Counterfactual Reasoning

Roberta Ferrario

Département de Philosophie – Université Marc Bloch de Strasbourg
Dipartimento di Filosofia – Università degli Studi di Milano
Via Festa del Perdono, 7 - Milano (Italy)
Email: ferrix@cs.unitn.it

Abstract. Primary goal of this paper is to show that counterfactual reasoning, as many other kinds of common sense reasoning, can be studied and analyzed through what we can call a cognitive approach, that represents knowledge as structured and partitioned into different domains, everyone of which has a specific theory, but can exchange data and information with some of the others. Along these lines, we are going to show that a kind of "counterfactual attitude" is pervasive in a lot of forms of common sense reasoning, as in theories of action, beliefs/intentions ascription, cooperative and antagonistic situations, communication acts. The second purpose of the paper is to give a reading of counterfactual reasoning as a specific kind of contextual reasoning, this latter interpreted according to the theory of MultiContext Systems developed by Fausto Giunchiglia and his research group.

1 Introduction

Counterfactuality has been the focus of a multitude of works, in philosophy [20, 25, 13, 16, 2, 23, 24, 19], in psychology [3, 22, 15], in artificial intelligence [9, 14, 4, 17] and in the cognitive sciences in general [8, 18]. In most approaches the study of counterfactuality is related to the problem of causality, and there is a wide agreement in describing counterfactuals as a powerful tool in explaining past events and predicting future outcomes.

In the literature, it is possible to find two different approaches to a theory of conterfactuals: a *metaphysical approach*, in which the problem is mainly to define the relationship between the actual world and a counterfactual world; and a *cognitive approach*, in which the emphasis is on the properties of counterfactual reasoning from the perspective of an agent in a given situation. In this paper we assume the cognitive approach to argue that counterfactual reasoning can be treated as a specific kind of contextual reasoning. This will be done as a preliminary step toward our long term goal, that is to build a formal system based on the logic of MultiContext Systems [12].

Our main interest is in how agents reason when they face scenarios in which actions can be influenced by the presence of other agents and have consequences for the other agents involved. We believe this is a dimension of counterfactual reasoning that has not been satisfactorily investigated in the literature. This is

shown, for example, by the fact that an interesting typology of counterfactual reasoning, namely counterfactuals of the form "If I were you ..." (called *counteridenticals*), has been almost neglected, even though it seems extremely useful to ascribe beliefs to other agents in multi-agent scenarios (e.g. in cooperative and antagonistic reasoning and, even more importantly, in communication acts).

The paper goes as follows. After a brief introduction on cognitive approaches, the main section of the paper is dedicated to an analysis of possible applications of counterfactual reasoning to other forms of common sense reasoning, showing that many of them have a (sometimes hidden) counterfactual dimension. Then we present our main thesis, namely that counterfactual reasoning can be studied as a type of contextual reasoning. In order to do this, we sketch the definitions of context and contextual reasoning as they are given in [12]. In the final part of the paper we present some preliminary ideas on a possible connection between counterfactual reasoning and contexts on one side and Game Theory on the other.

2 Cognitive approaches to counterfactual reasoning

As a general definition, we call *counterfactual reasoning* all those reasoning processes that an agent performs starting from a set of assumptions she believes to be *true*, with the addition of an hypothesis that she believes to be *false*, but that she treats as true *for the sake of the argument*. A simple example is the sentence:

"If I could turn back time, I would have studied economics"

On the "cognitive front", there are two theories that are extremely relevant for our approach to counterfactual reasoning: the theory of *mental spaces*, proposed by Gilles Fauconnier [7], and the theory of *partitioned representations*, proposed by John Dinsmore [5]. Both theories share the intuition that the cognitive state of an individual is better described as divided into multiple portions, called mental spaces in one case, partitioned representations in the other. Agents carry on reasoning processes locally to these portions of their mental state, trying to build a representation of reality. Dinsmore calls these processes *simulative reasoning* [6]:

> "Simulative reasoning requires a partitioning of knowledge into distinct spaces and additionally assumes that the contents of each space effectively simulate or model a possible reality, or a part of a possible reality, and therefore represents a meaningful domain over which normal reasoning processes work."

There are some elements in the language that work as *space builders*, because they introduce new partitions in the cognitive state. Counterfactuals are one of these space builders. In particular, in Fauconnier's view, they open a peculiar

kind of hypothetical space, whose structure is *analogical* [8], and not truth functional, as in Lewis' and Stalnaker's approach [20, 25]). According to Fauconnier, it is a *projection* of the structure of the base space.

Fauconnier calls *base space* the mental space from which it is originated the counterfactual space, through an analogy-based mechanism. This mechanism requires that some *matching conditions* are met by a counterfactual space in order to be related to a given base space. Fauconnier makes this point as follows:

> "[...] a counterfactual sets up an imaginary situation which differs from the actual one in one fundamental respect, expressed in the antecedent part (A, the protasis) of the *if A then B* construction. [...] In spite of appearences, the structure of counterfactuals is not truth functional (entailment from an alternative set of premises); it is analogical: projection of structure from one domain to another. [...] What is the use of C [the counterfactual space] in the discourse? It does not give direct information about actual situations, and it does not represent existing frame configurations. However, besides being counterfactual, C is also *conditional.* The semantics linked to C includes the general *matching* conditions on hypothetical spaces. The matching condition (an extended form of modus ponens) specifies in general that a space matching the defining structure of a conditional space fits it in all other respects." [8]

The *projection of structure* called for is the analogous of what Dinsmore calls the *default inheritance*:

> "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 [...] spaces. [...] The kind of inheritance involved in this case cannot be absolute. [...] Such cases require a weaker form of inheritance, *default inheritance.*" [6]

3 The counterfactual dimension of common sense reasoning

Our next step is to argue that there are many forms of common sense reasoning that involve reasoning processes with a counterfactual structure. Some of them, as practical reasoning, have already been mentioned in literature; however, there are many more that haven't been explored yet and that can reveal very useful applications.

3.1 Counterfactual reasoning and theories of action

In the philosophical tradition many authors stress the strong connection between counterfactuality and causality (see for instance [21, 2, 24]), whereas in AI people

have widely investigated the role of counterfactual reasoning in the diagnosis of artificial systems' failures and in the planning of future actions ([9] is a paradigmatic reference in this area). However, from our perspective, we can identify two general types of applications: one directed to reason about the past and the other one focused on reasoning about states of affairs. Each type can then be divided into two sub-categories, depending on the outcome of the reasoning. So we have four cases:

- **Past strategies to be changed**: the agent has previously planned an action that didn't reach the goal; she has to figure out different scenarios in which she alters one of the elements of the plan with the purpose of understanding what has gone wrong and has to be changed. The general schema is the following: "If I had performed that different action, I would have reached my goal". An instance of the schema is: "If I had come before, I would have met the President".
- **Past strategies to be confirmed**: the agent has previously planned a successful action; she can try to guess which elements of the plan were decisive for success, to be able to use them again in other plans. What she has to do is simply to imagine altered situations in which the lack of one or more of the elements influence negatively the outcome of the plan. The general schema is: "If I hadn't performed that action, this result wouldn't have been possible". A possible instance is: "If I hadn't waken up so early, I wouldn't have been able to arrive at the station on time"
- **State of affairs to be changed**: the agent can realize that she lacks some means to an end or some essential characteristics to obtain what she's looking for by imagining a situation in which she would have these means or characteristics. The general schema is: "If I were that way, I would do this thing". An example is: "If I were more corageous, I would ask my chief more money".
- **State of affairs to be confirmed**: the agent figures out a situation in which she lacks something (a means, a characteristic) she actually has and she realizes that she would not be able to do something she actually can do. The general schema is: "If I didn't have this property, I wouldn't be able to do this thing". An example is: "If I hadn't these mobility funds, I would not be able to travel so often"

The intuitive picture of counterfactual reasoning processes and of how it works in presence of plans and strategies is the following:

1. the agent wants to reason about a particular fact, event or problem. Thus she *selects*, inside her global knowledge base, a set of relevant assumptions which, in her opinion, are necessary and sufficient for the reasoning process at hand;
2. she builds a working context for the reasoning process, in which are contained all the assumptions she has previously selected and the effect that has been reached;

3. finally, a counterfactual context is constructed by:
   - importing all the assumptions from the working context, but changing the truth-value of one of them, or
   - importing all the assumptions from the working context and adding a new assumption that wasn't previously selected (this possibility is very important to show that counterfactual reasoning is a form of non monotonic reasoning).

   In both cases, the result of the reasoning performed inside the counterfactual context will be the negation of the one reached in the working context.

The final purpose of the reasoning process performed inside the counterfactual context is not only to show the relevance of the datum that has been changed (the counterfactual hypothesis), but also to show the importance of keeping all the other data unchanged. What the agent is trying to demonstrate with the counterfactual reasoning is the correctness of the choice she has made about the most relevant assumptions.

If we reconsider the classification given above about satisfactory/unsatisfactory strategies or states of affairs and the consequent strategy confirmation or revision, we can reformulate it and define a goal for counterfactual reasoning.

Counterfactual reasoning can be interpreted as a mechanism of verification and control of the selection function: counterfactual reasoning checks if the assumptions selected are *all* there is that is relevant for the reasoning and if they are *the only* that are relevant.

Strategy confirmation . The agent has elaborated a strategy that has reached the expected goal. Still, she wants to check if all the assumptions she has considered were necessary to the achievement of the result and if there was something else that she has not considered which could have prevented the outcome of the plan.
- In order to understand if all the assumptions were necessary, she can try to negate one or the other and verify if this change influences the result
- In order to understand if she has considered a sufficient set of assumptions, she can try to add some other assumption that could look relevant and see which is the result. If it doesn't change, this new assumption is redundant, if it does, this has to be added to the set of the relevant ones.

Strategy revision . The strategy hasn't reached the goal. The agent wants to understand if her statement of the problem was correct, if something she has done or thought has been an obstacle to the realization of the plan or if she has neglected some important assumption
- Analogously as in the case of strategy confirmation, the agent verifies if the change of truth-value of an assumption affects the result of the reasoning. If it does, this assumption can be considered responsible of the failure of the plan
- Similarly, the agent tries to guess if there was some unexpected obstacle she hasn't considered.

We can try to make these ideas clearer with an example.

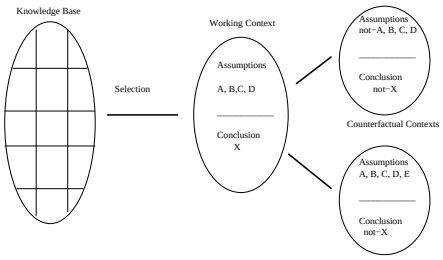

Fig. 1. The process of counterfactual strategy confirmation

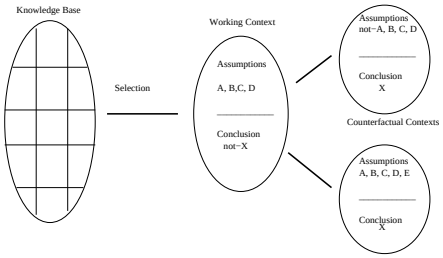

Fig. 2. The process of counterfactual strategy revision

Example: the flight to Paris The situation is the following: we have an agent, Anna, who wants to take a flight to Paris. She lives in a small town near Milan and she has selected certain assumptions in order to take the flight at 10 o'clock in the morning at the airport of Milan Malpensa. These are the assumptions she has selected. She has to:

- Pack the suitcase the evening before
- Close all the windows and back doors the evening before
- Check if there is a bus at 8 o'clock that arrives at the airport at 9.
- The alarm must ring at 7 o'clock

1. **Strategy confirmation** In this case, we can imagine that all the conditions have been fulfilled and in the end Anna has taken her flight to Paris. Now we have the two possible counterfactuals:
   - Alteration of the truth-value of a selected condition: "If the alarm hadn't rung, I wouldn't have caught the plane"
   - Addition of a new relevant condition: "If there had been a strike of the bus drivers, I wouldn't have caught the plane"

   Anna wants to show that, assuming that all the other conditions hold, the ringing of the alarm or the strike are very relevant elements.
2. **Strategy revision** In this case the situation is that the alarm actually didn't ring and Anna didn't catch the plane. Here are the two counterfactuals:

- Alteration of the truth-value of a selected condition: "If the alarm had rung, I would have caught the plane"
- Addition of a new relevant condition: "If I had set two alarms, instead of one, I would have caught the plane"

Anna is trying to show not only that the fact that the alarm didnt't ring is the cause of the failure of her plan, but also that it was the only reason why the plan failed: all the other important assumptions have been taken into consideration.

What the agent is interested in is to discover if her plan was correctly settled, if she has forgotten anything or if she has considered something that was inessential. The counterfactual reasoning is a tool to analyze the relevance of the assumptions previously selected from the global knowledge base.

There are a lot of other forms of counterfactual reasoning that are not so strictly connected with plans and actions that must be examined in more detail and this is undoubtedly a direction in which this analysis has to be deepened.

As we have seen, there are many situations in which a single individual is involved that require a counterfactual form of reasoning. What we are going to show next is how counterfactuality (and in particular a preculiar kind of counterfactual reasoning) can be considered a very important element in the processes of reasoning typical of multiagent situations.

3.2 Counterfactual reasoning and beliefs/intentions ascription

Before presenting our own suggestion, we have to set a preliminary fact: we think we can keep for granted that when agents have to interact with each other, eveyone of them has to ascribe beliefs and intentions to others and, since none of them has direct access to the cognitive state of the others, they have to find some means to figure out how others reason.

Each agent has two direct sources to detect in order to guess how other minds work: one external and the other internal.

Externally, an agent can listen to the descriptions other agents give of their own reasonings, hoping that they are sincere; moreover, she can observe their behavior and try to deduce their internal processes of reasoning from their "*modus operandi*".

Internally, an agent can try to be "selfconscious" of her own reasoning schemes and assume that all minds work in a similar way, i.e. like her own does.

Putting together these two amounts of data, the agent can perform a "mental act" consisting in her "walking in somebody else's shoes"; she constructs a new cognitive context containing knowledge, beliefs and intentions that she ascribes to the other agent and, from this context, she begins a process of reasoning using her cognitive tools as a substitute of those belonging to the other agent.

To make it clearer, here are the schema and an example:

"If I were X, knowing what he knows and having that particular belief, I would perform this action".

"If I were Giovanni, I would invite Anna to the party"

All reasonings of this kind have the purpose of predicting the actions and opinions of other agents in order to accomodate our behavior as to obtain the best result from the interactions with others.

There are three important applications of this "counterfactual ascription of beliefs/intentions": cooperative reasoning, antagonistic reasoning and, maybe the most interesting case (because in a way it applies to the others), communication.

Cooperative reasoning The first application we are going to consider is cooperative reasoning. Cooperative reasoning is essential for agents, because they often cannot execute with success what they have planned unless they don't ask the help of other agents.

But, before deciding whom to ask, the agent must consider capabilities and knowledge of the possible candidates and guess if they can be, directly or indirectly, interested in her plan.

To do this, once again, the agent has to perform the reasoning process "If I were X...". If the result of the reasoning is that the candidate is suitable and would probably agree to join the plan, she can proceed and ask the help of this agent; otherwise, she has to activate the same process "If I were X..." to think about something that the other would find appealing to propose as a "payoff" of the requested cooperation.

Another application in the domain of cooperative reasoning is when an agent considers retrospectively a plan (both individual and cooperative).

If the plan has failed, she can think about an alternative situation (that didn't take place) in which, with the help of some agent (or of a different agent if the plan was cooperative), it could have ended successfully.

Instead, if the plan was successful, she can try to imagine how things could have been without the cooperation of that particular agent (virtually substituting that agent with another one, or imagining an individual action instead of the cooperative) in order to understand how profitable the cooperation was.

Antagonistic reasoning The situation is dual to that of cooperative reasoning: the agent has settled a plan, but she realizes that there is an obstacle to overcome: the opposition of another agent.

To begin with, what she has to do is trying to guess why the other agent is against her plan. To do it, once again she has to think "Why would I be against this plan, if I were this agent?".

The first possibility is that the other agent has misunderstood some of her intentions, so she has only to try to explain her reasons to show the other that his opposition is not actually justified.

The other possibility is that the other agent has a strong and justified reason to oppose the plan. In this case the agent must "walk in the shoes of the other" ("If I were X...") and figure out what she can offer in exchange as to make him give up his opposition.

Both cooperative and antagonistic reasoning and in general all reasoning processes including multiagent scenarios are based upon the possibility for the agents to communicate.

As we are going to argue in the next paragraph, we think that the important function of communication rests on the capability of agents to ascribe each other certain ways of reasoning; counterfactual reasoning of the form "If I were X..." is an important tool in this direction.

Counterfactual reasoning in communication When an agent is to begin a communication, she has to check the conditions that will make her communication act effective. She has to elaborate a kind of *communication strategy.*

The first element to be considered is the form of the language that will be used. This language must not be too complicated or too simple relatively to the capabilities of the receiver: if it is too complicated, there is the risk of a lack of understanding; if it is too simple, it could be judged inappropriate by the receiver.

The language must match not only the cultural and cognitive features of the agent receiving the communication, but it has also to be fitted into the situation (in some cases it must be technical, in others informal and so on and so forth).

Second, the agent has to consider the degree of interest that the content of the communication can arise in the listener and she has to evaluate if the receiver has a minimal competence in the subject, otherwise the communication act would be pointless.

Finally, when agents communicate, usually they have the purpose of persuading the other agent to perform an action, to share the same opinion about something or to behave in a particular way. The goal of persuasion can be obtained only mixing together the right form with the right content, producing the convincing arguments.

All these evaluations can be stated thanks to counterfactual processes of reasoning of the kind "If I were X..." and through the ascription of certain beliefs and cognitive capabilities to other agents, that are fundamental to predict other agents' reactions to our communication acts.

Now we can think about an example that can summarize a series of situations in which an imaginary agent is involved in various sequences of counterfactual thoughts.

Mr.1 is a wealthy middle-aged man, who lives in a small villa with a little but beautiful garden, in common with his neighbour.

Last year, in this season, he had an accident with a tree that was in his garden: after a storm, the tree fell against the house, damaging the roof.

Now Mr.1 is in the garden looking at another tree of the same species that looks similarly ill; he looks also at the sky and notices that the weather is getting worse. Then he thinks: **"If last year I had cut the tree before the storm, it wouldn't have damaged the roof"**.

So, his next action is to manage to cut the tree, to avoid to pay for the repairing.

But let's suppose that, after having tryed to do it by himself, he realizes that he's not able to. But then it comes to his mind that his neighbour (Mr.2) was once a gardener, now retired, and that he could help him.

Then he thinks that, being the tree situated in a place such that it could fall even against the house of Mr.2, maybe Mr.2 would agree about the convenience of cutting the tree. The thought of Mr.1 will be something like: **"If I were Mr.2, I would be worried about the tree and I would agree to cut it"**.

Before asking Mr.2, Mr.1 has to decide in which way to express his thoughts to Mr.2; this will depend on the idea Mr.1 has about Mr.2: which are his basic beliefs, which is his level of education, which are his prejudices and so on and so forth. This operation is in most cases implicit, but it is expressible in counterfactual terms: **"If I were Mr.2, which argument would I understand and find convincing?"**.

Sometimes it can happen that the evaluation of Mr.1 relative to the thoughts of Mr.2 is not correct. In this case, we suppose the answer of Mr.2 is something like: "I don't think there is a need of cutting the tree, this coming storm won't be devastating like the one of last year."

In this case, Mr.1 has to find another way to persuade Mr.2 to help him, maybe proposing something in exchange. Even in this situation, Mr.1 can think **"If I were Mr.2, I would agree to cut the tree, provided that, after the tree is cut, we would build in the place now occupied by the tree a gazebo"**, because some months ago Mr.2 proposed this thing, that Mr.1 refused.

These are some of the possible examples of common sense reasoning conducted with a counterfactual attitude; we think they show some applications of counterfactual reasoning underestimated, at least untill now.

In our opinion, all these instances of counterfactuality can be studied and analyzed according to a contextual approach in a way that we will try to illustrate in the next paragraph.

4 Contexts and Contextual Reasoning in MultiContexts Systems

There are a great variety of works on contexts and contextual reasoning, but there is a particular interpretation of the notion of contexts and - consequently - of contextual reasoning that we find appropriate for the analysis of counterfactual reasoning that we want to give.

This interpretation is the one given by Fausto Giunchiglia and his research group (MRG: Mechanized Reasoning Group) and the formalization derived from this perspective is called MultiContext Systems.

There is a wide literature on MultiContext Systems [10, 12, 11, 1], so our goal here will not be that of giving precise formal definitions, but we want to give an idea about what contexts are with respect to those definitions that are given somewhere else and how counterfactual reasoning, analogously defined, works.

Firstly, contexts are theories (in the formal sense of the term: each of them has its language, axioms and inference rules).

They have three main features: *partiality*, *approximation* and *perspective*.

- They are *partial* because each context of reasoning utilizes only a subset of the knowledge base that is actually available to the agent;
- They are *approximate* because the representation expressed by a context can be presented at different and variable levels of detail. In other words, a set of parameters (time, space, agent, ...) defines each context and their number can be varied.
- They are *perspectival* because they always express the epistemic point of view of an agent.

Being contexts these partial objects, we have two forms of contextual reasoning: inside a single context and between different contexts.

- The reasoning performed inside a single context has been defined *local reasoning* and it utilizes only the language, axioms and inference rules peculiar of that specific context;
- The process of reasoning that begins with a premise stated in a context and that ends with a conclusion drawn in a different context needs an appropriate tool to switch from one context to another. This tool has been semantically defined as *compatibility relations*.

The logics of MultiContext Systems has revealed very useful in the resolution of some typical philosophical problems, such as the treatment of indexical expressions and the difficulties connected to belief ascription.

Our hope is that some good results could be reached even in the analysis of counterfactual reasoning; the reason that supports our hope is the intuition that countefactual reasoning can be viewed as a specific kind of contextual reasoning. In the next paragraph we will show why it is so.

5 Counterfactual reasoning as a particular kind of contextual reasoning

In this paragraph we will try to give a reading of counterfactual reasoning that could define it as an instance of contextual reasoning. This will be done in order to legitimate the use that we intend to make of MultiContext Systems as a paradigm inside which to develop the analysis of counterfactual reasoning.

Our first step will consist then in showing how we can find the three main features of contexts in what, from now on, we will call counterfactual context:

- it is *partial*: the agent performing a counterfactual reasoning is only interested in relevant information related to the counterfactual premise and this information is only a subset of the global knowledge base of the agent;
- it is *approximate*: the level of detail can be dynamically varied in the course of a counterfactual reasoning and these variations influence the outcome of the reasoning;

– it is *perspectival*: the centrality of the epistemic perspective of the reasoning agent can be deduced from the fact that different agents can reach different counterfactual conclusions starting from the same situation. The features of the counterfactual context built by an agent are strictly dependent on her set of beliefs about the "factual" situation.

Moreover, the two notions of locality and compatibility (the basis of every contextual reasoning) are crucial in the definition of a counterfactual reasoning:

– the core of the counterfactual reasoning is *local*, because it is performed entirely inside the countefactual context;
– but if we consider the whole process of counterfactual reasoning, it switches from the counterfactual context - that is defined by the counterfactual premise - and the working context, which is precisely what the agent performing the reasoning is actually interested in. For this reason, working context and counterfactual context must be *compatible* and all the assumptions that are relevant for the subject of reasoning (except the counterfactual premise) must be imported from the working context to the counterfactual one.

6 Counterfactual Reasoning and Game Theory

After having sketched the main features of the treatment of counterfactuals under a contextual theory, we want to give a hint of some possible conections with another theory emerged in a different discipline.

Another framework in which counterfactual reasoning can demonstrate its importance is a theory developed in economics, that has been widely applied and it is called Game Theory.

We will show how a lot of decision processes in Game Theory could be based upon a counterfactual reasoning and, moreover, we will try to compare the Game Theory framework with the one provided by MultiContext Systems.

When confronted with a decision about her future move in a game, an agent (or player) must consider the previous history of the game to try to guess which the future actions of her opponents could be.

All these considerations about the past history of the game can be used to build a "profile" of other agents. If the strategies played by an agent in the past are useful to reconstruct her profile, we cannot deny that the strategies that the same agent has decided not to follow are nearly as important.

The reasons why an agent has decided to reject the choice of a certain strategy can be very useful in the prediction of which strategies she will accept or reject in the future.

The importance of counterfactual reasoning is even greater in those cases in which the game is of imperfect information (something in the past history of the game is not common knowledge).

In such cases, the agent with the move is not perfectly aware of the precise situation she is in (this state is called in game theory information set, because a

set of "situations" are available to the agent and she has to guess which is the one she is in on the basis of the lacking information she possesses).

When a part of the game is uncertain, the importance of the strategies (realized or not) of which the agent is sure of, is increasingly high.

But predicting future actions of the opponents is not the only aim of counterfactual reasoning: the inquiry about the credibility, attitudes, beliefs, intentions of other agents has also the purpose of understanding which of them are fit to cooperate or, to use a more technical locution, to enter into a coalition.

As we have anticipated, the notion of information set (sets of possibilities determined by the lack of information) is very close to the notion of context as partial theory. Being partial, a context is a set of models representing "the possible ways the situation can be".

The use of contexts to represent information sets is promising for another reason: very often (if not in every situation) an agent has to consider not only the opinions and beliefs of other agents about the game, but also the opinions and beliefs these agents have toward her (and what they think she thinks of them and so on).

If we use a formalization based on MultiContext Systems, we have at our disposal all the tools developed within it to switch from one context (representation of the game of an agent) to another.

7 Conclusions

What we have tried to do with this paper is to give a preliminary and intuitive account of a cognitive approach to the subject of counterfactuality.

In doing so, we followed some intuitions coming from cognitive sciences and artificial intelligence, which make evident some points of distinction with the "traditional" works on counterfactuals developed by most philosophers in the study of the subject.

As we have shown, the primary difference has to be found in the goal: while these philosophers have concentrated their analyses on the semantics of counterefactual conditionals (what we have called the *metaphysical approach*), we are mainly interested in the way in which processes of reasoning having a counterfactual dimension develop in human or artificial "minds".

Another topic that we want to deepen is the one of the possible applications of the "counterfactual structure" to other forms of common sense reasoning.

Nevertheless, there are a series of points that are still open, the most important of which is the elaboration of a formal model able to illustrate and integrate counterfactual reasoning.

In this direction, our purpose is to apply the tools of Multicontext Systems and Local Models Semantics with their principles of locality and compatibility, that seem to fit the features of this peculiar kind of reasoning.

References

1. M. Benerecetti, P. Bouquet, and C. Ghidini. Contextual Reasoning Distilled. *Journal of Theoretical and Experimental Artificial Intelligence*, 12(3):279–305, 2000.
2. J. Bennett. Counterfactuals and temporal direction. *The Philosophical Review*, XCIII(1):57–91, January 1984.
3. R. Byrne and A. McEleny. Counterfactual thinking about actions. In P. Cherubini, editor, *Human Reasoning: Logical and Psychological Perspectives*. 1999.
4. 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/`.
5. J. Dinsmore. *Partitioned representations*. Kluwer Academic Publishers, 1991.
6. J. Dinsmore. Mental spaces from a functional perspective. *Cognitive Science*, 1987.
7. G. Fauconnier. *Mental spaces: aspects of meaning construction in natural language*. MIT Press, 1985.
8. G. Fauconnier. Analogical counterfactuals. In G. Fauconnier and E. Sweetser, editors, *Spaces, Worlds, and Grammar*, pages 1–28. The University of Chicago Press, 1996.
9. M. L. Ginsberg. Counterfactuals. *Artificial Intelligence*, 1986.
10. F. Giunchiglia. Contextual reasoning. *Epistemologia, special issue on I Linguaggi e le Macchine*, XVI:345–364, 1993.
11. F. Giunchiglia and P. Bouquet. Introduction to contextual reasoning. An Artificial Intelligence perspective. In B. Kokinov, editor, *Perspectives on Cognitive Science*, volume 3, pages 138–159. NBU Press, Sofia, 1997.
12. F. Giunchiglia and C. Ghidini. Local Models Semantics, or Contextual Reasoning = Locality + Compatibility. In *Proceedings of the Sixth International Conference on Principles of Knowledge Representation and Reasoning (KR'98)*, pages 282–289, Trento, 1998. Morgan Kaufmann.
13. N. Goodman. The problem of counterfactual conditionals. In *Conditionals*.
14. J. Y. Halpern. Hypothetical knowledge and counterfactual reasoning. *International Journal of Game Theory*, 28, 1999.
15. S. J. Hoch. Counterfactual reasoning and accuracy in predicting personal events. *Journal of Experimental Psychology*, 11(4):719–731, 1985.
16. F. Jackson, editor. *Conditionals*. Oxford Readings in Philosophy. Oxford University Press, 1991.
17. C. Ortiz Jr. Explanatory update theory: Applications of counterfactual reasoning to causation. *AI*, 1999.
18. G. Lakoff. Sorry, I'm not myself today: The metaphor system for conceptualizing the self. In G. Fauconnier and E. Sweetser, editors, *Spaces, Worlds, and Grammar*, pages 1–28. The University of Chicago Press, 1996.
19. M. Lange. Inductive confirmation, counterfactual conditionals, and laws of nature. *Philosophical Studies*, 1997.
20. D. Lewis. *Counterfactuals*. Blackwell, 1973.
21. D. Lewis. *Philosophical papers*. Oxford University Press, 1983. Two volumes.
22. M. G. Lipe. Counterfactual reasoning as a framework for attribution theories. *Psychological Bulletin*, 109(3):456–471, 1991.
23. M. McDermott. Counterfactuals and Access Point. *Mind*, 1999.
24. P. Noordhof. Probabilistic Causation, Preemption and Counterfactuals. *Mind*, 1999.
25. R. Stalnaker. A Theory of Conditionals. In F. Jackson, editor, *Conditionals*, Oxford Readings in Philosophy, pages 28–45. Oxford University Press, 1991.