlare di revisione delle preferenze, intendendo però sempre le preferenze "accessibili" e quindi l'oggetto delle intenzioni.

Proviamo ora a vedere un po' più nel dettaglio in cosa consiste questa riconsiderazione dei fini nei due casi di successo o fallimento di un piano.[6]

Riconsiderazione in casi di fallimento

Partiamo dunque dalla circostanza più probabile: quella di un agente che ha fallito un piano e che voglia approntarne uno nuovo per il futuro.

In maniera piuttosto ovvia, possiamo osservare che, se i mezzi e il fine non erano adeguati l'un l'altro, la soluzione era da ricercare in due possibilità: individuare mezzi più adeguati, oppure scegliere un obiettivo meno ambizioso[7].

In questo secondo caso, ciò che l'agente deve imparare per il futuro è che, con determinate risorse a disposizione, deve ridimensionare le sue aspettative e impegnarsi verso un fine più facilmente raggiungibile.

Un esempio potrebbe essere offerto dal solito viaggiatore che, con un certo budget, debba recarsi da Milano a Strasburgo per un weekend di vacanza. Poniamo che decida di comprare un biglietto aereo e, una volta arrivato a Strasburgo, si renda conto di non avere più soldi per pagare l'albergo.

Nel riconsiderare l'accaduto, l'agente può giungere alla conclusione che l'obiettivo di trascorrere un weekend a Strasburgo con quel budget era troppo ambizioso e, con i mezzi a lui disponibili, il comportamento più razionale da parte sua sarebbe stato (e presumibilmente sarà in futuro) quello di scegliere

---

[6] Analogamente a quanto succedeva nel caso strumentale, è statisticamente molto più frequente il caso di agenti che analizzano in dettaglio un piano fallito piuttosto che un piano andato a buon fine; tuttavia, poiché riteniamo interessante anche il secondo caso, anch'esso troverà spazio nella nostra analisi.

[7] Un caso non trascurabile è anche quello che si sia presentato un ostacolo inatteso; tuttavia il superamento di tale ostacolo può avvenire attraverso una revisione dei mezzi oppure delle preferenze e il caso è riconducibile quindi a uno dei due segnalati.

una destinazione che offra tariffe più economiche.

Se avessi deciso di andare a Bruxelles, non avrei speso tutti i soldi per il biglietto aereo (7.1)

Esiste però un altro senso nel quale è possibile e razionale "rivedere" gli obiettivi. Pensiamo al caso in cui un agente abbia un obiettivo prefissato e, in seguito all'esecuzione del piano che ha ideato, non lo raggiunga, ma raggiunga comunque uno stato diverso da quello dal quale era partito.

In questo caso potrebbe anche verificarsi l'eventualità che l'agente sia contento di questo nuovo e inatteso risultato e questo per almeno due ragioni.

In primo luogo, la situazione esterna potrebbe avere subito un'evoluzione, per cui ciò che prima appariva come non desiderabile ora al contrario lo sia e il fine attualmente raggiunto dall'agente si riveli più appropriato all'ambiente per come si presenta sotto le nuove condizioni, in rapporto al fine inizialmente perseguito.

In secondo luogo, ciò che potrebbe anche accadere è che il fine attualmente raggiunto non fosse stato nemmeno preso in considerazione dall'agente, che non si era neppure posto il problema della sua desiderabilità. Ciò che accade quindi nel momento in cui il piano viene portato a compimento e uno stato diverso da quello identificato dall'agente come obiettivo viene raggiunto è che l'agente lo include nel suo orizzonte di ragionamento e ne valuta la desiderabilità, scoprendo che è un fine degno di essere perseguito in circostanze analoghe.

Quindi, che sia l'ambiente esterno a cambiare o che sia la prospettiva dell'agente, una modifica nel quadro di riferimento può determinare un diverso ordinamento delle preferenze dell'agente e, di conseguenza, l'individuazione di nuovi obiettivi e fini.

Cambiamento nell'ambiente

Consideriamo, come esempio della prima possibilità, il caso che l'agente fosse particolarmente interessato a visitare Strasburgo avendo saputo di una mostra di arte contemporanea che avrebbe dovuto svolgersi là in concomitanza con la sua visita. Poniamo anche che, alla fine, il nostro viaggiatore avesse rinunciato all'idea di Strasburgo e avesse ripiegato su Bruxelles (considerata

meno "appetibile") e là avesse assistito a un'altra bellissima mostra. Inoltre supponiamo che all'agente capiti anche di leggere sul giornale che la mostra di Strasburgo è stata annullata.

Tutti questi fattori esterni potrebbero portare l'agente a riconsiderare la situazione sotto una nuova luce:

Se fossi andato a Strasburgo non avrei potuto assistere a nessuna mostra (7.2)

Cambiamento nella prospettiva cognitiva dell'agente

Un esempio del secondo caso potrebbe invece essere dato da una riconsiderazione dell'agente che, una volta giunto a Bruxelles, considera il fatto che in fondo questa è una città nuova per lui, mentre Strasburgo l'aveva già visitata in passato.

Se fossi andato a Strasburgo, non avrei visitato una nuova città (7.3)

Nei due casi precedenti, l'individuazione di un nuovo fine è conseguenza di un ribilanciamento del rapporto tra benefici attesi (calcolati anche tenendo conto delle probabilità di riuscita) e perdite attese (effetto del principio dei costi affondati) e, quindi, in qualche modo anche la "minore desiderabilità a posteriori" dell'obiettivo originario contribuiva alla scelta dell'agente.

Esistono però anche casi in cui, nonostante l'immutata propensione dell'agente verso il fine originario, questo può essere abbandonato a vantaggio di un nuovo fine *solamente* a causa di un aumento netto e spesso inatteso dei costi affondati.

Per esempio, l'agente potrebbe avere erroneamente calcolato all'inizio che il budget potesse essere sufficiente per il suo viaggio a Strasburgo, ma l'offerta di cui aveva pensato di usufruire potrebbe non essere più valida. A quel punto, però, avendo già chiesto le ferie al lavoro, decide di partire ugualmente alla volta di Bruxelles, pensando:

Se il bigiletto fosse costato meno, sarei andato a Strasburgo (7.4)

Riconsiderazione in casi di successo

Analizziamo ora simmetricamente la possibilità di riconsiderare controfattualmente un piano che abbia avuto successo. L'utilità di eseguire questo tipo di riconsiderazione discende dal fatto che in realtà potrebbe anche esistere una sproporzione tra mezzi e fini tale per cui le risorse impiegate per il raggiungimento dell'obiettivo eccedano quelle realmente necessarie.

Le soluzioni che l'agente ha a disposizione sono di due tipi: o, come abbiamo visto nel paragrafo 6.2, l'agente decide di "risparmiare" alcuni mezzi, evitando di utilizzarli perché superflui, oppure può decidere di dirigersi verso un obiettivo più ambizioso, sfruttando cosí al massimo le proprie potenzialità. Questa seconda eventualità è un esempio di riconsiderazione controfattuale che porta alla definizione di un nuovo obiettivo.

Pensiamo all'agente di prima che sia riuscito, però, questa volta, a recarsi a Strasburgo come previsto. Modificando leggermente l'esempio, potremmo pensare che l'agente avesse vinto a un'estrazione un bonus per acquistare biglietti aerei su rotte europee e, dopo aver comprato il biglietto per Strasburgo, si fosse reso conto che gli avanzavano dei soldi e avrebbe quindi potuto comprare un biglietto per un'altra, più costosa, destinazione.

Se fossi andato a Parigi avrei speso meglio il bonus (7.5)

Cambiamento nell'ambiente

Come per il fallimento, anche in caso di successo può accadere che le condizioni esterne cambino, rendendo in seguito insoddisfacente un fine che inizialmente era appetibile e, conseguentemente, eccessivi i mezzi predisposti per raggiungerlo. Questo, ancora una volta, potrebbe indurre l'agente a impiegare in futuro i mezzi a sua disposizione in vista di un fine più adeguato.

Nell'esempio di prima, potrebbe darsi il caso che l'agente non avesse inizialmente considerato Parigi perché non era a conoscenza delle offerte sui voli aerei diretti lì.

Se non avessi deciso di andare a Strasburgo, avrei potuto permettermi un viaggio a Parigi (7.6)

Cambiamento nella prospettiva cognitiva dell'agente

Ancora una volta in analogia con il caso del fallimento, a cambiare, invece delle circostanze esterne, potrebbe essere la prospettiva cognitiva dell'agente, cambiamento questo che potrebbe indurlo a rivedere la sua valutazione delle preferenze, in seguito all'ingresso di un fattore nuovo e precedentemente non considerato nella sua scala delle preferenze.

Nel solito esempio, l'agente potrebbe essersi ricordato, solo in seguito all'acquisto del biglietto per Strasburgo, della presenza a Parigi di un suo vecchio amico, trasferitosi là per lavoro.

Se avessi deciso di andare a Parigi, avrei potuto visitare Marco (7.7)

Anche in caso di successo, i soli costi affondati (in questo caso il bonus) possono essere sufficienti a rendere più appetibile un fine piuttosto che un altro (il viaggio a Parigi, normalmente più costoso, rispetto al viaggio a Strasburgo).

Abbiamo quindi visto in questo paragrafo come il riconsiderare retrospettivamente le alternative che erano possibili al momento della scelta ma che non sono state sottoscritte (e sono rimaste quindi controfattuali) sia un modo che gli agenti hanno a disposizione per essere *reattivi*, nel senso di porsi nuovi fini di fronte a cambiamenti nell'ambiente circostante, ma anche e soprattutto *proattivi*, nel senso di generare per se stessi dei nuovi obiettivi sulla base di una mutata prospettiva cognitiva sul problema.

7.3 Esempio riassuntivo

In questo paragrafo si cercherà, attraverso un esempio, di mostrare come un agente possa decidere di volta in volta quale tipo di razionalità adottare nella riconsiderazione dei propri piani e si tenterà al tempo stesso di fornire una rappresentazione schematica, fondata sul sistema introdotto nel capitolo 4, dei diversi tipi di situazione emersi dall'analisi dei capitoli 6 e 7.

L'esempio che verrà utilizzato è quello di un giocatore di scacchi che, alla fine di una partita, rifletta su quanto fatto alla luce dei risultati ottenuti e consideri se, a un certo punto della partita, non avrebbe potuto abbandonare la strategia che di fatto ha portato a termine e ingaggiarne una migliore.

Il gioco degli scacchi è un esempio particolarmente appropriato, innanzitutto perché è un dominio chiuso e poi, fornendo qualche regola supplementare, come a volte può essere fatto in occasione di qualche torneo, è possibile rappresentare in maniera abbastanza fedele tutti i concetti precedentemente introdotti.

Per esempio, ipotizziamo che la vittoria finale dia un certo punteggio (beneficio) e la patta dia un punteggio minore, ma a entrambi i giocatori; a questo punteggio vanno però sottratti dei punti per ogni mossa compiuta (costo) e per ogni pezzo perso (costo affondato)[8].

Se un giocatore si ritrova a riconsiderare retrospettivamente la strategia che ha adottato, dovrà possedere un metodo che gli consenta di decidere in ogni situazione se la riconsiderazione più appropriata per la circostanza sia quella basata sulla razionalità strumentale o su quella *ex-post*.

In base alla sua esperienza passata, un giocatore saprà che, superata una certa soglia di punti persi, la proabilità di ottenere un punteggio finale soddisfacente sarà molto bassa. In particolare, la soglia sarà relativa ai pezzi persi (ovvero ai costi affondati) perché non più recuperabili.

In sostanza, se ϵ è la soglia, quando i costi affondati della strategia sono minori di ϵ, il giocatore metterà in atto una riconsiderazione strumentale e, perseverando nella volontà di raggiungere il fine che si era inizialmente posto, ricercherà mezzi "meno costosi" per raggiungerlo (nella fattispecie, percorsi più brevi); d'altro canto, quando tali costi sono maggiori di ϵ, il buon esito della strategia può essere considerato compromesso e quindi il giocatore dovrà compiere una riconsiderazione *ex-post* fissando un nuovo obiettivo che gli permetterà di recuperare almeno in parte i costi (scegliendo un obiettivo più "a portata" del percorso intrapreso fino a quel momento).

Una volta deciso quale dei due tipi di razionalità adottare nella riconsiderazione, il giocatore metterà a confronto la strategia portata a termine con una strategia ipotetica alternativa (controfattuale), con percorsi o obiettivo diversi a seconda dei casi. Il confronto tra le due strategie verrà compiuto

[8]Si considerano affondati i costi derivati dalla perdita di un pezzo perché questo non è più recuperabile in nessun modo, mentre una mossa può sempre essere "riusata" in strategie diverse.

sulla base del rapporto tra benefici e costi. Se tale rapporto (che chiameremo beneficio netto) è maggiore per la strategia effettivamente perseguita, questa verrà confermata, se invece è maggiore per la strategia alternativa, il giocatore rivedrà l'ordine di preferibilità delle strategie e in futuro, in situazioni analoghe, propenderà per la nuova strategia ipotetica.

Riassumendo, il processo che stiamo descrivendo si compie secondo i seguenti passi:

- un giocatore porta a termine una strategia con un determinato risultato
- individua una fase della strategia da sottoporre a revisione
- calcola quanti punti ha perso in termini di pezzi eliminati dal gioco
- se la cifra è minore di ϵ, avvia una riconsiderazione strumentale, altrimenti avvia una riconsiderazione *ex-post*
- in entrambi i casi, costruisce una strategia alternativa ipotetica
- mette a confronto strategia "attuale" e strategia "controfattuale"
- quella che consente di ottenere il migliore punteggio finale è quella che sarà riprodotta in futuro.

Riprendendo quanto precedentemente presentato nel capitolo 4, potremmo rappresentare una strategia come un contesto costituito da un certo insieme di modelli locali (sottoinsieme di tutte le possibili combinazioni di mosse dei due giocatori), definito da due parametri: la disposizione dei pezzi "superstiti" sulla scacchiera e l'obiettivo (che può essere di vincere la partita o di raggiungere un pareggio).

Una volta costruito il contesto della strategia "attuale", il contesto della strategia controfattuale sarà, come mostrano le figure 7.1, 7.2, 7.3 e 7.4, nel caso della razionalità strumentale, un contesto al cui interno verrà cambiato il valore di una formula (un passo della strategia) mentre, nel caso della razionalità *ex-post*, un contesto definito dal diverso valore di un parametro (l'obiettivo).

Una volta costruito il contesto della strategia controfattuale, si ragiona al suo interno per vedere i possibili esiti:

- se tutti i modelli locali del contesto della strategia controfattuale ottengono un punteggio migliore della strategia attuale, il giocatore inferirà che "se avesse cambiato strategia avrebbe ottenuto un risultato migliore" e presumibilmente userà la strategia "rivista" in futuro;

- se tutti i modelli locali del contesto della strategia controfattuale ottengono un punteggio peggiore della strategia attuale, il giocatore inferirà che "se avesse cambiato strategia avrebbe ottenuto un risultato peggiore" e presumibilmente confermerà la vecchia strategia e la userà in futuro;

- se alcuni modelli locali del contesto della strategia controfattuale ottengono un punteggio migliore della strategia attuale e altri peggiore, il giocatore inferirà che "se avesse cambiato strategia avrebbe potuto ottenere un risultato migliore" e la sua scelta per il futuro dipenderà dall'euristica (si ricordi la differenza tra razionalità sostanziale e procedurale enunciata nella sezione 5.2.1) che deciderà di adottare, che determinerà quanti o quale percentuale di modelli locali "vincenti" il giocatore giudicherà sufficienti a imporre un cambio di strategia.

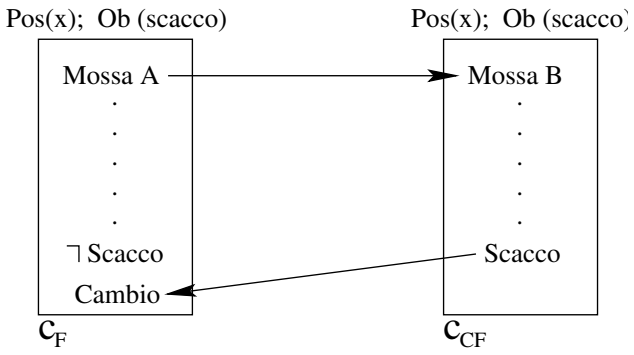

Figura 7.1: Riconsiderazione strumentale in caso di fallimento

Le figure 7.1, 7.2, 7.3 e 7.4, sono altrettanti possibili scenari semplificati che potrebbero presentarsi. Come si vede, nel caso strumentale il contesto controfattuale viene costruito attraverso il cambiamento di un fatto che si

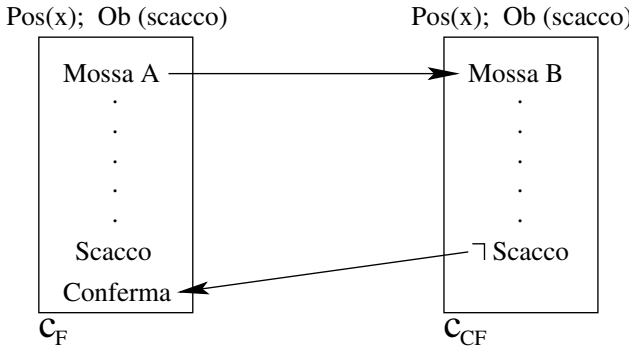

Figura 7.2: Riconsiderazione strumentale in caso di successo

trova *all'interno* del contesto fattuale, mentre nel caso retrospettivo il cambiamento ha luogo nei parametri che definiscono i due contesti. Ovviamente, i quattro scenari presentati non esauriscono tutte le possibili situazioni: solo per fare un esempio, la riconsiderazione strumentale in caso di fallimento può dare un risultato peggiore nel contesto controfattuale, oppure un risultato migliore rispetto a quello del contesto fattuale, ma neppure quello coronato da successo.

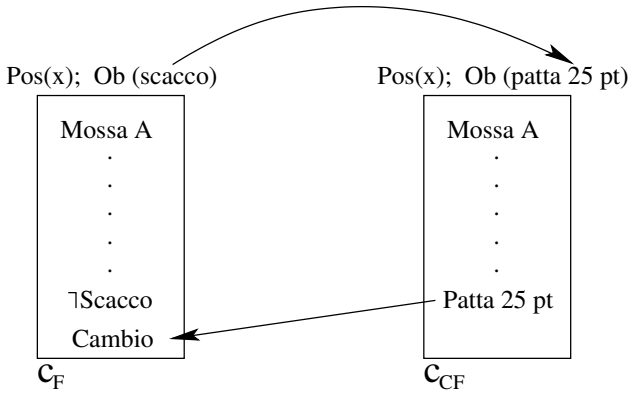

Figura 7.3: Riconsiderazione *ex-post* in caso di fallimento

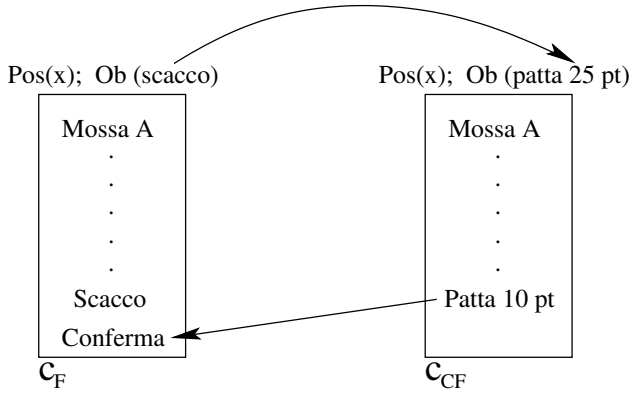

Figura 7.4: Riconsiderazione *ex-post* in caso di successo

Parte III

Sviluppi futuri

Capitolo 8

Sviluppi futuri

> Rational reconstruction [...] cannot be comprehensive since human beings are not *completely* rational animals; and even when they act rationally they may have a false theory of their own rational actions[1].
>
> [Imre Lakatos, History of science and its rational reconstructions, p.114]

Questo capitolo ha lo scopo di mostrare delle applicazioni specifiche molto interessanti di quanto è stato presentato finora. Sebbene i temi presentati non siano sviluppati dettagliatamente in questa sede, tuttavia ci sembra importante mostrare, a volte anche solo con degli accenni, che questo ambito di ricerche è stato a oggi esplorato solo in minima parte ed è a nostro avviso ancora molto fecondo.

Nella sezione 8.1 cerchiamo di mostrare come la bidirezionalità tra mezzi e fini nella razionalità decisionale trovi un analogo nella razionalità scientifica: la bidirezionalità tra evidenza sperimentale e nucleo teorico. La scelta della direzione da prendere – quindi del tipo di razionalità da utilizzare – e, di conseguenza, la scelta tra muoversi all'interno del paradigma teorico tradizionale o mettere in atto una rottura con esso, viene effettuata attraverso un confronto tra sistemi teorici accettati (e quindi fattuali) e alternative ipo-

[1] La ricostruzione della scienza [...] non può essere onnicomprensiva, dal momento che gli esseri umani non sono animali *completamente* razionali; e anche quando agiscono razionalmente possono avere una falsa teoria delle loro azioni razionali. [*tr. it. di Marcello D'Agostino*: [96]]

tetiche (in un certo senso controfattuali); questo processo comparativo può considerarsi un'istanza di ragionamento controfattuale.

Nella sezione 8.2 si esamina la possibilità di costruire agenti artificiali in grado di ragionare secondo entrambi i tipi di razionalità precedentemente individuati e si propone il ragionamento controfattuale come supporto metodologico per l'agente nella scelta di quale dei due tipi di razionalità sia da impiegarsi in ogni specifico caso; si individua infine in questa capacità di scelta sistematica un elemento in favore dell'autonomia degli agenti.

Infine, la sezione 8.3 propone di allargare l'uso delle nozioni e dei metodi utilizzati per gli agenti singolarmente presi anche a scenari multiagente, affinandoli in modo da rendere conto dei processi di attribuzione di credenze, intenzioni ecc. anche ad altri agenti.

8.1 Razionalità scientifica e controfattuale

In questa sezione applichiamo lo schema esplicativo presentato nei capitoli precedenti in relazione alla razionalità su azioni intesa nel senso più ampio a un dominio specifico: quello della razionalità scientifica.

Verrà quindi tentata una ricostruzione di alcune celebri teorie epistemologiche (principalmente quelle di Kuhn e Lakatos) parallela rispetto alla dicotomia razionalità strumentale/*ex-post* e si cercherà di mostrare come questa direttrice possa essere intersecata con quella della controfattualità anche nell'ambito specifico della razionalità scientifica.

8.1.1 I due tipi di razionalità nell'impresa scientifica

Nella sezione 7.1 è stata presentata una lettura bidirezionale della razionalità pratica mostrando come sia possibile per un agente mantenere fisse le sue preferenze e ragionare sui mezzi a sua disposizione, oppure mantenere fissi tali mezzi e ragionare sulle proprie preferenze. Allo stesso modo, questo paragrafo ha lo scopo di rintracciare una bidirezionalità anche all'interno dell'impresa scientifica, i cui attori possono mantenere fisso il nucleo teorico ragionando sull'esito degli esperimenti, ma anche mantenere fisso l'esito degli

esperimenti per ragionare sul nucleo teorico, aprendo così la strada a una rivoluzione scientifica.

L'interesse di sottoporre a un tale tipo di analisi la razionalità scientifica è duplice. Da un lato, se per azione razionale, da quanto affermato nella sezione 5.1, si intende un'azione conseguente a un processo di ragionamento e pianificazione che segua un metodo rigoroso e delle procedure corrette, allora un'azione compiuta in ambito scientifico dovrebbe essere massimamente razionale, essendo il ragionamento scientifico per definizione guidato da regole rigorose.

La seconda ragione per la quale questo genere di trattazione dovrebbe risultare di un certo interesse è il dibattito che, soprattutto negli ultimi trent'anni, ha ricoperto un ruolo centrale in filosofia su quali dovrebbero essere le caratteristiche del modo di procedere della scienza.

Il tentativo che viene fatto in questa sede è quello di, in un certo senso, parafrasare il lavoro di studiosi del calibro di Karl Popper, Imre Lakatos, Thomas Kuhn, Hilary Putnam, Norwood Russell Hanson e numerosi altri, cercando di esplicitare le due direzioni (dal nucleo teorico agli esperimenti e dagli esperimenti al nucleo teorico) servendoci come chiave di lettura dell'analisi già compiuta sulla razionalità intesa in senso più generale.

Gli esperimenti possono in senso lato essere considerati come parte dell'apparato strumentale della scienza, insieme alle strumentazioni vere e proprie, quelle fornite dalla tecnologia e quindi il ragionamento sugli esperimenti può essere considerato una sorta di applicazione della razionalità strumentale alla scienza.

D'altra parte, il nucleo teorico è un sistema di riferimento che struttura la visione scientifica della realtà e gioca un ruolo molto simile a quello che occupa l'insieme di preferenze di un agente a livello di decisioni di senso comune, poiché proprio attraverso di esso l'agente costruisce la propria visione della realtà. Il ragionamento sugli assunti teorici può essere dunque avvicinato alla razionalità *ex-post*.

8.1.2 Dagli esperimenti alla teoria

Il primo filosofo a porre esplicitamente due fasi ben distinte all'interno dell'impresa scientifica è stato Thomas Kuhn. Tuttavia, questa stessa distinzione è presente in maniera più implicita anche nei lavori di altri filosofi a lui contemporanei, come ad esempio Popper e Lakatos. Noi manterremo la distinzione, che ci sarà utile per mostrare come ciascuna delle due fasi sia caratterizzata da una delle due forme di ragionamento appena introdotte.

Kuhn parla di periodi che definisce di scienza *normale* e periodi di *rivoluzione scientifica*. Nei periodi di scienza normale la comunità scientifica si muove all'interno di un paradigma teorico che si è affermato sugli altri perché più atto a risolvere una serie di problemi [91]:

> Closely examined, whether historically or in the contemporary laboratory, that enterprise seems an attempt to force nature into the preformed and relatively inflexible box that the paradigm supplies[2].
>
> [*The Structure of Scientific Revolutions*, p.24]

Il compito degli scienziati nei periodi di scienza normale è quello di continuare a risolvere problemi (i cosiddetti *rompicapo*) "accomodandoli" nell'apparato teorico consolidato [91].

> Bringing a normal research problem to a conclusion is achieving the anticipated in a new way, and it requires the solution of all sorts of complex instrumental, conceptual, and mathematical puzzles[3].
>
> [*The Structure of Scientific Revolutions*, p.36]

[2]Esaminata da vicino, storicamente o nel laboratorio contemporaneo, quell'impresa sembra tentare di forzare la natura all'interno della preformata e relativamente inflessibile scatola che il paradigma fornisce. [*traduzione mia*]

[3]Portare a conclusione un problema di ricerca normale è raggiungere ciò che si era anticipato in un modo nuovo e richiede la soluzione di ogni sorta di complesso rompicapo strumentale, concettuale e matematico. [*traduzione mia*]

Similmente, Popper parla di teorie che vengono fornite insieme a una serie di *falsificatori potenziali*, esperimenti che, nel momento in cui avessero un certo esito, falsificherebbero la teoria. Durante i periodi di scienza normale, i falsificatori mantengono il loro carattere potenziale e la teoria riesce a non essere inficiata da essi.

Lakatos, invece, non parla di teorie ma di programmi di ricerca, che possono essere progressivi o regressivi; quando essi sono percepiti dalla comunità scientifica come *progressivi*, ogni volta che si presenta un risultato sperimentale anomalo, i suoi effetti vengono limitati alla cosiddetta *cintura protettiva*, cioè un insieme di assunti e principi di importanza non fondamentale rispetto invece a quello che è il vero "cuore" della teoria, ossia il *nucleo* che, fintanto che il programma viene giudicato progressivo, non viene intaccato. Quando invece un programma viene giudicato *regressivo*, è il suo nucleo stesso a subire modificazioni.

Quindi, nei periodi in cui un programma di ricerca è giudicato progressivo (dalla comunità scientifica) o, in altre parole, nei periodi di scienza normale, il sistema di riferimento teorico viene mantenuto fisso e la ricerca, di fronte agli insuccessi, tende ad elaborare nuove strumentazioni e a modificare gli esperimenti in modo che non contrastino con la teoria e anzi ne fungano il più possibile da conferma.

Per utilizzare una metafora di carattere economico, ogni teoria scientifica (o programma di ricerca) ha dei benefici attesi, nella forma di problemi che si propone di risolvere e al tempo stesso comporta dei costi per i ricercatori, costi legati ai mezzi – quindi alle strumentazioni e agli esperimenti – sia, banalmente, dal punto di vista economico, sia dal punto di vista cognitivo. Tutti questi costi, comprendenti le spese per acquistare apparecchiature adibite all'esecuzione di determinati esperimenti e al tempo stesso gli investimenti in termini di risorse umane che dedicano il loro lavoro alla risoluzione di problemi interni al paradigma, normalmente vengono "ammortizzati" ogniqualvolta un esperimento conferma la teoria. Può però accadere che a un certo punto questi "ammortamenti" diminuiscano e certi costi divengano, quindi, affondati[4].

[4]È importante notare, sempre in analogia con il caso generale, che si danno entrambe

Ricapitolando, in situazioni di scienza normale, l'"agente razionale scientifico" formula un piano al fine di ottenere una conferma alla propria teoria (l'obiettivo è di portare alla massima realizzazione la teoria che lui sostiene), oppure di risolvere dei rompicapo che aggiungano dei "nuovi pezzi" alla teoria. Tale piano prevede l'utilizzo di una strumentazione apposita ed è composto di una serie di azioni da compiere, ovvero di esperimenti. Ogni piano comporta dei costi (sia fissi, consistenti nei costi effettivi degli esperimenti che confermano la teoria, sia affondati, conseguenti all'accumulo di evidenze negative che non si inquadrano nella teoria)[5].

Fino a quando gli utili dalla teoria (e cioè i problemi che risolve) sono maggiori dei suoi costi affondati, la razionalità impiegata continua a essere quella strumentale, che sottopone al vaglio gli esperimenti stessi. Fuori di metafora, finché è possibile accomodare le anomalie all'interno del sistema teorico, si tenta di reinterpretare o di ricontrollare gli esperimenti e si permane in un periodo di scienza normale.

A un certo punto, però, può verificarsi l'eventualità che tali costi affondati (determinati dagli insuccessi) arrivino a superare i benefici (cioè i problemi ai quali viene data soluzione) e quindi un programma di ricerca, come direbbe Lakatos, cessa di essere progressivo e diviene regressivo. A questo stadio l'agente razionale è costretto, per poter recuperare i suoi costi affondati (attraverso il riuso, dirigendoli verso teorie alternative), ad abbandonare quella che abbiamo definito prospettiva strumentale e ad abbracciare quella *ex-post*.

Questo significa che, pur di "recuperare" i costi investiti nella mole di esperimenti che non rientrano nella teoria dominante (i costi affondati del caso), l'agente si vede costretto a rivedere i suoi presupposti teorici e a op-

---

le possibilità: sia che l'empiria minacci di falsificare la teoria, sia che la confermi ma a costo troppo elevato, cioè che risolva problemi modesti a costo di grossi "accomodamenti". Ovviamente, la prima possibilità è quella che più direttamente dà luogo a ripensamenti del piano, ma anche la seconda va presa in considerazione.

[5]Chiaramente, quando si parla di costi nella scienza ciò che si ha in mente non sono solo i costi "monetari" per apparecchiature e allestimento di esperimenti, ma in senso più lato anche i costi "cognitivi" impiegati nell'elaborazione e conseguente corroborazione delle teorie scientifiche.

tare per un'altra teoria (o programma di ricerca), ciò che Kuhn definisce *rivoluzione scientifica*.

La nozione di costi affondati può essere interpretata secondo una duplice prospettiva: da una parte ci sono i costi affondati intrinseci a una teoria (quelli che abbiamo appena spiegato e che, una volta accumulati oltre una certa soglia, determinano l'abbandono della teoria) e dall'altra ci sono i costi affondati che gli agenti (in questo caso i singoli scienziati) hanno personalmente affrontato per confermare la teoria, in termini di energie intellettuali spese.

Non è difficile dedurre che proprio gli agenti che hanno personalmente meno costi affondati relativamente alla vecchia teoria (nel senso che hanno speso meno energie e capitali in esperimenti volti a confermare quella teoria) saranno più propensi ad abbandonarla per compiere la rivoluzione, come lo stesso Kuhn ha fatto notare in [91]:

> Almost always the men who achieve these fundamental inventions of a new paradigm have been either very young or very new to the field whose paradigm they change. And perhaps that point need not have been made more explicit, for obviously these are the men who, being little committed by prior practice to the traditional rules of normal science, are particularly likely to see that those rules no longer define a playable game and to conceive another set that can replace them[6].
>
> [*The Structure of Scientific Revolutions*, p.90]

Per concludere, gli scienziati che personalmente hanno meno costi affondati in una teoria sono quelli più propensi a compiere la rivoluzione, ma questa è possibile solo grazie alla presenza di teorie alternative che permettono di

[6]Quasi sempre gli uomini che realizzano queste fondamentali invenzioni di un nuovo paradigma sono o molto giovani oppure molto nuovi al campo il cui paradigma cambiano. E forse questo punto non ha bisogno di essere esplicitato, poiché ovviamente questi sono gli uomini che, essendo poco legati dalle pratiche precedenti alle tradizionali regole della scienza normale, sono particolarmente propensi a vedere che quelle regole non definiscono più un gioco giocabile e a pensare a un altro insieme [di regole] che le sostituisca. [*traduzione mia*]

"riusare" i costi affondati della vecchia teoria, che in essa superano ormai i benefici attesi. La vecchia teoria è allora abbandonata a favore di un'altra che ammortizzi meglio i costi sostenuti. Il tipo di razionalità utilizzato in questi casi (che sono, secondo la definizione di Kuhn, rivoluzioni scientifiche) è, come argomenteremo nel prossimo paragrafo, la razionalizzazione *ex-post*.

8.1.3 Dalla teoria agli esperimenti

Abbiamo finora approfondito il tema della forma di ragionamento adottata dagli scienziati afferenti al paradigma dominante durante i periodi di scienza normale. Come abbiamo visto, però, a un certo punto può accadere che la teoria (o il programma di ricerca) attraversi un periodo di crisi, nel quale i benefici attesi (i rompicapo risolti) siano superati dai costi (connessi agli esperimenti – soprattutto quelli falliti) sostenuti.

Quando molti sforzi, compiuti in direzione della corroborazione di una teoria, non sono andati a buon fine (o non hanno ottenuto un risultato all'altezza delle aspettative), la comunità scientifica può esprimere la volontà di "recuperare" i costi legati alla mole di evidenze che erano state accantonate, mantenere queste ultime fisse e, sulla base di esse, ricercare una teoria che ne renda conto. Dunque, durante i periodi di crisi che precedono una rivoluzione scientifica, un cambiamento di paradigma o il passaggio da un programma di ricerca a un altro, il tipo di razionalità impiegata sembra essere quella *ex-post*.

A riprova di questo, possiamo anche indicare il fatto che la revisione delle preferenze nell'esposizione fatta da March [112] e il cambiamento di paradigma nell'ottica kuhniana [91] condividono la peculiarità di portare con sé una nuova interpretazione del mondo circostante.

Sono facilmente rilevabili le analogie presenti in quanto scrive March in [111]:

> We leave a decision world with coherent intentions, expectations, identities, and rules. Decisions are seen as vehicles for constructing meaningful interpretations of fundamentally con-

> fusing worlds, not as outcomes produced by a comprehensible environment[7].
>
> [*A Primer on Decision Making: how Decisions Happen*, p.179]

e quanto affermato da Kuhn in [91]:

> [...] during revolutions scientists see new and different things when looking with familiar instruments in places they have looked before.
>
> [...] paradigm changes do cause scientists to see the world of their research-engagement differently[8].
>
> [*The Structure of Scientific Revolutions*, p.111]

Un cambiamento di paradigma, quindi, equivale a una rilettura di tutti i fenomeni in esame alla luce di un'interpretazione che, fino a quel momento, era stata *non standard* e da quel momento in poi diventa standard.

Riallacciandoci allo schema presentato nella sezione 7.1, possiamo fornire una rappresentazione più dettagliata delle varie fasi della razionalizzazione *ex-post* nella pratica scientifica.

Per prima cosa bisogna ricordare che nella scienza una razionalizzazione *ex-post* prende le mosse da un'evidenza empirica non in linea con le aspettative, ossia con una serie di fatti che non trovano una sistemazione nella teoria "dominante".

Quando questa evidenza empirica "discordante" supera una certa soglia, per gli scienziati diventa piuttosto difficile accomodarla; essi decidono dunque di *costruire* a partire da essa un nuovo scenario teorico (anche più di

---

[7]Le nuove teorie di decisione lasciano un mondo basato su intenzioni, aspettative, identità e regole coerenti; le decisioni sono viste come veicoli per costruire interpretazioni significative di mondi fondamentalmente confusi, non come esiti prodotti da un ambiente comprensibile. [*tr. it. di Stefano Micelli in*: [112]]

[8][...] durante le rivoluzioni gli scienziati vedono nuove e differenti cose quando guardano con strumenti familiari in posti dove avevano guardato prima. [...] i cambiamenti di paradigma portano gli scienziati a vedere il mondo dove applicano la loro ricerca in maniera diversa. [*traduzione mia*]

uno), che all'inizio contravviene a quelli che vengono accettati come "fatti" (è quindi l'analogo di un *contesto controfattuale*, concetto che è stato spiegato diffusamente nella sezione 4.3)[9].

Arrivati a questo punto, gli scienziati, mantenendo fissa l'evidenza empirica problematica (ormai accettata), metteranno a confronto i vari sistemi teorici per capire quale tra essi ottimizzi il rapporto tra problemi risolti ed evidenza contraria.

Se dal confronto a uscire vittoriosa sarà la nuova teoria, essa diverrà dominante e a partire da essa verranno formulati nuovi obiettivi per la ricerca e verranno individuati nuovi problemi da risolvere.

Possiamo infine riassumere, con uno schema analogo a quello del paragrafo 7.1, comprendente sia la razionalità strumentale che quella *ex-post*, il discorso fin qui presentato sulla razionalità scientifica.

- All'interno di un paradigma scientifico dominante si presenta un'anomalia nella forma di evidenza empirica contraria.
- Si riconsidera il rapporto tra i problemi risolti e le anomalie irrisolte dalla teoria dominante, tenendo conto dell'influenza negativa degli esperimenti compiuti con esito fallimentare, che risultano inutili ai fini della corroborazione della teoria dominante e invece potrebbero rivelarsi importanti per delle teorie rivali (comportando dei costi affondati).
- Si ripete questo calcolo con le altre teorie rivali che rendono conto proprio di quella specifica evidenza negativa che mette in crisi la teoria dominante:
  - se nella teoria dominante le anomalie superano i problemi risolti e altre teorie invece si comportano meglio, la comuinità scientifica abbraccia la teoria "migliore" e rivede le proprie assunzioni di

[9]Per precisare meglio l'intuizione possiamo aggiungere che in un primo momento la teoria è pensata come alternativa a quella che dovrebbe descrivere adeguatamente i "fatti"; essa sarà dunque percepita come *controfattuale*. Seguirà poi una fase in cui si sospenderà il giudizio su quale sia la teoria più adeguata ed essa sarà dunque identificata come *supposizionale* e, se alla fine sarà accettata, il suo status diventerà in un certo senso *fattuale*.

conseguenza ed elaborando una nuova prospettiva sui fenomeni, selezionando nuovi obiettivi per la ricerca;

– se la teoria dominante continua a essere quella che mostra il miglior rapporto problemi risolti/anomalie irrisolte, la comunità persevera nel tentativo di raggiungere gli obiettivi da essa posti e cerca di elaborare dei nuovi esperimenti che possano corroborare la teoria.

8.1.4 Il ragionamento controfattuale nella ricerca scientifica

Analizziamo ora l'applicazione del ragionamento controfattuale a un particolare tipo di razionalità, quella impiegata in ambito scientifico.

Il fatto che si possa parlare di ragionamento controfattuale in ambito di scienza è cosa perlomeno opinabile, soprattutto alla luce di tutta la produzione epistemologica dell'ultimo secolo, che ha messo seriamente in crisi l'idea che la scienza abbia a che fare con i *fatti*. Nell'ottica delle correnti antinduttivistiche che hanno preso le mosse dai lavori di Popper, Kuhn, Lakatos, Feyerabend, Hanson e altri, l'esistenza stessa dei "puri fatti", svincolati da qualsiasi teoria, risulta pressoché insostenibile.

Ricordiamo le parole di Lakatos in [94]:

> The proposition 'the Proutian programme was carried through' looks like a 'factual' proposition. But there are no 'factual' propositions: the phrase only came into ordinary language from dogmatic empiricism. *Scientific 'factual' propositions* are theory-laden: the theories involved are methodological theories[10].
>
> [History of science and its rational reconstruction, p.119, nota 1]

[10]La proposizione "il programma [di Newton] venne condotto in porto" assomiglia a una proposizione "fattuale". Ma non esistono proposizioni "fattuali": questa espressione è stata indotta nel linguaggio ordinario dall'empirismo dogmatico. Le proposizioni "fattuali" scientifiche sono cariche di teoria: le teorie coinvolte sono "teorie osservative". [*tr. it. di Marcello D'Agostino in*: [96], pp.151–152, nota 61]

Anche Baas Van Fraassen in [60] delinea due posizioni generali in filosofia della scienza che, pur differenziandosi massimamente nei confronti di quello che ritengono essere lo scopo della scienza, in un certo senso concordano nello spogliare i fatti di quell'alone di sacralità di cui il realismo scientifico ingenuo li aveva rivestiti.

La prima posizione che descrive è quella del realismo scientifico che potremmo chiamare "sofisticato":

> Science aims to give us, in its theories, a literally true story of what the world is like; and acceptance of a scientific theory involves the belief that it is true[11].
>
> [*The Scientific Image*, p.8]

La seconda è quella dell'*empirismo costruttivo*:

> Science aims to give us theories which are empirically adequate; and acceptance of a theory involves as belief only that it is empirically adequate. [...] a theory is empirically adequate exactly if what it says about the observable things and events in this world, is true – exactly if it 'saves the phenomena'[12].
>
> [*The Scientific Image*, p.12]

Al di là di quello che può essere il rapporto tra la credenza nella verità della teoria e la verità stessa, oppure tra la verità dei fenomeni osservati e la verità stessa, in un caso accettare una teoria significa *credere* che sia vera e nell'altro significa ritenere che questa descriva adeguatamente i fenomeni. In entrambi i casi, la verità – e quindi i fatti – resta al massimo un ideale regolativo e

[11] La scienza ha lo scopo di fornirci, nelle sue teorie, una storia letteralmente vera di come va il mondo; e l'accettazione di una teoria scientifica implica la credenza che questa sia vera. [*traduzione mia*]

[12] La scienza ha lo scopo di fornirci teorie che siano empiricamente adeguate; e l'accettazione di una teoria implica come credenza solo che questa sia empiricamente adeguata. [...] una teoria è empiricamente adeguata esattamente se ciò che dice delle cose osservabili e degli eventi di questo mondo è vero – esattamente se 'salva i fenomeni'. [*traduzione mia*]

nessuna teoria scientifica può quindi avanzare la pretesa di parlare dei "puri fatti".

Per prima cosa va dunque precisato che, prendendo noi le mosse dalla tradizione antinduttivistica e dal costruttivismo di Van Fraassen, quando parliamo di ragionamento controfattuale con riferimento a una teoria scientifica, in realtà non stiamo parlando di un ragionamento che parte da premesse che negano dei fatti che sono veri *nella realtà*, ma parliamo di un ragionamento che prende le mosse dalla negazione di asserti che sono veri o validi *all'interno della teoria* stessa.

Per questo motivo, quando si ragiona all'interno di un paradigma teorico (*à la* Kuhn), le teorie alternative sono, in questa accezione debole, controfattuali.

Sotto questa prospettiva, come fanno notare Tetlock e Belkin in [156], il confine tra fattuale e controfattuale diventa molto labile:

> As a result of this vigorous research program, many scientists argue that a once highly speculative counterfactual conjecture is now better viewed as a quite-probable fact of natural history – yet another illustration of how blurry the boundary between factual and counterfactual can be[13].
>
> [Counterfactual Thought Experiments in World Politics]

In altri termini, ogni dato periodo storico ha una teoria scientifica dominante che è trattata come se presentasse i "fatti" e tutti gli asserti che non rientrano in essa sono trattati come falsi ed eventualmente come potenziali ipotesi controfattuali. Se però la comunità scientifica decide di abbandonare tale teoria per abbracciarne una alternativa nella quale gli asserti di cui sopra possono essere accomodati, la vecchia teoria diventa a sua volta un'alternativa controfattuale alla nuova teoria fattuale nella quale gli asserti che prima erano considerati falsi trovano una sistemazione.

[13]Come risultato di questo vigoroso programma di ricerca, molti scienziati sostengono che quella che una volta era una congettura controfattuale altamente speculativa è oggi vista piuttosto come un fatto abbastanza probabile della storia naturale – ancora un'altra dimostrazione di quanto possa essere indistinto il confine tra fattuale e controfattuale. [*traduzione mia*]

Fatte tutte queste premesse, quello che si vuole ora sostenere è che, in accordo con quanto già rilevato sia da Kuhn che da Popper, una teoria scientifica "dominante" non viene mai abbandonata dalla comunità scientifica se non c'è già a disposizione una teoria alternativa.

È quindi proprio la presenza di teorie alternative che possono essere messe a confronto con la teoria comunemente accettata che permette quello che si è soliti chiamare il *progresso* della scienza.

Popper in [133] afferma:

> Quindi, possiamo anche caratterizzare una *teoria sottoposta a indagine* come quella parte di un vasto sistema, per la quale abbiamo in mente, sia pure vagamente, un'alternativa, e per la quale cerchiamo di elaborare dei controlli cruciali.
>
> [*Congetture e Confutazioni*, p.194]

Anche Kuhn in [91] similmente:

> [...] once it has achieved the status of paradigm, a scientific theory is declared invalid only if an alternate candidate is available to take its place [14].
>
> [*The Structure of Scientific Revolutions*, p.77]

Razionalità strumentale

Secondo questi punti di vista, lo scenario che si presenta agli scienziati che si trovano a operare in un periodo di crisi della scienza "ufficiale", che prelude a una cosiddetta rivoluzione scientifica, è composto da una teoria "normale" (il cui nucleo fino a quel momento era stato trattato *come se* fosse aproblematico), da una teoria alternativa e da alcune osservazioni "critiche" (nel senso che rappresentano delle anomalie che devono essere risolte dalla teoria che supererà la prova).

---

[14] [...] una volta che ha acquisito lo status di paradigma, una teoria scientifica è dichiarata non valida solo se un candidato alternativo è disponibile a prendere il suo posto. [*traduzione mia*]

L'interpretazione che si vorrebbe suggerire è che le osservazioni empiriche, nei periodi che Kuhn definisce di scienza normale, vengono studiate e rielaborate al fine dell'accettazione della teoria che ha raggiunto un certo grado di corroborazione ed è assunta come dominante, ovvero la più adeguata alla spiegazione dei fenomeni.[15]

Così Kuhn in [91]:

> We have already seen, however, that one of the things a scientific community acquires with a paradigm is a criterion for choosing problems that, while the paradigm is taken for granted, can be assumed to have solutions [16].
>
> [*The Structure of Scientific Revolutions*, p.37]

Il tipo di ragionamento dominante nei periodi di scienza normale è dunque quello che abbiamo precedentemente definito strumentale, che può essere esemplificato dal seguente schema:

$$\text{Se l'esperimento } E \text{ avesse dato il risultato } R, \text{ allora } T \text{ sarebbe stata confermata (o refutata)} \tag{8.1}$$

consistente nella valutazione delle risorse sperimentali disponibili per la conferma o falsificazione di una teoria.

Ex-post

Viceversa, nei periodi di crisi che precedono ogni rivoluzione e il conseguente cambiamento di paradigma, il tipo di ragionamento utilizzato assomiglia a

[15] Parliamo in questo caso genericamente di corroborazione, poiché, pur riconoscendo l'importanza di distinguere tra modelli verificazionisti o falsificazionisti, in questa sede ci stiamo muovendo a un livello differente, per cui l'esperimento ha valore strumentale sia che abbia come fine la verificazione, sia la falsificazione della teoria.

[16] Abbiamo già visto, tuttavia, che una delle cose che una comunità scientifica acquisisce insieme a un paradigma è un criterio per scegliere problemi che, mentre il paradigma è dato per scontato, possono essere considerati dotati di soluzione. (*traduzione mia*)

quello *ex-post*: viene mantenuto fisso l'apparato sperimentale, mentre il nucleo teorico viene variato. In questo modo la comunità scientifica *crea* un nuovo paradigma teorico.

A questo punto è interessante mettere in luce un'analogia tra il generico meccanismo di valutazione che sottende le decisioni degli agenti razionali e lo specifico meccanismo di valutazione che, nella prospettiva di Kuhn, starebbe alla base della scelta tra due teorie rivali che la comunità scientifica si trova a fronteggiare.

Ricordiamo che, nel caso generico, l'agente razionale di fronte a una scelta doveva compiere un calcolo sul rapporto tra il beneficio atteso dal raggiungimento di un obiettivo "di partenza" e l'effetto dei costi affondati (cioè la perdita derivante dal non utilizzo di altri mezzi che si avevano a disposizione). Quello che succedeva di norma era che il beneficio atteso avesse un valore sufficientemente alto da ignorare l'effetto dei costi affondati, per cui l'obiettivo veniva mantenuto fisso ed eventualmente l'agente ricercava nuovi mezzi per raggiungerlo. D'altro canto, se veniva superata una certa soglia, questo effetto poteva indurre l'agente a decidere di riutilizzare i mezzi (o risorse) e a porsi quindi un nuovo obiettivo.

Analogamente, una comunità scientifica si trova sempre di fronte a delle anomalie sperimentali che deve decidere se trattare come dei rompicapo che la teoria non ha *ancora* risolto, oppure come dei controesempi che invalidano la teoria. Durante i periodi di scienza normale, le anomalie vengono viste come rompicapo e la comunità scientifica si impegna a risolverle *nell'ambito* della teoria, compiendo nuove misurazioni, modificando gli strumenti di misura ecc. Quando però si supera una certa soglia e le anomalie diventano troppo importanti, la comunità scientifica conferisce lo status di controesempi alle anomalie e abbandona la teoria che non è in grado di spiegarle e crea una nuova teoria per la quale le evidenze osservative (che prima erano viste come anomalie) siano un idoneo mezzo di dimostrazione.

Nelle parole di Kuhn ([91]):

> Sometimes a normal problem, one that ought to be solvable by known rules and procedures, resists the reiterated onslaught of the ablest members of the group within whose competence it falls.

> On other occasions a piece of equipment designed and constructed for the purpose of normal research fails to perform in the anticipated manner, revealing an anomaly that cannot, despite repeated effort, be aligned with professional expectation. In these and other ways besides, normal science repeatedly goes astray. And when it does – when, that is, the profession can no longer evade anomalies that subvert the existing tradition of scientific practice – then begin the extraordinary investigations that lead the profession at last to a new set of commitments, a new basis for the practice of science. The extraordinary episodes in which that shift of professional commitments occurs are the ones known in this essay as scientific revolutions[17].
>
> [*The Structure of Scientific Revolutions*, p.6]

Questo processo di valutazione è stato presentato a partire dalla prospettiva kuhniana per motivi di semplicità, ma può essere ugualmente rappresentato all'interno della trattazione più sofisticata che Lakatos fa dei programmi di ricerca progressivi e regressivi: quando la comunità scientifica giudica un programma di ricerca come progressivo, le anomalie vengono relegate alla cosiddetta *cintura protettiva*, cioè a quell'insieme di ipotesi ausiliarie, teorie osservative ecc. che possono essere modificate e accomodate affinché possa essere mantenuto il nucleo del programma di ricerca.

In [97] si legge:

---

[17] A volte un problema normale, uno che dovrebbe essere risolvibile con regole e procedure note, resiste i reiterati assalti dei membri di un gruppo all'interno della cui competenza cade. In altre occasioni un pezzo di equipaggiamento disegnato e costruito per lo scopo della ricerca normale non si comporta nel modo atteso rivelando un'anomalia che non può, a dispetto dei ripetuti sforzi, essere allineata all'aspettativa dei professionisti. In questi e altri modi, la scienza normale si smarrisce. E quando ciò accade – quando, cioè, i professionisti non possono più ignorare le anomalie che sovvertono la tradizione esistente della pratica scientifica – allora cominciano le investigazioni straordinarie che portano i professionisti alla fine a un nuovo insieme di impegni, a una nuova base per la pratica della scienza. Gli episodi straordinari nei quali si verifica quella trasformazione degli impegni professionali sono quelli noti in questo saggio come rivoluzioni scientifiche. [*traduzione mia*]

> [...] the positive heuristics consists of a partially articulated set of suggestions or hints on how to change, develop the 'refutable variants' of the research-programme, how to modify, sofisticate, the 'refutable' protective belt[18].
>
> [Falsification and the methodology of scientific research programmes, p.50]

Quando però il programma non è più in grado di superare le difficoltà con questi aggiustamenti, è il nucleo stesso del programma di ricerca a essere scardinato; quest'ultimo viene dichiarato regressivo e abbandonato in favore di un nuovo programma di ricerca, che si dimostri "più progressivo".

Lakatos, a differenza di Kuhn, sempre in [97], fissa un criterio per distinguere un programma di ricerca progressivo da uno regressivo:

> [...] a scientific theory T is *falsified* if and only if another theory T' has been proposed with the following characteristics: *(I)* T' has excess empirical content over T: that is, it predicts *novel* facts, that is, facts improbable in the light of, or even forbidden, by T; *(2)* T' explains the previous success of T, that is, all the unrefuted content of T is included (within the limits of observational error) in the content of T'; and *(3)* some of the excess content of T' is corroborated[19].
>
> [Falsification and the methodology of scientific research programmes, p.32]

---

[18] [...] l'euristica positiva consiste di un insieme parzialmente espresso di proposte e suggerimenti su come cambiare e sviluppare le "varianti confutabili" del programma di ricerca, su come modificare e complicare la cintura protettiva "confutabile". [*tr. it. di Marcello D'Agostino in*: [95], p.64]

[19] [...] una teoria scientifica T è *falsificata* se e solo se è stata proposta un'altra teoria T' con le seguenti caratteristiche: *(I)* T' ha un contenuto empirico addizionale rispetto a T: cioè essa predice fatti *nuovi*, ossia fatti improbabili alla luce di T o addirittura vietati da quest'ultima; *(2)* T' spiega il precedente successo di T, cioè, tutto il contenuto non confutato di T è incluso (entro i limiti dell'errore osservativo) nel contenuto di T'; e *(3)* parte del contenuto addizionale di T' è corroborato. [*tr. it. di Marcello D'Agostino in*: [95], p.42]

Tuttavia, Paul Feyerabend in [58] critica questo criterio affermando che esso è vano se non fissa dei limiti di tempo oltre i quali un programma che non manifesti la propria "progressività" debba essere dichiarato regressivo.

A nostro avviso una discriminante di tipo temporale risulterebbe troppo rigida e per tale motivo vorremmo proporre un'interpretazione che trae ispirazione da quanto detto finora relativamente alla scelta di un agente se utilizzare la razionalità strumentale o quella *ex-post*. Nella pratica scientifica, come in tutti gli altri processi di ragionamento, deve esistere una soglia in corrispondenza della quale i costi affondati superano i benefici attesi dalla teoria[20] e rendono quindi *più razionale* cambiare il paradigma teorico piuttosto che perseverare nel riaggiustamento delle anomalie.

È evidente che determinare il valore di questa soglia non è impresa facile, ma almeno questo valore è funzione di una serie di parametri e non unicamente del tempo.

A questo punto, quando davanti al tribunale della comunità scientifica si presentano due teorie alternative, che si affrontano sul campo della concordanza con i risultati osservativi, una delle due, quella sostenuta dal paradigma fino a quel momento dominante, verrà trattata *come se* fosse vera e contenente asserzioni "fattuali", mentre la teoria "sfidante", prima di passare il vaglio dell'esperienza, viene considerata falsa e contenente asserzioni "controfattuali".

Il processo di ragionamento muove dall'assunto che, partendo dalla teoria consolidata, i risultati osservativi non sono giustificati; allora pone l'ipotesi (controfattuale) che valga invece la nuova teoria. Vengono quindi tratte delle conseguenze *nel contesto controfattuale* della nuova teoria scientifica e si verifica se, alla luce di questo ragionamento e all'interno di questo contesto, gli stessi risultati osservativi sono giustificati. Nel caso in cui lo siano, la nuova teoria viene assunta a spiegazione dei fenomeni e passa dallo status di controfattuale a quello di fattuale, nel senso "indebolito" illustrato sopra.

Prendiamo il celeberrimo esempio, tratto dalla cosmologia, dello sposta-

[20]Ovviamente, come già segnalato in precedenza, quando si ha a che fare con teorie scientifiche i costi affondati e i benefici attesi sono da intendersi in senso molto più ampio (che include anche i costi cognitivi) e non limitato al valore economico.

mento delle righe spettrali verso il rosso, che rappresentava un'anomalia per la teoria della gravitazione universale newtoniana e al quale venne data una spiegazione all'interno della teoria della relatività generale di Einstein.

Il fenomeno anomalo era l'osservazione della deflessione verso il rosso delle righe spettrali dell'emissione di determinati corpi luminosi, nella fattispecie delle cosiddette "stelle fisse".

Dato che la teoria della gravitazione universale non poteva fornire nessuna spiegazione per questo fenomeno, la comunità scientifica si sarà presumibilmente trovata a un certo punto a formulare un ragionamento (in quel momento storico) controfattuale, del tipo:

> Se la frequenza della luce emessa da un atomo dipendesse dal campo gravitazionale, allora le righe spettrali delle stelle sarebbero spostate verso il rosso (8.2)

Asserto che non era giustificabile a partire dalla teoria della gravitazione newtoniana, ma trovava spiegazione all'interno della teoria della relatività generale einsteiniana. Una volta passata al vaglio e accettata questa conclusione, la teoria della relatività generale ha potuto diventare "fattuale"[21].

Le teorie scientifiche, come ogni altro processo di ragionamento, possono essere sottoposte a riconsiderazione attraverso il ragionamento controfattuale e, in particolare, nei periodi che preludono a uno slittamento di paradigma, sono le teorie a essere sottoposte a critica, più che le osservazioni, e il particolare procedimento adottato è quello del ragionamento controfattuale *ex-post*.

Lo schema del ragionamento controfattuale *ex-post* in ambito scientifico è dunque il seguente:

> Se ragionassimo all'interno della teoria T' (invece che all'interno di T), allora potremmo spiegare l'evidenza sperimentale E. (8.3)

consistente appunto nel valutare quale sia la teoria che riesce a sfruttare al meglio l'evidenza sperimentale accumulata.

[21] Ovviamente, questo è un modo molto semplicistico di trattare la questione. In realtà non è mai una sola osservazione isolata a inficiare una teoria, ma sono necessarie numerose e significative anomalie.

La proposta appena illustrata relativa a un'interpretazione delle modalità attraverso le quali può manifestarsi la razionalità scientifica resta per il momento solo un vago suggerimento di una chiave di lettura per l'epistemologia. Restano ancora da precisare i parametri che entrano nel calcolo del rapporto costi-benefici nell'impresa scientifica, precisazione che è fondamentale per capire le modalità del passaggio da una forma di razionalità strumentale a una *ex-post*.

Questo compito, insieme a quello di rappresentare un esempio di processo della razionalità scientifica attraverso il formalismo illustrato nella prima parte della tesi, è una delle linee di sviluppo possibili di questo lavoro.

8.2 Razionalità e controfattuale per agenti artificiali intelligenti

In questa sezione ci proponiamo di analizzare l'uso delle due forme di razionalità presentate nei capitoli 6 e 7 negli studi sugli agenti artificiali intelligenti, per vedere quanto sia già stato fatto e quanto resti ancora da fare. Inoltre, cercheremo di mostrare che il possesso di entrambe queste forme di razionalità e la capacità di passare dall'una all'altra – grazie anche all'utilizzo del ragionamento controfattuale – aumentano l'autonomia degli agenti.

8.2.1 Modelli di razionalità strumentale nell'intelligenza artificiale

Gli agenti artificiali intelligenti sono una delle più grandi scoperte dell'informatica dell'ultima decade. Si tratta di programmi di computer che posseggono una certa dose di autonomia e sono in grado di agire in maniera indipendente in ambienti dinamici e imprevedibili.

Uno degli scopi degli studi in intelligenza artificiale è quello di simulare – o, più precisamente – emulare i processi di ragionamento umani attraverso delle macchine, affinché queste ultime siano in grado di assistere, e in certi casi sostituire, gli esseri umani nello svolgimento di alcuni compiti. Per esempio, aree di applicazione in cui il paradigma degli agenti si è dimostra-

to particolarmente interessante sono il commercio elettronico su Internet, il controllo di prototipi utilizzati nelle missioni astronautiche, la progettazione di interfacce facilmente utilizzabili dagli utenti, il controllo dei processi industriali, solo per citarne alcune.

Uno dei modelli di agente artificiale più conosciuti e studiati è il modello BDI (da *belief, desire, intention*, ossia credenza, desiderio, intenzione). Il motivo del suo successo è probabilmente da ascrivere alla solidità del modello filosofico che ne è alla base (sviluppato a partire da idee originariamente portate avanti da Michael Bratman), unita al successo di molte delle sue applicazioni.

Paradigmi alternativi al BDI sono, come indicato da Michael Wooldridge [163], le architetture logiche, quelle reattive (come quelle descritte da Rodney Brooks in [26], [27] e [28]) e gli agenti ibridi delle architetture stratificate. Per lo scopo che ci proponiamo in queste sezioni, ovvero quello di analizzare il tipo di razionalità sottostante, il BDI fungerà da modello di riferimento, essendo quello forse più interessante dal punto di vista filosofico poiché in esso vengono maggiormente esplicitati i processi razionali e non presentando gli altri modelli – sotto il preciso e specifico rispetto della rappresentazione della razionalità – differenze sostanziali con il BDI.

Il modello BDI è basato su queste tre fondamentali componenti:

- **credenze**: rappresentano la conoscenza che l'agente ha sul mondo;
- **desideri**: rappresentano lo stato a cui l'agente aspira, quindi in senso lato il suo obiettivo;
- **intenzioni**: rappresentano la persistenza nel perseguire un obiettivo.

Le intenzioni, in particolare, giocano un ruolo abbastanza importante, poiché sono la componente che guida gli agenti all'azione e la scelta relativa a quando continuare a perseguire un'intenzione o quando abbandonarla determina le strategie che l'agente seguirà nei suoi comportamenti.

Per esempio, Philip Cohen e Hector Levesque in [36] stabiliscono che un agente abbandona un'intenzione quando questa è stata soddisfatta, ossia quando l'obiettivo è stato raggiunto, oppure quando questo è diventato irraggiungibile, ovvero quando le credenze dell'agente sono cambiate.

Molta attenzione viene prestata alla capacità che gli agenti devono possedere di scomporre i piani in sotto-piani e quindi di costruire delle gerarchie di obiettivi e sotto-obiettivi. Questa capacità è in parte traducibile nei processi della razionalità strumentale che, una volta deciso un obiettivo da perseguire, hanno il compito di identificare una strategia per raggiungerlo e di modificarla nel caso l'agente giudichi insoddisfacente il suo esito, ricercando di volta in volta dei mezzi più idonei allo scopo.

Di conseguenza, i processi della razionalità strumentale sono rappresentati in maniera esauriente nei modelli BDI; tuttavia, abbandonare un obiettivo quando esso viene raggiunto o quando viene percepito come non più raggiungibile può non essere sufficiente: esistono dei casi in cui non è lo status "assoluto" dell'obiettivo a essere determinante, ma quello "relativo all'agente", ossia la sua preferibilità; può accadere che l'agente, in base a cambiamenti percepiti nell'ambiente o inferiti attraverso il ragionamento, si renda conto di propendere per un altro obiettivo alternativo.

Sembra insomma mancare un meccanismo che giustifichi la perseveranza nell'inseguire un obiettivo anche sulla base della costante preferibilità rispetto ad altri obiettivi alternativi e ugualmente raggiungibili.

8.2.2 Razionalità *ex-post* per agenti artificiali

Il secondo tipo di razionalità da noi preso in considerazione è stato affrontato in maniera sicuramente più limitata rispetto a quanto sia stato fatto per la più comune nozione di razionalità strumentale.

Tuttavia, già in uno dei primi articoli sui fondamenti dei modelli BDI di Anand Rao e Michael Georgeff [135], veniva definita "interessante" una strategia basata su quella che noi abbiamo chiamato razionalità *ex-post*.

Dopo aver definito tre tipi possibili di atteggiamento che un agente può intrattenere nei confronti di un'intenzione, definiscono sulla base della combinazione di due di questi atteggiamenti una prospettiva che potrebbe essere interessante far assumere agli agenti.

Questi tre possibili tipi di impegno verso il raggiungimento di un fine sono:

- impegno **cieco**: le intenzioni vengono mantenute finché l'agente non riconosce di averle *effettivamente* realizzate;
- impegno **univoco**: le intenzioni vengono mantenute finché l'agente ritiene che siano realizzabili;
- impegno **aperto**: le intenzioni vengono mantenute finché si applicano a obiettivi effettivi per l'agente.

In [135] si legge:

> [...] a particularly interesting commitment strategy is one in which the agent is open-minded with respect to ends but single-minded with respect to the means towards those ends. Such an agent is free to change the ends to which she aspires, but once committed to a means for realizing those ends, will not reconsider those means[22].
>
> [Modeling Rational Agents within a BDI-Architecture, p.482]

La descrizione di questa strategia corrisponde esattamente a quanto detto della razionalità retrospettiva, ossia della capacità degli agenti di impegnarsi a mantenere invariati i mezzi e sottoporre a revisione la preferibilità dei fini e, di conseguenza, l'intenzione di realizzarli.

Sebbene venga fatto questo rilievo, in seguito alla questione non viene più dedicata molta attenzione, né viene spiegato più precisamente quando e perché un obiettivo viene abbandonato nei casi in cui un agente mostra un impegno di tipo "aperto". Molta più importanza è stata attribuita alla scomposizione dei piani e all'identificazione di sotto-obiettivi che disegnano i possibili percorsi che portano gli agenti verso il loro obiettivo.

Da questi presupposti consegue una nozione ancora piuttosto limitata di agente autonomo come agente in grado di scegliere da sé la particolare strategia per raggiungere un fine, ossia quella che Castelfranchi in [31] definisce

[22] [...] una strategia di impegno particolarmente interessante è una in cui l'agente è aperto rispetto ai fini ma univoco rispetto ai mezzi rivolti a quei fini. Tale agente è libero di cambiare i fini ai quali aspira ma, una volta stabiliti i mezzi per realizzare quei fini, non riconsidererà quei mezzi. [*traduzione mia*]

autonomia esecutiva (*executive autonomy*). A questa Castelfranchi contrappone l'autonomia negli obiettivi (*goal autonomy*), ossia la capacità che gli agenti autonomi dovrebbero possedere, di generare endogenamente, di volta in volta, degli obiettivi. Le preferenze dell'agente non sarebbero più dunque imposte dall'esterno (dal costruttore una volta per tutte o dagli utenti continuativamente), ma sarebbero un ordinamento provvisorio che varia al variare dell'ambiente esterno e degli stati interni dell'agente stesso.

Se è pur vero che in molti casi l'utente umano ha tutto l'interesse di mantenere un certo livello di controllo sull'agente, decidendo dall'esterno l'obiettivo dei suoi sforzi, in altri casi può essere comodo che l'agente si muova con un maggior grado di autonomia, evitando all'utente di dover continuamente intervenire o osservare il suo operato.

Resta però un problema aperto capire come l'agente possa decidere in quali casi adottare un processo di razionalità strumentale e quando un processo retrospettivo. La sezione 8.2.3 illustra in linee molto generali una possibile proposta.

8.2.3 Agenti artificiali autenticamente autonomi

Immaginiamo ora di avere un agente capace di utilizzare entrambi i tipi di razionalità, cioè capace sia di mantenere la sua intenzione di raggiungere il fine e quindi perseverare nello sforzo di raggiungerlo cercando nuovi mezzi da impiegare in questa ricerca, ma capace anche al tempo stesso di abbandonare un obiettivo quando si verificano alcune precise circostanze.

Il punto era già stato messo in luce fin dall'inizio da Cohen e Levesque in [36]:

> An autonomous agent should act on its intentions, not in spite of them; adopt intentions it believes are feasible and forego those believed to be unfeasible; keep (or commit to) intentions, but not forever; discharge those intentions believed to have been satisfied; alter intentions when relevant beliefs change; and adopt subsidiary intentions during plan formation[23].

[23]Un agente autonomo dovrebbe agire secondo le sue intenzioni, non nonostante queste;

[Intention Is Choice with Commitment, p.214]

Ma come può un agente decidere quale dei due tipi di razionalità è più appropriato in un caso specifico? A nostro avviso tale agente dovrebbe possedere un metodo generale di scelta da applicare ai casi particolari.

Nel ricercare questo metodo possiamo prendere le mosse dall'osservazione dei metodi di valutazione dei decisori umani. Una prima osservazione da fare è di carattere molto generale: quando un agente persevera nella ricerca di un obiettivo, significa che lo considera – per così dire – "in cima alla sua scala delle preferenze", metre quando lo abbandona in favore di un nuovo obiettivo, normalmente ciò significa che il nuovo obiettivo è ora preferibile rispetto al vecchio[24].

La nozione di preferibilità è una nozione relazionale, quindi se la perseveranza o l'abbandono di un obiettivo dipendono dalla sua preferibilità, questa andrà definita in relazione ad altri obiettivi considerati (o che avrebbero dovuto essere considerati) al momento della scelta.

Se accettiamo l'idea che sia possibile imparare dal passato, gli agenti dovrebbero essere in grado di ragionare a posteriori su una scelta fatta (di perseguire un obiettivo o di perseguirlo in un determinato modo) e di confrontare il piano e il risultato ottenuto con altri piani che erano anch'essi plausibili al momento della decisione e i loro rispettivi – ipotetici – risultati attesi.

Questo genere di riconsiderazioni sono l'oggetto del ragionamento controfattuale, secondo quanto spiegato diffusamente nei capitoli 6 e 7. In altri termini, il ragionamento controfattuale dovrebbe permettere a un agente artificiale:

---

adottare intenzioni che crede realizzabili e abbandonare quelle ritenute irrealizzabili; mantenere le intenzioni (o impegnarsi in esse), ma non per sempre; affrancarsi da quelle intenzioni che ritiene siano state soddisfatte; alterare le intenzioni quando le credenze rilevanti cambiano; e adottare intenzioni sussidiarie durante la formazione dei piani. [*traduzione mia*]

[24]Tralasciamo in questa analisi le situazioni in cui l'obiettivo sia stato raggiunto o sia diventato irraggiungibile poiché esse comportano automaticamente il cambio dell'obiettivo, come già fatto rilevare correttamente dai sostenitori del modello BDI.

- di confrontare il piano scelto con piani in cui i mezzi a disposizione erano differenti, ma l'obiettivo era lo stesso (razionalità strumentale)

  Se avessi usato il mezzo B invece del mezzo A avrei ottenuto l'obiettivo x (8.4)

- di confrontare il piano scelto con piani in cui l'obiettivo finale è differente, ma i mezzi a disposizione sono gli stessi (razionalità *ex-post*)

  Se avessi tentato di ottenere l'obiettivo y invece dell'obiettivo x, con i mezzi A l'avrei raggiunto (8.5)

Questo processo può essere rappresentato nel formalismo illustrato nel capitolo 4 nel modo seguente:

- l'agente si focalizza sul piano che ha portato a termine e costruisce un contesto fattuale contenente tutta l'informazione rilevante;
- elabora l'ipotesi controfattuale e costruisce il contesto controfattuale in cui almeno una condizione rilevante è stata cambiata;
- controlla l'esito del piano rivisto svolgendo il processo inferenziale interno al contesto controfattuale;
- confronta l'esito dei due piani e di conseguenza costruisce il nuovo contesto per il prossimo piano.

L'esito di questo tipo di ragionamento controfattuale determina, almeno in parte, se l'agente persevererà nel suo piano o se lo rivedrà acquistando nuovi mezzi o modificando le sue preferenze e quindi scegliendo un obiettivo diverso; tuttavia, questo ragionamento da solo non è sufficiente e per due motivi.

In primo luogo, è possibile che esistano molti piani plausibili o obiettivi desiderabili per un agente che si accinge a compiere una scelta e deve esistere dunque un ulteriore modo per distinguere il migliore tra essi; secondariamente, il ragionamento controfattuale ancora non dice quando la riconsiderazione debba applicarsi ai mezzi e quando alle preferenze.

Lo strumento che, a nostro parere, può svolgere questi due compiti è il calcolo dei costi e benefici (reali nel caso del piano effettivamente messo

in atto dall'agente e attesi nel caso dei piani "controfattuali"), che tenga conto dell'effetto dei costi affondati (illustrato nella sezione 7.1), la cui funzione è precisamente quella di integrare l'influenza che la valutazione di piani alternativi può avere sulla valutazione del piano messo in atto.

In altre parole, il valore di un obiettivo (e quindi il suo posizionamento nella scala delle preferenze) viene calcolato togliendo al beneficio ricavato dal raggiungimento del fine il costo dei mezzi effettivamente impiegati e il costo derivante dal non utilizzo di mezzi che erano disponibili e che avrebbero potuto essere utilizzati in un piano alternativo (l'effetto costi affondati). Ricordiamo che utilizzando un mezzo se ne "ammortizza" il costo e quindi un mezzo acquistato che giace inutilizzato costituisce un costo[25].

Anche queste idee sono state presentate solo sotto forma di intuizioni e sarebbe certamente interessante sistematizzarle in un modello più rigoroso per verificare se possano effettivamente essere implementate in un agente artificiale e fornire dei vantaggi concreti.

8.3 Gli scenari multiagente

Quanto finora detto sugli agenti artificiali è ancora incompleto, poiché è stato tralasciato un aspetto molto importante, ovvero che difficilmente un agente si trova a operare in ambienti isolati, dove non esistono altre entità "intelligenti"; la maggior parte delle volte molteplici agenti condividono un ambiente e si ritrovano quindi a dover interagire tra di loro, oltreché con gli utenti umani.

Questa interazione fa nascere delle problematiche nuove rispetto a quanto detto finora, problematiche legate al coordinamento, che deve essere studiato in maniera tale da massimizzare l'efficacia delle azioni compiute dagli agenti.

Un'altra problematica molto importante che emerge non appena ci si sposta dal livello del singolo agente a quello del multiagente è quella della comunicazione, poiché dal momento che gli agenti condividono uno spazio e che risulta quindi necessario trovare dei meccanismi di coordinamen-

[25] Un'analisi più dettagliata di questo argomento è contenuta in [17]

to, la comunicazione è probabilmente il mezzo più efficace in favore del coordinamento.

Per quanto riguarda la comunicazione, il problema al quale è stata rivolta la massima attenzione da parte degli studiosi di intelligenza artificiale è quello del linguaggio, ovvero quale tipo di linguaggio debbano parlare gli agenti, costruito secondo quale sintassi e, una volta che costruttori diversi abbiano compiuto scelte diverse relativamente al linguaggio da far parlare agli agenti, creare dei protocolli di traduzione o di negoziazione del significato per far sì che anche agenti dotati di linguaggi diversi possano comunicare.

Tuttavia, anche se questo ordine di problemi venisse risolto, gli agenti dovrebbero essere in grado di decidere come rendere più efficace la comunicazione anche a livello di contenuto: un agente che volesse ottenere della collaborazione da parte di un altro agente dovrebbe poter prevedere cosa sia meglio dire e cosa omettere in ogni singolo caso.

Per fare questo e per formulare previsioni di qualsiasi genere sul comportamento degli altri, gli agenti devono essere dotati della capacità di attribuire credenze, desideri, intenzioni ecc. agli altri agenti, ossia devono essere in grado di costruire una "teoria della mente" degli altri. Nel prossimo paragrafo mostreremo come, a nostro parere, un particolare tipo di ragionamento controfattuale, quello legato a quelli che Nelson Goodman in [78] aveva definito "controidentici" (*counteridenticals*), possa suggerire una possibile rappresentazione.

8.3.1 Il controfattuale di immedesimazione: "Se io fossi in te"

"Se io fossi in te farei A", "Al suo posto, io avrei fatto B", "Fosse successo a me, io avrei detto C"; queste e altre sono formule abbreviate che condensano un processo in realtà molto più articolato che parte dall'attribuzione di un certo stato cognitivo all'altro agente per giungere alla previsione dei suoi comportamenti futuri.

Anche nel caso degli esseri umani, nessun agente ha accesso diretto allo stato cognitivo degli altri e quindi ogni agente si ritrova a doversi costruire una teoria su "come l'altro ragiona"; tuttavia, spesso queste teorie sono

alquanto approssimate e sono costruite su dati frammentari. Tali dati comprendono le informazioni che l'altro agente fornisce spontaneamente su di sé e l'osservazione dei comportamenti. Il primo tipo di dati è influenzato dal possibile interesse che l'altro agente ricava a non essere sincero e può essere una fonte molto scarsa in presenza di agenti reticenti; il secondo tipo può avvalersi del confronto continuo con comportamenti messi in atto da agenti terzi e con le modalità di ragionamento che l'agente che sta compiendo l'indagine associa a quel tipo di comportamento. Lo stesso deve valere anche per gli agenti artificiali.

In altri termini, se un agente riscontra un'analogia tra il proprio comportamento e quello di un altro agente in una data circostanza, può ricavarne (a volte anche erroneamente) che l'altro agente metta in atto un processo di ragionamento simile al suo.

Questo genere di riflessioni permette agli agenti di prevedere il comportamento di altri agenti; immaginiamo che l'agente x nella situazione S abbia fatto A. L'agente y ne osserva il comportamento e cerca di ricavarne informazione che gli permette di prevedere come x si comporterà in una situazione S' che presenta dei tratti comuni rispetto a S. Potrebbero aversi due casi:

1. L'agente y pensa:

   - nella situazione S, x ha fatto A;
   - al suo posto, anch'io avrei fatto A;
   - nella situazione S' io farei B;
   - prevedo che anche x in S' farà B.

2. L'agente y pensa:

   - nella situazione S, x ha fatto A;
   - al suo posto, io avrei invece fatto B;
   - nella situazione S' io farei C;
   - prevedo che x in S' farà D.

Questo processo si applica in maniera abbastanza diretta anche al caso specifico della comunicazione: l'agente x, che vuole convincere l'agente y relativamente all'argomento A, si forma un'idea di quali siano le sue credenze, il suo carattere, i suoi pregiudizi ecc. sulla base di quanto y ha detto e fatto in passato e, prima di comunicare con lui pensa: "al posto di y quali argomenti troverei convincenti ed espressi in quale forma?".

A livello di rappresentazione, quanto detto sia relativamente alle azioni che relativamente alla comunicazione corrisponde a costruire un contesto fattuale con le proprie credenze e regole di inferenza relative a un determinato problema e parallelamente costruire un contesto controfattuale "di attribuzione" di uno stato cognitivo a un altro agente. Con delle opportune regole per trasferire informazione da un contesto all'altro, un agente dovrebbe essere in grado di fare dei pronostici sui comportamenti futuri degli altri agenti.

Prevenire il comportamento degli altri presenta degli indubbi vantaggi sia in scenari di cooperazione che in scenari di competizione, come vedremo meglio nei prossimi due paragrafi.

8.3.2 Il ragionamento controfattuale in situazioni di cooperazione

Molto spesso si possono verificare situazioni in cui, laddove un singolo agente è impossibilitato a portare a termine un compito, un gruppo di agenti può invece farlo abbastanza agevolmente e questo avviene sia per gli agenti umani che per quelli artificiali. Da ciò nasce la necessità di costruire agenti capaci di cooperare gli uni con gli altri.

Nei casi di cooperazione, il ragionamento controfattuale può tornare utile in due diverse circostanze: in primo luogo, il ragionamento controfattuale di immedesimazione può aiutare nella selezione dei candidati per la cooperazione; secondariamente, una semplice riconsiderazione controfattuale di un piano cooperativo può fornire una valutazione dell'opportunità della cooperazione.

Partiamo dal primo caso: quando un agente capisce che da solo non è in grado di portare a termine un compito, si pone alla ricerca di candidati adatti a cooperare con lui. Una volta individuati uno o più agenti che posseg-

gono le caratteristiche e le capacità richieste per la cooperazione, si procede a valutare l'interesse che tali agenti possono avere a collaborare. Se è immediatamente evidente che anche gli altri agenti hanno interessi a raggiungere l'obiettivo del piano cooperativo, il passo successivo è quello di formulare una richiesta di cooperazione utilizzando una tecnica di comunicazione il più possibile efficace, secondo quanto specificato nel paragrafo 8.3.1.

> Se fossi in lui, sarei interessato all'obiettivo della cooperazione e vorrei che mi venisse chiesto di cooperare. (8.6)

Se invece gli altri agenti non ricevono vantaggi diretti dal conseguimento dell'obiettivo, si può di nuovo ricorrere al ragionamento controfattuale di immedesimazione per cercare di individuare una "moneta di scambio" congrua per ricompensare tali agenti della loro collaborazione.

> Se fossi in lui, coopererei se in cambio mi venisse offerto x. (8.7)

Per quanto riguarda il secondo caso, di fronte a un piano cooperativo coronato dal successo, l'agente può riconsiderare se non avrebbe potuto portare a termine il compito anche da solo o rinunciando a qualcuno dei suoi collaboratori, oppure se cambiando qualcuno dei collaboratori o coinvolgendone degli altri non avrebbe potuto ottenere un risultato migliore; la riconsiderazione è dunque finalizzata all'aumento del beneficio netto ricavato dal piano.

> Se avessi portato avanti il piano da solo, avrei comunque raggiunto l'obiettivo. (8.8)

> Se avessi collaborato con x invece che con y, avrei raggiunto un risultato migliore. (8.9)

> Se avessi collaborato con x oltreché con y, avrei raggiunto un risultato migliore. (8.10)

Di fronte a un piano fallimentare, la riconsiderazione è finalizzata al raggiungimento dell'obiettivo nel futuro e quindi di nuovo si ipotizzeranno controfattualmente delle nuove coalizioni, nelle quali i collaboratori verranno cambiati, aggiunti o eliminati nel caso l'agente valuti come dannoso il loro contributo.

> Se non avessi chiesto a x di collaborare, avrei raggiunto l'obiettivo. (8.11)

Se avessi collaborato con x invece che con y, avrei raggiunto l'obiettivo. (8.12)

Se avessi collaborato con x oltreché con y, avrei raggiunto l'obiettivo. (8.13)

8.3.3 Il ragionamento controfattuale in scenari di competizione

Se le architetture in cui gli agenti cooperano sono le più diffuse e la loro utilità è più immediatamente evidente, non mancano gli scenari in cui gli agenti si trovino in posizione antagonistica e costretti a competere per portare a termine i loro compiti.

Anche in questo caso, come in quello della cooperazione, il ragionamento controfattuale agisce su due livelli: per prima cosa, con il ragionamento controfattuale di immedesimazione, l'agente cerca di capire le ragioni dell'opposizione degli altri agenti e di convincerli a desistere; poi la riconsiderazione controfattuale può aiutare l'agente a capire determinate mosse dell'avversario e a prevedere come si muoverà in futuro.

Cominciamo dalla situazione in cui l'agente può ancora tentare di eliminare l'ostacolo costituito dall'opposizione di altri agenti al suo piano. In primo luogo l'agente deve sincerarsi che l'opposizione da parte degli altri agenti sia reale e non frutto di un'incomprensione.

Io al suo posto quali motivi avrei per essere contro questo piano? (8.14)

In caso di incomprensione, l'agente procederà a fornire chiarimenti, altrimenti dovrà immaginare che cosa potrebbe fare per l'altro agente (o gli altri agenti) che per questo possa avere un valore superiore alla perdita che subisce lasciando che il primo agente porti a termine il suo piano.

Io al suo posto lascerei che il piano fosse portato a termine purché mi fosse dato x. (8.15)

Infine, ragionando sul comportamento passato degli avversari (cosa hanno fatto e cosa avrebbero invece potuto fare), gli agenti possono acquisire informazioni e formulare previsioni sulle azioni future degli avversari.

> Se in quella circostanza avesse fatto A avrebbe ottenuto x, ma non lo ha fatto, quindi forse non desidera ottenere x, quindi ora tra x e y dovrebbe scegliere y. (8.16)

Come si può facilmente intuire da questo schema, lo studio delle decisioni in ambienti caratterizzati da forte antagonismo può beneficiare del ragionamento controfattuale; di conseguenza, la teoria dei giochi, che si occupa proprio di formalizzare le scelte dei decisori in scenari di competizione, sembrerebbe un terreno di applicazione piuttosto promettente per il ragionamento controfattuale.

Esistono già una serie di lavori molto interessanti che si muovono in questa direzione; alcuni dei contributi più importanti sono quelli di Cristina Bicchieri ([15], [14], [13]) e Robert Stalnaker ([151]). Tuttavia, l'utilizzo di un nuovo formalismo basato sui contesti potrebbe portare il vantaggio di permettere di rappresentare scenari più complicati, come quelli di gruppi al cui interno si coopera che si oppongono ad altri gruppi ecc., oppure potrebbe essere utile per formalizzare da una parte la visione che un giocatore ha del gioco, accanto alla visione del gioco che questo giocatore attribuisce all'avversario, accanto alla visione che egli attribuisce all'avversario di quella che quest'ultimo pensa essere la sua visione e così via.

Infine, i contesti, in quanto oggetti parziali, potrebbero rivelarsi utili nel rappresentare le situazioni di gioco a informazione imperfetta, in cui la strategia prende le mosse non da una rappresentazione completa dello stato del gioco, ma piuttosto da un insieme parziale di possibili rappresentazioni. Analogamente, anche le strategie che non sono state messe in atto sono spesso caratterizzate da stati in cui l'informazione è incompleta e un contesto controfattuale potrebbe essere lo strumento formale con cui caratterizzarli e trarne delle informazioni.

Il livello di approfondimento al quale sono state presentate le intuizioni raggruppate in questo capitolo è ancora ampiamente insoddisfacente,

ma ognuna di esse dovrebbe segnare una direzione che questo lavoro sul ragionamento controfattuale può seguire.

Conclusioni

Il messaggio principale contenuto in questa tesi è che la ricerca sul ragionamento controfattuale è ben lungi dall'essere giunta a un punto morto, anche in ambito filosofico; essa può beneficiare delle intuizioni avanzate in settori anche molto lontani della ricerca e molto spesso l'esigenza di fornire delle soluzioni a problemi sorti in altri ambiti disciplinari può contribuire – e di fatto sta già contribuendo – a creare nuove prospettive anche all'interno dei confini del dibattito filosofico.

Nella prima parte della tesi si è dunque compiuta una sorta di ricognizione all'interno della letteratura, partendo da posizioni più "assolutiste", che si ponevano come obiettivo quello di definire quando gli enunciati controfattuali sono "oggettivamente" veri, osservando come stanno le cose nel mondo (o nei mondi), fino a giungere ad approcci più "relativisti", che si proponevano più semplicemente di verificare quando un enunciato controfattuale è derivabile in (o coerente con) un sistema teorico creduto (o assunto come) vero da un agente ragionante.

In questa ricognizione abbiamo cercato di individuare all'interno di tutte queste variegate posizioni – mutuate dalla filosofia, ma anche da altre discipline come l'intelligenza artificiale, la psicologia o l'economia – gli aspetti più interessanti e più rispondenti alla visione intuitiva che ci perviene dall'osservazione quotidiana del fenomeno del ragionamento controfattuale, per tentare poi di riprodurli nel formalismo che abbiamo presentato nel capitolo conclusivo della prima parte.

Scopo del formalismo è di rappresentare il processo di ragionamento attraverso il quale un agente razionale può, sulla scorta di quelle che sono le sue assunzioni di sfondo, inferire che sussista un "legame" controfattuale tra due fatti espressi da due enunciati. Naturalmente, una delle direzioni in cui que-

sto lavoro può e deve essere ampliato è verso la formulazione di teoremi che permettano di mettere in relazione l'informazione ricavata dai processi controfattuali con altra informazione disponibile per far sì che il ragionamento controfattuale divenga uno strumento cognitivo di una certa utilità.

Nei capitoli della seconda parte è stata quindi presa in esame l'applicazione che ci è sembrata più immediata e più utile: quella al ragionamento pratico, dominio nel quale il ragionamento controfattuale può assurgere a strumento di controllo e revisione di piani già portati a termine, con il fine ultimo di approntare dei piani sempre più appropriati per il futuro. All'interno del ragionamento su azioni, particolare attenzione è stata prestata a due specifiche forme di razionalità (strumentale ed *ex-post*), prese a prestito dalla letteratura economica e utilizzate come casi paradigmatici di applicazione del ragionamento controfattuale.

Infine, nella terza e ultima parte, si è cercato di suggerire alcuni terreni sui quali questa analisi potrebbe essere condotta, restituendo in alcuni casi delle intuizioni a settori disciplinari dai quali aveva precedentemente preso ispirazione, come l'epistemologia, lo studio sugli agenti artificiali o sui sistemi multiagente.

Un ulteriore e importante scopo, che però esula da quanto contenuto in questa tesi, è quello di mostrare come tutte queste applicazioni specifiche, che qui sono state illustrate solamente in linee molto generali, possano beneficiare di una sistematizzazione rigorosa, condotta attraverso il sistema formale che questo lavoro ha presentato.

Bibliografia

[1] J. L. Austin. *How to Do Things with Words*. Harvard University Press, Cambridge, Mass., 1962.

[2] A. Balke and J. Pearl. Counterfactual Probabilities: Computational Methods, Bounds and Applications. In R. Lopez de Mantras and D.Poole, editors, *Proceedings of the Conference on Uncertatinty in Artificial Intelligence (UAI 94)*, pages 46–54, San Mateo (CA), July 1994. Morgan Kaufmann.

[3] S. Barker. Counterfactuals, probabilistic counterfactuals and causation. *Mind*, 108:427–469, July 1999.

[4] S. Baron-Cohen, P. Howlin, and J. Hadwin. *Teoria della mente e autismo*. Erickson, Trento, 1999.

[5] J. Barwise. Conditionals and conditional information. In E.C. Traugott, C.A. Ferguson, and J.S. Reilly, editors, *On Conditionals*, pages 21–54. Cambridge University Press, Cambridge (UK), 1986.

[6] J. Barwise and J. Perry. *Situations and Attitudes*. MIT Press, Cambridge, MA, 1983.

[7] M. Benerecetti, P. Bouquet, and C. Ghidini. Formalizing opacity and transparency in belief contexts. In *Practical reasoning and rationality*. AISB, 1997.

[8] M. Benerecetti, P. Bouquet, and C. Ghidini. A multi context approach to belief report. In S. Buvač and L. Ivanska, editors, *AAAI*

Fall 1997 symposium on context in KR and NL. AAAI, 1997. Also IRST-Technical Report 9607-12, IRST, Trento, Italy.

[9] M. Benerecetti, P. Bouquet, and C. Ghidini. Formalizing belief report – the approach and a case study. In F. Giunchiglia, editor, *Artificial Intelligence: Methodology, Systems, and Applications (AIMSA'98)*, volume 1480 of *Lecture Notes in Artificial Intelligence*, pages 62–75. Springer, 1998.

[10] M. Benerecetti, P. Bouquet, and C. Ghidini. Contextual Reasoning Distilled. *Journal of Theoretical and Experimental Artificial Intelligence*, 12(3):279–305, July 2000.

[11] M. Benerecetti, F. Giunchiglia, and L. Serafini. Modeling multiagent systems with local model semantics. In *Proceedings of the AAAI'99 Workshop on Reasoning in Context for AI Applications*, July 19 Orlando, Florida, USA, 1999.

[12] J. Bennett. Counterfactuals and temporal direction. *The Philosophical Review*, XCIII(1):57–91, January 1984.

[13] C. Bicchieri. Strategic behavior and counterfactuals. *Synthese*, 76:135–169, 1988.

[14] C. Bicchieri. Counterfactuals and backward induction. *Philosophica*, 44:101–118, 1989.

[15] C. Bicchieri. Counterfactuals, belief changes, and equilibrium refinements. *Philosophical Topics*, 21, 1994.

[16] N.A. Blue. A metalinguistic interpretation of counterfactual conditionals. *Journal of Philosophical Logic*, 10:179–200, 1981.

[17] M. Bonifacio, P. Bouquet, R. Ferrario, and D. Ponte. Rationality, autonomy and coordination: the sunk costs perspective. In *Proceedings of ESAW'02, Engineering Societies in the Agents World, Third International Workshop*, Madrid, September 2002. Universidad Rey Juan Carlos.

[18] P. Bouquet. A mechanized multi-context solution to mccarthy's glm problem. In *Proceedings AIA-94, 2nd Intnl. Round-Table on Abstract Intelligent Agent*, Rome, Italy, 1994. Also IRST-Technical Report 9406-12, IRST, Trento, Italy.

[19] P. Bouquet. *Contesti e ragionamento contestuale. Il ruolo del contesto in una teoria della rappresentazione della conoscenza*. PhD thesis, Dipartimento di Filosofia, Università di Genova, Genova, Italy, 1997.

[20] P. Bouquet and F. Giunchiglia. Reasoning about theory adequacy: A new solution to the qualification problem. *Fundamenta Informaticae*, 23(2–4):247–262, June,July,August 1995. Also IRST-Technical Report 9406-13, IRST, Trento, Italy.

[21] P. Bouquet and M. Warglien. Mental models and local models semantics: the problem of information integration. In *European Conference on Cognitive Science (ECCS'99)*, Siena (Italy), October 27–30 1999.

[22] A. Bouvier and A. Oliviero. *Azioni, Razionalità e decisioni*. Luiss Edizioni, 2000.

[23] M. E. Bratman. Davidson's theory of intention. In *Faces of Intention*, pages 58–90. Cambridge University Press, Cambridge, 1999.

[24] J. Brockner and J. Z. Rubin. *Entrapment in escalating conflicts: A social psychological analysis*. Springer-Verlag, New York, 1985.

[25] J. Brockner, J. Z. Rubin, and E. Lang. Face saving and entrapment. *Journal of Experimental Social Psychology*, 17:68–79, 1981.

[26] R. A. Brooks. A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, RA-2(1):14–23, 1986.

[27] R. A. Brooks. Intelligence Without Reasoning. In *Proceedings IJCAI-91, 12th Int. Joint Conf. on Artificial Intelligence*, pages 569–595, Sydney, Australia, 1991.

[28] R. A. Brooks. Intelligence Without Representation. *Artificial Intelligence*, 47:139–160, 1991.

[29] R.M.J. Byrne and A. Tasso. Deductive reasoning with factual, possible and counterfactual conditionals. *Memory & Cognition*, 27(4):726–740, 1999.

[30] L. Camaioni. *La teoria della mente*. Laterza, Bari, 1995.

[31] C. Castelfranchi. Guarantees for autonomy in cognitive agent architecture. In N. R. Jennings, editor, *Intelligent Agent: Theories, Architectures, and Languages*, volume 890 of *Lecture Notes in Artificial Intelligence*, pages 56–70, Heidelberg, 1995. Springer-Verlag.

[32] C. Castelfranchi. Modelling social action for AI agents. *Artificial Intelligence*, 103:157–182, 1998.

[33] P. Catellani and P. Milesi. Counterfactuals and roles: Mock victims' and perpetrators' accounts of judicial cases. *European Journal of Social Psychology*, 31:247–264, 2001.

[34] M. L. Dalla Chiara and G. Toraldo di Francia. *Le teorie fisiche*. Boringhieri, Torino, 1981.

[35] A. Cimatti and L. Serafini. Reasoning about belief with multi language systems - a case study. Technical Report 9304-10, IRST, Trento, Italy, 1993.

[36] P.R. Cohen and H.J. Levesque. Intention Is Choice with Commitment. *Artificial Intelligence*, 42:213–261, 1990.

[37] H. Arló Costa. Epistemic conditionals, snakes, and stars. In Oxford University Press, editor, *Conditionals from Philosophy to Computer Science*, pages 203–249. G. Crocco and L. Farinas del Cerro and H. Herzig, 1995.

[38] H. Arló Costa. Belief revision conditionals:basic iterated systems. *Annals of Pure and Applied Logic*, 96:3–28, 1999.

[39] H. Arló Costa and S. J. Shapiro. Maps between nonmonotonic and conditional logic. In B. Nebel, C. Rich, and W. Swortout, editors,

Principles of Knowledge Representation and Reasoning, pages 553–564. Morgan Kaufmann, 1992.

[40] H.L. Arló Costa and I. Levi. Two notions of epistemic validity. *Synthese*, 109:217–262, 1996.

[41] T. Costello and J. McCarthy. Useful counterfactuals. Technical Report Vol. 3 (1999): nr 2, Linköping University, Articles in Computer and Information Science, 1999. `http://ep.liu.se/ea/cis/1999/002/`.

[42] R. Cowley, editor. *What if.* American Historical Publications, 1999. tr. it. in [43].

[43] R. Cowley, editor. *La storia fatta con i se.* Rizzoli, Milano, 2001. tr. it. di [42].

[44] D. Davidson. *Essays on Actions and Events.* Oxford University Press, Oxford, 1980.

[45] J. Dinsmore. Mental spaces from a functional perspective. *Cognitive Science*, 1987.

[46] J. Dinsmore. *Partitioned Representations.* Kluwer Academic Publishers, 1991.

[47] P. Duhem. *La théorie physique. Son objet et sa structure.* Chevalier et Rivière, Paris, 1906. tr. it in [48].

[48] P. Duhem. *La teoria fisica.* Il Mulino, Bologna, 1978. tr. it. di [47].

[49] T. Eiter and G. Gottlob. On the complexity of propositional knowledge base revision, updates, and counterfactuals. *Artificial Intelligence*, 57:227–270, 1992.

[50] J. Elster. *Ulysses and the Sirens.* Cambridge University Press, Cambridge, 1979. tr.it. in [51].

[51] J. Elster. *Ulisse e le sirene.* Il Mulino, Bologna, 1983. tr. it. di [50].

[52] G. Fauconnier. *Mental Spaces: aspects of meaning construction in natural language*. MIT Press, 1985.

[53] G. Fauconnier. *Mappings in thought and language*. Cambridge University Press, 1997.

[54] G. Fauconnier and E. Sweetser, editors. *Spaces, Worlds, and Grammar*. The University of Chicago Press, 1996.

[55] N. Ferguson. *Virtual History: Alternatives and counterfactuals*. Picador, London, 1997.

[56] D. Ferrante. Gli effetti del pensiero controfattuale nella'attività decisionale. *Sistemi Intelligenti*, XII(3):401–414, dicembre 2001.

[57] R. Ferrario. Counterfactual reasoning. In V. Akman, P. Bouquet, R. Thomason, and R. A. Young, editors, *Modeling and Using Context*, volume 2116 of *Lecture Notes in Artificial Intelligence*, pages 170–183, Dundee, UK, July 2001. Springer.

[58] P. Feyerabend. *Against Method: Outline of an Anarchist Theory of Knowledge*. New Left Books, London, 1975. tr. it. in [59].

[59] P. Feyerabend. *Contro il metodo. Abbozzo di una teoria anarchica della conoscenza*. Feltrinelli, 1979. tr. it. di [58].

[60] B. C. Van Fraassen. *The Scientific Image*. Clarendon Press, Oxford, 1980.

[61] D. Gabbay. A general theory of the conditional in terms of a ternary operator. *Theoria*, 38(3):97–104, 1972.

[62] D. Galles and J. Pearl. An axiomatic characterization of causal counterfactuals. Technical report. Prepared for Foundations of Science, Kluwer Academic Publishers.

[63] P. Gärdenfors. *Knowledge in Flux*. Bradford Book. MIT Press, Cambridge, Mass., 1988.

[64] C. Ghidini. Semantiche a modelli locali per logiche multicontestuali. Technical Report Thesis 9404-02, IRST, Trento, Italy, 1994.

[65] C. Ghidini. Modelling (Un)Bounded Beliefs. In P. Bouquet, L. Serafini, P. Brezillon, M. Benerecetti, and F. Castellani, editors, *Modelling and Using Context – Proceedings of the 2nd International and Interdisciplinary Conference, Context'99*, volume 1688 of *Lecture Notes in Artificial Intelligence*, pages 145–158. Springer Verlag - Heidelberg, 1999.

[66] C. Ghidini and F. Giunchiglia. Local Models Semantics, or Contextual Reasoning = Locality + Compatibility. *Artificial Intelligence*, 127(2):221–259, April 2001.

[67] M. L. Ginsberg. Counterfactuals. *Artificial Intelligence*, 30:35–79, 1986.

[68] E. Giunchiglia, P. Traverso, and F. Giunchiglia. Multi-Context Systems as a Specification Framework for Complex Reasoning Systems. In J. Treur and T. Wetter, editors, *Formal Specification of Complex Reasoning Systems*. Ellis Horwood, 1993. Also IRST-Technical Report 9206-22, IRST, Trento, Italy.

[69] F. Giunchiglia. Multilanguage systems. In *Proceedings of AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, 1991. Also IRST-Technical Report 9011-17, IRST, Trento, Italy.

[70] F. Giunchiglia. Reasoning with contexts. Technical Report 9204-19, IRST, Trento, Italy, 1992.

[71] F. Giunchiglia. Contextual reasoning. *Epistemologia, special issue on I Linguaggi e le Macchine*, XVI:345–364, 1993. Short version in Proceedings IJCAI'93 Workshop on Using Knowledge in its Context, Chambery, France, 1993, pp. 39–49. Also IRST-Technical Report 9211-20, IRST, Trento, Italy.

[72] F. Giunchiglia and P. Bouquet. A Context-Based Framework for Mental Representation. In *Proceedings of the Twentieth Annual Meeting of the Cognitive Science Society - CogSci'98*, pages 392–397, Madison,

Wisconsin (USA), August 1998. Cognitive Science Society, Lawrence Erlbaum Associates.

[73] F. Giunchiglia and C. Ghidini. A Local Models Semantics for Propositional Attitudes. In P. Bonzon, M. Cavalcanti, and R. Nossum, editors, *Formal Aspects of Context*, volume 20 of *Applied Logic Series*. Kluwer Academic Publishers, July 2000. Also IRST-Technical Report 9607-12, IRST, Trento, Italy.

[74] F. Giunchiglia and L. Serafini. Multilanguage systems (provably equivalent to modal logics). Technical Report 9002-05, IRST, Trento, Italy, 1990.

[75] F. Giunchiglia and L. Serafini. Multilanguage first order theories of propositional attitudes. In *Proceedings 3rd Scandinavian Conference on Artificial Intelligence*, pages 228–240, Roskilde University, Denmark, 1991. IOS Press. Also IRST-Technical Report 9001-02, IRST, Trento, Italy.

[76] F. Giunchiglia and L. Serafini. Multilanguage hierarchical logics or: how we can do without modal logics. *Artificial Intelligence*, 65(1):29–70, 1994. Also IRST-Technical Report 9110-07, IRST, Trento, Italy.

[77] F. Giunchiglia, L. Serafini, E. Giunchiglia, and M. Frixione. Non-Omniscient Belief as Context-Based Reasoning. In *Proc. of the 13th International Joint Conference on Artificial Intelligence*, pages 548–554, Chambery, France, 1993. Also IRST-Technical Report 9206-03, IRST, Trento, Italy.

[78] N. Goodman. The problem of counterfactual conditionals. In F. Jackson, editor, *Conditionals*, pages 9–27. Oxford University Press, 1991.

[79] R.V. Guha and D.B. Lenat. Counterfactuals. In *Proc. Stanford Spring Workshop on Logical Formalizations of Commonsense Reasoning.*, 1990.

[80] J. Y. Halpern. Hypothetical knowledge and counterfactual reasoning. *International Journal of Game Theory*, 28, 1999.

[81] J. Y. Halpern and Y. Moses. Using counterfactuals in knowledge-based programming. In *Proceedings of the Seventh Conference on Theoretical Aspects of Rationality and Knowledge*, pages 97–110, 1998.

[82] N. R. Hanson. *Patterns of Discovery. An Inquiry into the Conceptual Foundations of Science*. Cambridge University Press, Cambridge, 1958. tr. it. in [83].

[83] N. R. Hanson. *I modelli della scoperta scientifica. Ricerca sui fondamenti concettuali delle scienze*. Feltrinelli, Milano, 1978. tr. it. di [82].

[84] R. Harris. *Fatherland*. Mondadori, Milano, 1992. tr. it. di [85].

[85] R. Harris. *Fatherland*. Random House, 1992. tr. it. in [84].

[86] S. J. Hoch. Counterfactual reasoning and accuracy in predicting personal events. *Journal of Experimental Psychology*, 11(4):719–731, 1985.

[87] C.L. Ortiz Jr. Explanatory update theory: Applications of counterfactual reasoning to causation. *AI*, 108:125–178, 1999.

[88] D. Kahneman and A. Tversky. The simulation heuristic. In D. Kahneman, P. Slovic, and A. Tversky, editors, *Judgement under uncertainty: Heuristics and biases*, pages 201–208. Cambridge University Press, New York, 1982.

[89] D. Kahneman and C. A. Varey. Propensities and counterfactuals: The loser that almost won. *Journal of Personality and Social Psychology*, 59(6):1101–1110, 1990.

[90] A. Kratzer. Partition and revision: the semantics of counterfactuals. *Journal of Philosophical Logic*, 10:201–216, 1981.

[91] T. Kuhn. *The structure of Scientific Revolutions*. University of Chicago Press, 1979.

[92] I. Kvart. *A Theory of Counterfactuals*. Hackett, Indianapolis, 1986.

[93] I. Kvart. Counterfactuals. *Erkenntnis*, 36:139–179, 1992.

[94] I. Lakatos. History of science and its rational reconstructions. In J. Worrall and G. Currie, editors, *The methodology of scientific research programmes. Philosophical Papers*, volume 1. Cambridge University Press, Cambridge Mass., 1978.

[95] I. Lakatos. La falsificazione e la metodologia dei programmi di ricerca scientifici. In M. D'Agostino, editor, *La metodologia dei programmi di ricerca scientifici. Scritti filosofici*, volume 1. Il Saggiatore, Milano, 1985. tr. it. di [97].

[96] I. Lakatos. La storia della scienza e le sue ricostruzioni razionali. In M. D'Agostino, editor, *La metodologia dei programmi di ricerca scientifici. Scritti filosofici*, volume 1. Il Saggiatore, Milano, 1985. tr. it. di [94].

[97] I. Lakatos and A. Musgrave. Falsification and the methodology of scientific research programmes. In G. Currie J. Worrall, editor, *The Methodology of Scientific Research Programmes: Philosophical Papers*, Cambridge, 1978. Cambridge University Press.

[98] M. Lange. Inductive confirmation, counterfactual conditionals, and laws of nature. *Philosophical Studies*, 1997.

[99] D. Lewis. General Semantics. *Synthese*, 22:18–67, 1970. Reprinted in [102].

[100] D. Lewis. *Counterfactuals*. Blackwell, 1973.

[101] D. Lewis. Ordering semantics and premise semantics for counterfactuals. *Journal of Philosophical Logic*, 10:217–234, 1981.

[102] D. Lewis. *Philosophical papers*. Oxford University Press, 1983. Two volumes.

[103] D.K. Lewis. Counterfactuals and comparative possibility. *Journal of Philosophical Logic*, 2:418–446, 1973.

[104] D.K. Lewis. Controfattuali e possibilità comparativa. In C. Pizzi, editor, *Leggi di natura, modalità, ipotesi. La logica del ragionamento controfattuale*, pages 233–263. Feltrinelli, Milano, 1978. tr. it. di [105].

[105] D.K. Lewis. Counterfactuals and comparative possibility. In *Philosophical Papers*, chapter 16, pages 3–31. Oxford University Press, 1983. tr. it. in [123].

[106] M. G. Lipe. Counterfactual reasoning as a framework for attribution theories. *Psychological Bulletin*, 109(3):456–471, 1991.

[107] L.B. Lombard. Causes, enablers, and the counterfactual analysis. *Philosophical Studies*, 1990.

[108] E.J. Lowe. The truth about counterfactuals. *The Philosophical Quarterly*, 1995.

[109] J. G. March. How decisions happen in organizations. *Human Computer Interaction*, 6:95–117, 1991.

[110] J. G. March. *Decisioni e organizzazioni*. Il Mulino, 1993.

[111] J. G. March. *A primer on decision making : how decisions happen*. The Free Press, 1994.

[112] J. G. March. *Prendere decisioni*. Il Mulino, 1998. Traduzione di Stefano Micelli.

[113] M. McDermott. Counterfactuals and Access Point. *Mind*, 1999.

[114] A. McEleny and R. M. J. Byrne. Consequences of counterfactual reasoning and causal reasoning. In S. Bagnara, editor, *European Conference on Cognitive Science '99*, pages 199–205, Siena (Italy), 1999.

[115] M. N. McMullen, K. D. Markman, and I. Gavanski. Living in neither the best nor the worst of all possible worlds: Antecedents and consequences of upward and downward counterfactual thinking. In N. J. Roese and J. M. Olson, editors, *What might have been: The social psychology of counterfactual thinking*. Erlbaum, Mahwah, NJ, 1995.

[116] P. Menzies. Difference-making in context. In J. Collins, N. Hall, and L. Paul, editors, *Counterfactuals and Causation*. MIT Press, 2002.

[117] J. S. Mill. *Principi di economia politica*. UTET, 1954.

[118] D. T. Miller and W. Turnbull. The counterfactual fallacy: Confusing what might have been with what ought to have been. *Social Justice Research*, 4:1–19, 1990.

[119] M. W. Morris and P. C. Moore. The lessons we (don't) learn: Counterfactual thinking and organizational accountability after a close call. *Administrative Science Quarterly*, 45:737–765, 2000.

[120] D. Nute. Conditional logic. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, volume 2, pages 387–439. Reidel, 1984.

[121] J. Pearl. Causation, action and counterfactuals. In *Proceedings of TARK 1996*, pages 51–73, 1996.

[122] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[123] C. Pizzi, editor. *Leggi di natura, modalità, ipotesi. La logica del ragionamento controfattuale*. Feltrinelli, Milano, 1978.

[124] C. Pizzi. Fictionalism and the logic of "as if" conditionals. In L. Magnani, N. J. Nersessian, and P. Thagard, editors, *Model-based Reasoning in Scientific Discovery*, New York, 1999. Kluwer A. P.

[125] C. Pizzi. Deterministic models and the "unimportance of the inevitable". In L. Magnani and N. J. Nersessian, editors, *Model-Based*

Reasoning: Science, Technology, Values, New York, 2002. Kluwer Academic/Plenum Publishers.

[126] C. Pizzi. Il ragionamento controfattuale. *Problémata*, 1(1):85–94, 2002.

[127] H. Poincaré. *La science et l'hypothèse*. Flammarion, Paris, 1902. tr. it. in [128].

[128] H. Poincaré. *Opere Epistemologiche*, volume I. Piovan, Abano Terme, 1989. tr. it. di [127].

[129] J. L. Pollock. Interest driven suppositional reasoning. *Journal of Automated Reasoning*, 6:419–462, 1992.

[130] J. L. Pollock. The phylogeny of rationality. *Cognitive Science*, pages 563–588, 1993.

[131] J.L. Pollock. A refined theory of counterfactuals. *Journal of Philosophical Logic*, 10:239–266, 1981.

[132] J.L. Pollock. New foundations for practical reasoning. *Minds and Machines*, 2:113–144, 1992.

[133] K. R. Popper. *Conjectures and Refutations*. Routledge and Kegan, London, 1969.

[134] F. P. Ramsey. *Foundations of Mathematics and other Logical Essays*, chapter General Propositions and Causality, pages 237–257. New York, 1950.

[135] A. S. Rao and M. P. Georgeff. Modeling rational agents within a BDI architecture. In J. Allen, R. Fikes, and E. Sandewall, editors, *Proceedings of the 2nd International Conference on Principle of Knowledge Representation and Reasoning*. Morgan Kaufmann Publishers, 1991.

[136] A. Reboul. If i were you, i wouldn't trust myself: indexicals, ambiguity and counterfactuals. In *Time, Space and Identity, Acts of the 2nd International Colloquium on Deixis*, pages 151–175, Nancy, 1996.

[137] N. Rescher. *The Coherence Theory of Truth*. Oxford University Press, London, 1973.

[138] N. Rescher. L'analisi coerentista dei controfattuali. In C. Pizzi, editor, *Leggi di natura, modalità, ipotesi. La logica del ragionamento controfattuale*, pages 114–129. Feltrinelli, Milano, 1978. tr. it. del cap.XI di [137].

[139] N. J. Roese. Counterfactual thinking. *Psychological Bulletin*, 121:133–148, 1997.

[140] A. R. Schotter. *Microeconomia*. Giappichelli Torino, 1995.

[141] L. Serafini and C. Ghidini. Local Models Semantics for Information Integration. Technical Report 9702-04, IRST, Trento, Italy, February 1997. Extended abstract presented as poster at the 15th IJCAI.

[142] L. Serafini and F. Giunchiglia. Ml systems: A proof theory for contexts. *To appear in the Journal of Logic Language and Information*, 2000. Also Technincal Report 0006-01, ITC-IRST, Trento (Italy).

[143] H. A. Simon. *La ragione nelle vicende umane*. Il Mulino, 1984.

[144] H. A. Simon. *Causalità, razionalità, organizzazione*. Il Mulino, 1985.

[145] H. A. Simon. Dalla razionalità sostanziale alla razionalità procedurale. In M. Egidi and M. Turvani, editors, *Le Ragioni delle Organizzazioni Economiche*, pages 291–317, Torino, 1994. Rosenberg e Sellier.

[146] H. A. Simon. *La ragione delle Organizzazioni Economiche*. Rosenberg e Sellier, 1994.

[147] H. A. Simon, M. Egidi, R. Marris, and R. Viale. *Economics, bounded rationality and the Cognitive Revolution*. Il Mulino, 1992.

[148] R. Sobel. *For Want of a Nail...; If Burgoyne Had Won at Saratoga*. Greenhill/Stackpole, 1997.

[149] B. A. Spellman and D. R. Mandel. When possibility informs reality: Counterfactual thinking as a cue to causality. *Current Directions in Psychological Science*, 8:120–123, 1999.

[150] R. Stalnaker. A Theory of Conditionals. In F. Jackson, editor, *Conditionals*, Oxford Readings in Philosophy, pages 28–45. Oxford University Press, 1991.

[151] R. Stalnaker. Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy*, 12:133–163, 1996.

[152] R. Stalnaker and R.M. Thomason. A semantic analysis of conditional logic. *Theoria*, 36:23–42, 1970.

[153] R. Stalnaker and R.M. Thomason. Analisi semantica della logica condizionale. In C Pizzi, editor, *Leggi di natura, modalità, ipotesi*, pages 215–232. Feltrinelli, Milano, 1978. tr. it. di [152].

[154] M. Swain. A counterfactual analysis of event causation. *Philosophical Studies*, 34:1–19, 1978.

[155] E. Sweetser and G. Fauconnier. Cognitive links and domains. In G. Fauconnier and E. Sweetser, editors, *Spaces, Worlds, and Grammar*, pages 1–28. The University of Chicago Press, Chicago and London, 1996.

[156] P. E. Tetlock and A. Belkin. Counterfactual thought experiments in world politics. *Social Science Research Council*, 50(4), December 1996.

[157] P. E. Tetlock and A. Belkin, editors. *Counterfactual thought experiments in world politics: Logical, methodological, and psychological perspectives*. Princeton University Press, Princeton, NJ, 1996.

[158] P. G. Tsouras. *Gettysburg: An Alternate History*. Greenhill/Stackpole, 1997.

[159] T.S. Ulen. Rational choice theory in law and economics. *Encyclopedia of Law and Economics*, pages 790–818, 1999.

[160] K. Warmbrod. Counterfactuals and substitution of equivalent antecedents. *Journal of Philosophical Logic*, 10:267–289, 1981.

[161] R. L. Wiener, M. Gaborit, C. C. Pritchard, and E. M. McDonough. Counterfactual thinking on mock juror assessments of negligence. *Behavioral Sciences and the Law*, 12:89–102, 1994.

[162] W. Wobcke. A theory of conditionals based on hierarchies of situations. citeseer.nj.nec.com/128705.html.

[163] M. Wooldridge and N. Jennings. Intelligent agents: Theory and practice. *The Knowledge Engineering Review*, 10(2):115–152, 1995.

[164] D. G. Yarlett and M. J. A. Ramscar. Structural determinants of counterfactual reasoning. In Lawrence Erlbaum Associates, editor, *Proceedings of the 23rd Annual Meeting of the Cognitive Science Conference*, Mahwah, NJ, 2001.

[165] F. Znaniecki. *Social Action*. Farrar and Reinehart, New York, 1936.

Elenco delle figure

Indice dei nomi