What does Game Theory have to do with Plans?

Olivier Roy
ILLC, Universiteit van Amsterdam
oroy@science.uva.nl

July 12, 2005

Abstract

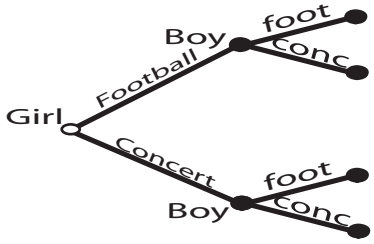
The belief-desire model underlying game-theoretical (GT) analysis of strategic interaction has been criticized by philosopher of action, such as Michael Bratman, who stressed that it cannot account for the role played by intentions and plans. How should GT react to this? I will argue that it already has what it takes to cope with intentions and plans of *ideal* agents: the concept of strategy. On the other hand, I will try to show that plans and intentions can play a key role in game-theoretical modeling of cognitively bounded agents because, for such agents, the task of *constructing* a decision problem is as important as finding what is the best move to play.

1 Introduction

Decision and game theory[7] (GT) are formal models of rational agency. They both explain the behavior of rational agents in terms of their beliefs and preferences. This *belief-desire* conception of rational agency has been criticized by Michael Bratman [1], [2]. He has argued that it cannot account for intentions and plans, two essential characteristics of human agency. Does it means that something fundamental is missing in GT? In other words, does GT have something to do with plans and intentions? In this paper, I will explain why I think that the answer to this question depends on the type of agent that GT tries to model. I will argue, first, that plans of *ideal* agents boils down to something that is already present in GT: strategies. Thus, from that perspective, one would not get much new out of GT by enriching it with plans and intentions. But the focus on ideal agency is by no means forced on GT by its own tools. Moreover, it seems that plans and intentions get their full *raison d'être* for bounded agents : they help us to simplify entangled decision problems by filtering the available options. Thus, plans not only play a part in the decision process itself, they are active vectors in the shaping of decision problems for limited agents. I will argue in the second part of the paper that to shift attention from optimal strategies *given* a decision problem to optimal deliberation strategies that *leads* to the formation of manageable decision problems can prove to be a fertile move for the analysis of bounded rationality. But before all this, I will provide a sketchy and informal account of the GT models of agency, and of Bratman's conception of intentions and plans.

2 Game theoretical models : an informal overview

In GT, interaction situations are modeled either in *strategic* or *extensive* forms. A game in extensive form is a tree, where each node represents a decision point for a player, and the edges spreading from a node are the possible actions that the player has at that node. The game starts at the root of the tree and each leave represent an outcome of the game. Let's take an example[1] :



Here, a couple has to decide what they will do in the evening. They have two options: go to the concert or to the football match. The lady decides first; her decision point is at the root and the two edges represent her options. Then the man decides what he will do.

A move for a player at a decision point consists simply in the choice of one of his available actions. A *strategy* for a player is a complete set of moves: it specifies what to do at each of his decision point. I will come back in more details on the concept of strategy in section 3[2]. A *play* of the game is a path from the root to one of the leave. A play can be viewed as induced by a *strategy profile*, which is a combination of strategies, one for each player.

Extensive game forms are detailed models of the game situation. *Strategic* game forms abstract from the particular moves of each player, and considers full strategies as the object of choice. Thus, games in strategic form can be represented as matrices:

↓Boy, Girl →	Football	Concert
Football	Football together	Boy football, girl concert
Concert	Boy concert, girl football	Concert together.

In what follows, I will mainly focus on extensive game forms. Yet, I think that most of what I will argue for applies to strategic representations as well.

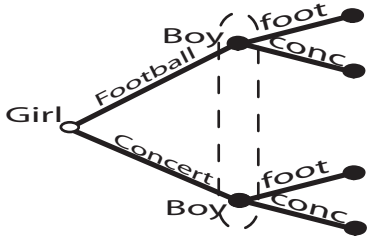
So far, we have models that allow us to represent the different options available to the players. But we need more than options to play a game. We also need a specification

[1] This example is a simplified version of a well-known example in game theory named "battle of the sexes".

[2] In this paper I will restrict myself to what is called "pure strategies". *Mixed* strategies are randomization over pure strategies. I think that my argument applies to mixed strategies has well.

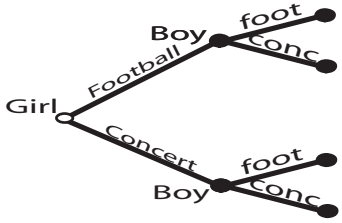
of the *information* the players have as the game unfolds, and we need to know what are the *preferred* outcomes for each player.

To account for the availability of information, extensive games are equipped with *information sets*. Intuitively, an information set contains the nodes that a player cannot distinguish. To see how it works, suppose that, in our example, the man has to choose what he will do without knowing what his partner has chosen before. He cannot distinguish the situation where he chose after a "football" choice by the girl, upmost node, from the situation where he choose after a "concert" choice from the girl. Those two nodes are in the same information set (the dashed box in the image below). On the other hand, he knows that when time comes for him to choose, the girl had chosen before, and so the root is not in the same information set.



Games where there is no undistinguishable nodes are called *games of perfect information*. In those games, all information sets are singletons. Games of *imperfect* information are games where there is some uncertainty, games where some information sets contains more than one node. It is important to see that the distinction between perfect and imperfect information is *not* intended to represent the player's subjective limitations of memory or failure to notice certain salient characteristics of decision nodes. If two nodes are in the same information set for one player, this is not because he is not smart enough to distinguish them. Rather, what is encoded by information sets is the information structurally available in the game.

The last components we need is preference over the outcomes. It is usually represented by "values of utility" assigned, for each player, to the leaves of the game tree. In our example, we can suppose that they both prefers to go to the match, and they also prefers to go together than alone. Those preferences are indicated between brackets in the figure, the first value being for the girl, the second for the boy.



(2, 2)

(0, 0)

(0, 0)

(1, 1)

Intuitively, an outcome gets a higher utility value than another if the former is preferred to the latter. Note the order of the implication: outcomes get higher utility *because* they are preferred, they are not preferred because they have a higher utility. Although the preference ordering used in extensive games can be seen are representing subjective preference, there are no assumptions about how the players came to have such preferences. It can be because of personal inclinations, cultural bias, moral or religious conviction, etc.

This terminates our short tour of game models. We have, so far, a model of the players' options, the information they have and their preferences over the outcomes. But, GT is not just here to build models for games. It gives insights about what the players will/should do when they face a given interaction situation. This is where the *solution concepts* for games come into the picture. Among the most famous ones are the *Nash equilibrium* and the *backward induction* algorithm. Both rest on the concept of *rationality*.

"Rationality", as used in GT, is often called "instrumental rationality". Roughly, the idea is that a rational agent is an agent who tries to reach the best feasible outcome, given what he believes and according to his preferences. This is the well-known conception of rationality in terms of maximization of expected utility.

Let's examine how this works through the backward induction solution concept. The idea here is that each player finds out what his best move is from what he expects the others to play after each of his moves. In our example, suppose we try to see what is the best move for the girl. The backward induction algorithm tells her to think about what will happen if she chooses "football". Then the boy will also go for "football", because he prefers to go out with her. Similarly, if she chooses "concert" she expects the boy to chose also "concert". So, if she chooses "football" she will go out with the boy to the football match, and if she chooses "concert" she will go to the concert with the boy. As she prefers going to the match than to the concept, the best move for her is

"football". One can see why this solution concept is called "backward induction": the players find out what to do at a decision point by reasoning backward from the leaves along to subtree starting at that decision point.

Although I didn't mentioned it explicitly, this reasoning clearly rested on the beliefs of the girl. First, she used information about the *rationality* of the boy. That is, she assumed that he would choose his preferred option. But, she also used the fact that the boy would know what she has chosen. In this simple case, this led her to the conclusion that he would copy her move. In the imperfect information variant of this game, things get more complicated. The boy's beliefs[3] about what the girl has chosen would play a crucial role in computing his optimal strategy.

I will not push further the presentation of GT solution concepts here[4]. The point that I wanted to make should be clear by now: GT analyze of strategic interaction is grounded in a belief-desire model of agency. That is, rational GT agency is function of the preferences and the beliefs of the players.

3 Bratman on intentions and plans.

As I mentioned in the introduction, Michael Bratman has argued that the belief-desire model cannot account for plans and intentions. So let's leave GT for a moment, and take a look at Bratman's ideas.

Bratman's perspective is functionalist; intentions and plans are mental states characterized by their function. Traditionally[5], in philosophy of action, intentions are conceived as *compounds* of beliefs and desires.[6] Bratman rejects this view. He thinks that intentions are irreducible to beliefs-desires compounds.

The two keywords to Bratman's functional definition of intentions are *commitment* and *stability*. Intentions are (relatively) stable mental states that commit agents both in action and in deliberation. The stability of intention is the first thing that distinguishes them from desire. Typically, what we like and want change a lot, while we tend to stick to our intentions once we have adopted them. This stability is accountable for the *reason-centered commitment* of intention. Once an agent has the intention to achieve something, he will only consider options that are compatible with this intention when deliberating about other actions[7]. But intentions also function, in the first place, as guides to action. This is what Bratman call their *volitive commitment*; an agent with a certain intention will act to carry it out. Bratman thinks that this action-oriented committing force of intentions is much stronger than that of desires. Finally, it worth noting that intentions are subjects of consistency constraints from which desires are exempt. Contradictory desires are common while contradictory intentions are simply irrational.

Intentions are the building blocks of plans, for which Bratman distinguishes two

[3]Represented by a probability distribution over the nodes of the information set.

[4]I leave aside the famous Nash equilibrium.

[5]The canonical example being Davidson's [3, Chap.1]

[6]The distinction between intention *in action*[9] and future-directed intention is also assumed, the latter being explained in terms of the former.

[7]As long has he doesn't reconsider his intention, of course.

concepts. First, plans can be viewed as "abstract structures of sort that can be represented by some GT notations"[1, p.28]. Second, they can be viewed as "sets of intentions about future action". I will focus on the second view here, the argument that I will develop in section 4 being precisely that, under the "ideal agent" paradigm, they are equivalent.

The two main features of plans viewed as sets of intentions are their *hierarchical structure* that compensate for their *incompleteness*. A plan contains general intentions, for example getting my Ph.d, upon which more specifics intentions are subordinated, for example choosing a thesis topic and a supervisor. This hierarchical structure prevents us from having to settle in advance every single detail of the plan, which would be an impossible task for limited agents like us. In that sense, plans are typically incomplete. They are, in a nutshell, "intentions writ large" [1, p.29]. They thus share, more or less, the same functions as intentions. However, because of their more general character, they must meet new consistency constraints to fulfill those functions.

Endogenous consistency require the intentions they are made of to be coherent. But those intentions have also to be *exogenously* consistent with what the agent believes and takes for granted. That is, if this cognitive background happen to be the case, the plan should be executable. Finally, it should be possible to extend consistently partial plans to more precise ones by adding new sub-intentions about means to reach the intended end. This is what Bratman call *means-end* consistency.

This sketchy account of plans and intentions already contains part of Bratman's argument for their importance to a theory of rational agency. Because of their stability and deliberative commitment, intentions are not only the outcome of practical reasoning, they take active part in it. They also impose strong consistency constraints for agents who persist through time. Along the same lines, Bratman has argued that intentions provide a key tool for self-identification. Plans, as we just saw, are extremely useful for limited agents. But, just as intentions do on a more limited scale, they also impose strong constraint on deliberations. All those features of intention and plans show how important they are from the perspective the individual rationality. But, their role in interactive situations cannot be ignored. Plans and intentions are prime vectors of *coordination* and *cooperation* between rational agents.

So, if intentions and plans are so important for the analysis of rational decision making, why is there no mention of them in GT? Can GT do without them, or does it miss something vital? In the next two sections, I will address those questions. I will argue, first, that GT can do without intentions and plans as long as it stays to the abstract level of ideal rational agency. I then sketch a way to re-introduce those notions in order to model rational interactions of limited, that is, non-ideal agents.

4 Plans and strategies for ideal agents.

4.1 Ideal agency

The example of backward induction presented in section 2 implicitly assumed that our agents were rational players with certain cognitive and computational capacities. To appreciate better the sort of analysis GT provides, and to finalize setting the stage for

my argument, let's return to GT and make those assumptions explicit.

From its foundation by von Neumann and Morgenstern[6], a large part of GT operated under the *ideally rational agent* paradigm[8]. This view of the player is characterized, as Roger Myerson[5] puts it, by assumptions about their *intelligence* and *rationality*.

The assumptions about the players' intelligence regard their capacity to represent the game situation and to compute the appropriate strategy. An ideal agent is viewed as having unlimited cognitive capacities. He can represent any game situation as precisely as required, no matter how big and complicated this representation gets. In the analysis of simple games, such as our example of section 2, this assumption is not so involved. But one can appreciate its importance when applied to more complicated games such as chess. An ideal agent is also equipped with unlimited computational power. That is, no calculation of the optimal strategy is too difficult for him. Not only that, but he can perform any such calculation without time or energy cost. Real life agents, or even quite powerful computers, can solve very difficult optimization problems if they "sit down and think about it". But such calculations require time and effort. To idealize the agents in GT means to abstract from such complications concerning the time/energy costs. [9]

The other type of assumption made about ideal agents is that they are flawless utility maximizers. This means two things. First, that such agents have total preference ordering on every possible outcome of the game. If we take any pair of outcome, the players is always able to tell which one he prefers. Futhermore, the preferences of the players are assumed to be transitive and acyclic. This is indeed a very strong idealization, given what we just said about the cognitive unboundedness of such agents. Second, ideally rational players always take up the action that is the best outcome, given what they believe about the game and the other players. Unlike most of us, they do not suffer from apparent irrationality or "weakness of the will". So, the assumption of ideal rationality boils down to saying that the players always do what is best for them.

The ideal agent paradigm was intended to simplify GT analysis. Intelligence assumptions avoid complications related to limited representational capacities. It also removes the need to take into account the computational time/energy cost. On the other hand, rationality assumption rules out difficulties generated by intransitive or cyclic preference. It also makes it easier to model players' *expectations* about each other. We already came across such simplifying effect when we computed the optimal move for the girl on our example in section 2. Thus, working with ideal agents really reduces GT analysis to the computation of optimal strategies in interaction situation. It is now time to explain why I think that, for such ideal agents, the concepts of plans and strategies are equivalent.

[8] I leave aside evolutionary game theory, for it is an open question to me whether the argument applies to it as well. Note that attempt to move out of the ideal agency paradigm, while continue to use non-evolutionary GT tools exist, see [8].

[9] Note that assuming ideal intelligence doesn't mean eliminating uncertainty in games. An ideal agent can play a game of imperfect information. As I have already mentioned, the uncertainty aspect of those games is not the player's faults, but rather a consequence of the game structure. Being ideally intelligent in such game typically requires more powerful minds, as the computations and representations get more complicated.

4.2 Strategies and plans for ideal agents

To show that plans of ideal agents are just GT strategies, I have to show that they are functionally equivalent. I will proceed "top-down", working my way from plans to intentions.

First, recall that a strategy for a player in an extensive game is can be viewed as a function that assigns to every decision point where the player has to move an action to take. It is *complete* in the sense that every eventuality is covered. "Every" has to be understood in a very strong sense here: a strategy specifies what to do even at decision points that are excluded by the strategy itself. Hence, there may be "superfluous instructions" in a strategy. I should be clear that, because of the "over-completeness" of strategies, once one is chosen by a player, he doesn't have to fill the details as the game unfolds. He just has to follow mechanically what it tells him to do at each decision point (or information set) that he encounters.

Let's start the comparison between plans and strategies by noticing that both share a basic function: the volitive commitment. Once they are adopted, they "control" what the agent will do in the sense that the agent will act in accordance with his plan/strategy.

Now there is already a striking difference between plans and strategies: plans, unlike strategies, are incomplete. I don't think that this difference is of any importance for ideal agents. First recall that, according to Bratman, plans are incomplete precisely because we don't have the resources to settle in advance a fully detailed plan of action. But, it is assumed that an ideal agent has all the cognitive and computational resources to build such a detailed plan. So we cannot argue for the importance of incompleteness on the ground of cognitive capacities for ideal agents. One can try to point out that a complete strategy almost always includes some superfluous instruction, and thus that, even for an ideal agent, it can be better to restrict himself to an incomplete but "detailed enough" plan: a plan that specifies what to do at every possible move, leaving aside nodes that cannot be reached by the plan. But in what sense are such plans better? As GT ignore the costs of computing a full strategy, it is surely not better because it saves precious time and energy. More importantly, I think that a partial but detailed enough plan has to fulfill the same role than a complete strategy with superfluous instruction. Both will specify what the player has to do in every decision point that he may encounter. A less detailed plan would lack instruction that an ideal agent could have effortlessly computed before, and an over-detailed strategy will include some extra instructions that could be computed and "stored" without effort. So, narrowed to their *effective* role in games for ideal agents, I think that incomplete plans and complete strategies are equivalent.

This equivalence partly takes care of another apparent difference between plans and strategies: the hierarchical structure of the former. Recall that the main function of such structures was precisely to make it possible to specify general intentions with more precise ones, when needed. If there is no reason for an ideal agent to leave details unspecified, the hierarchical aspect of plans seems rather useless. But, one can argue that the function of a hierarchically ordered set of the intention is not only to leave room for sub-intention, but also to encode a kind of priority ordering between those intentions. The more general intentions are related to more deeply desired objectives, while very "low" mean-related intentions only rely on fugacious and volatile desires.

If this function is really a function of hierarchically structured plans, I think that it is completely fulfilled, for ideal agents, by the very abstract concept of a preference encoded in the utility values. Recall that the utility scale "sums up", so to speak, all the evaluative considerations that an agent may have relative to an outcome. True, for a limited agent, it can be an impossible task to come to such a definitive evaluation, given all possible preference shifts and unexpected situations that one may encounter. But, again, there is no such concern relative to ideal agents. No matter how complicated it can be to come up with a complete utility rating of outcome, an ideal agent can do it and, similarly as for the building of a complete strategy, I don't see why he wouldn't. So, again, this function of hierarchical organization is rather a useless device for an ideal agent.

Now let's turn to the consistency constraint imposed on plans, and see whether we have equivalent constraints on the side of strategies. I've mentioned, following Bratman, endogenous, exogenous and mean-end consistency. Let's begin with the latter, for it is the easiest to deal with. Mean-end consistency means that a partial plan has to be consistently extendable to a more detailed one. But, as partial plans are of no use for an ideal agent, this constraint is trivially satisfied. In other word, strategies are already maximally extended plans, and thus the mean-end constraint would stay *lettre morte*. Let's now turn to exogenous consistency. This constraint states that a plan should be executable if the beliefs and facts took for granted that form its cognitive background would happen to be the case. An agent should not settle for what he thinks would be an impossible plan. Again, this constraint is ruled out from the start for ideal agents. Our ideal agents cannot "forget" about one fact of the game when building a strategy, nor can they miscalculate the consequences of what they know and believe. Hence, they never come up with such an impossible strategy or, conversely, their strategies are always exogenously consistent. The last consistency constraint is more related to intentions than to plans as whole. Endogenous consistency means that there should not be contradictory intentions in a plan. There are two things to notice here. First, if we assume that the hierarchical structure of plans is useless for an ideal agent that means that we can consider all the intentions of his plans on the same "level" of generality. As it turns out, plans for ideal agents seem to spelled in full detail, and so the intentions they are made of specify what action to take into a single situation. This is just what happens with strategies, where acting according to a strategy at a decision point can be seen as doing what was intended at that decision point. But, in that context, what would be two "inconsistent intentions"? Simply, two different instructions for a same decision point. So, for ideal agents, the consistency constraint imposed by plans on their own component boils down to avoiding two different instructions for a single situation. But clearly game-theoretical strategies fulfill this function, simply because the "one instruction per node" constraint is built-in in them: they are *functions* from decision nodes to action possible at that node. It shows, I think, that strategies and plans for ideal agents are functionally equivalent on that respect.

This last consistency constraint on plans has slightly moved us toward intentions. Let me finish this move, and the argument as a whole, by turning to the functions of intentions. One of them is clearly fulfilled by strategies: the control of actions. It remains to check what happens with the reasoning commitment of intentions. As Bratman stresses repeatedly, intentions are stable and thus can function as filters for

further deliberations by ruling out options that contradict them. Again, I think that this function is of no use for an ideal agent, and so, for two reasons. First, as we saw, the completeness of strategies renders further deliberations useless. Once an agent has decided in favor of such a strategy, he just has to switch on the auto-pilot, so to speak. So the "filtration" function of intention is not called for help for ideal agents. Second, the "stable" character of intentions is a clear feature of strategies. There is no question of reconsidering the choice of a strategy in game theory, mainly because it is assumed that every calculation that an agent could make as the game unfolds he could make it before committing himself to a strategy. So, strategies are even more stable than intentions. Are they too stable? I don't think so, again because we deal with ideal agents. Bratman argues that a radical "non-reconsiderator" would simply be an irrational stubborn, because he would stick to his intention even in the face of the most important new information. But, has we saw, this idea of groundbreaking new information is ruled out by our conception of ideal agent. If there is any information to be available at all in the game, the agent is assumed to be able to take it into account before the game starts. So, in such context, I think that strategies are as stable as intentions of ideal agent should be.

This terminates my argument. Let me wrap it up. I've argued that plans for ideal agents are just game-theoretical strategies. To do so, I've tried to show that they fulfill the same functions of controlling action, imposing endogenous consistency, and that they are equivalently stable. On the other hand, I have tried to show that the other specific functions and characteristics of plans apparently not fulfilled by strategies are simply non-starters for ideal agents. Those were incompleteness, hierarchical organization, along with exogenous and mean-end consistency. This, I think, shows that plans and strategies are functionally equivalent for ideal agents, and thus that, at that level of analysis, GT does not need to be enriched with intentions and plans. But, as I have noted in the introduction, the ideal agent paradigm is not at all enforced by game-theoretical tools. They can also be used to analyze limited or bounded agents. In the next section, I stress a suggestion of Bratman about how intentions and plan can play a role in the modeling of such bounded agents.

5 Plans, intentions and deliberative strategies

The idea I want to develop here is that of *deliberative strategies*, a process by which agents go from complicated decision problems to simple ones, and in which specific functions of plans are of prime importance. Indeed, I will not present a full theory of deliberative strategies here. I will rather give a sketchy account of the main ideas related to this conception of bounded rationality, focusing on the place of plans. But before that, I should make clear what kind of bounded agent I have in mind.

5.1 Bounded Agents

There are many ways to depart from the ideal agent paradigm. The agents that I will talk about in what follows are bounded in two aspects.

First, they have limited cognitive and computational capacities. That is, I will assume that there is a maximum to the size of the game model that they can represent. Similarly, our non-ideal agents are not capable of performing arbitrarily complex calculations. I will not, however, be explicit about the extend of those limits. It is enough to assume that they exist and are roughly fixed.

Second, I will suppose that the agents' preferences are not always total, which means that some alternatives may be incomparable. I will not assume that an agent may have intransitives or cyclic preferences. As we will see, one of the goals of deliberative strategies will be to come up with a set of options upon which agents have a total preference ordering. In that context, including intransitive or acyclic preferences would only bring worthless complications.

So, by bounded agent, I mean agents with partial preference ordering and limited cognitive and computational capacities. Note that I *still* conceive of agents as utility maximizers, although it seems to be one of the most criticized idealization of GT. The conception of bounded rationality I have in mind is one where, once they have zoomed to simple and manageable problems, agents do try to reach what seems to be the best outcome. I will come back to this picture of bounded rationality in the concluding remarks of this section.

5.2 Deliberative commitment of intentions revisited

Recall that one of the main functions of intentions identified by Bratman is the *deliberative commitment*, inherited from their stability. Traditionally, intentions are seen as the output of practical reasoning: after comparing different outcomes according to their preferences, agents form the intention to do what has to be done to get what they prefer. What Bratman has called to attention with his idea of deliberative commitment is that intentions are also important *inputs* in practical reasoning. A prior plan acts as *filter*, so to speak, which rules out options that are incompatible with the intentions it contains. As we saw in the previous section, this function of intentions isn't of much use for an ideal agent. But, it is quite important for an agent with limited cognitive capacities. Intuitively, considering every single possible action seems to be a painful and, in the end, quite worthless enterprise. It seems much wiser for agent with limited capacities to rule out options, before even starting comparing them, to focus on what really matters.

But, obviously, ruling out options is different from optimizing expected utility. Finding what is the best action to take *given* a set of options is one thing, *coming out* with an admissible set of option is another. While GT is mostly concerned with the former activity, the deliberative commitment of intention points toward a kind of pre-processing of decision problems. That is, it suggests an active process *prior* to the decision making, where agents "build up", so to speak, the situation they face. As Bratman puts it, intentions and plans "help to answer a question that tends to be unasked within traditional decision theory, namely : where do decision problems come from"[1, p.33].

Let's return to the example of section 2 to see what's going on here. We had two agents who had to decide how to spend their evening. They had two options, go to a concert or to a football game, and each knew what they preferred. The optimal

solution was quite easily found, using the backward induction algorithm, for such a simple decision problem. So far, so good: if all decision situations where as simple as that, life would be quite easy. But, how come it is so simple? One possible answer is that we were focusing on a toy example that is simple for the reason that this is how we built it. Another is to see this decision problem as the *result* of another deliberation process in which many irrelevant and incompatible options have been trimmed. Let's retell the story along this second perspective.

Try, first, to list *everything* that our agents could have done that evening. The girl could have stayed home to work, called other friends and go out with them, gone to the grocery store, visit her mother, etc. This list grows almost indefinitely, and the same happens with the boy's options. Generally, trying to come out with a complete list of every possible action soon proves to be a doomed enterprise. Our world is replete with possibilities, probably with more than our limited cognitive capacities can represent. So, to come up with the simplified version of the "what to do this evening" problem, both the boy and the girl must have done some work on their coarse, and maybe open, set of options.

How can we characterize this work? Among other things, probably by focusing on their prior intentions. We can assume that both had, for instance, the intention *not* to stay home, with the subordinate intention to go to the concert or the football game. This alone rules out a huge bulk of options: stay home, indeed, but also visit the mother and go to the grocery store. But even then, there are almost indefinitely many ways to fulfill the intention to go to the concert or the football game: go by tram or by bike, wearing the blue or the red shirt, arriving early or not, and so on. Most of those admissible courses of actions are, indeed, irrelevant variations on what seems to be the two obvious candidates to deliberate upon: football or concert. The specification level of those "two obvious candidates" correspond quite sharply to the specification level of the "filtering" intention: going to the concert wearing a blue shirt and going to the concert wearing a red shirt seems to fall into the same "equivalence class modulo the intention", so to speak. Once incompatible options have been ruled out, and compatible ones have been classified into equivalent sets of alternatives, what remains seems to be the simple decision problem that we started with in section 2. Each player has retained two classes of courses of action, those that boil down to go to the concert and those that boil down to go to the game. In this simplified decision context, it seems to me more plausible to assume that they have a total preference order upon which they will choose what seems to be the best option.

So, looking at how intentions act as input in practical reasoning leads to an examination of how agents shape the decision problem they are facing. The analysis of our example suggests how this shaping takes place, by trimming and regrouping options. Let's now take a step back from the example to get a more general picture of the situation.

5.3 Deliberative strategies

First, let me give a very general definition: I will call a *deliberative strategy* any process that takes a set of options as input and gives another (not necessarily different) set of options as output. The name "deliberative strategies" is adapted from what Bratman

has called "strategies for reasoning"[2, p.21][10]. As I shall explain briefly in the concluding remarks of this section, such processes are called "strategies" on purpose: like strategies in game theory, they seem to be subject to *rationality standards*.

This definition is purposely general. It is meant to encompass the widest possible scale of deliberative strategies. For instance, it includes the filtering process described in our example, but also, the trivial process that gives back the same set of options that it has been fed with, the "complicating" processes that return more options and the "over-zooming" processes that return only one. What matters is that among all those possible deliberative strategies some are of interest for bounded agents.

In the first place, agents with limited cognitive capacities will probably use deliberative strategies that return sets of options not bigger than they can handle. If, for example, an agent cannot handle more than three or four different options[11], a deliberative strategy that returns dozens of alternatives will more a burden than a tool for him[12]. Along the same lines, an agent with partial preference ordering will take advantage of a deliberative strategy that returns a set of comparable options. Deliberative strategies exhibiting those two features would surely be of use for a bounded agent, and thus it worth trying to characterize them more precisely. As the example of the previous subsection suggests, plans and intentions give valuable insights for this characterization.

Remember that plans have a hierarchical structure and that they are typically incomplete; they consist of general intentions upon which more specific ones are subordinated. But, the latter intentions leave unspecified many important details which will have to be filled as the execution of the plan goes along. We can see a moment of decision as the moment where an agent have to fill a general plan. Let's say that a deliberative strategy *complies with a plan* if, first, it doesn't return options that are incompatible with the fulfillment of the plan and, second, it regroups in the same family options that are equivalent up to the specification level of the intentions that have to be filled at the decision point. In other words, a deliberative strategy complies with a plan if, first, it eliminates actions that would go against the plan and, second, if it doesn't bother with details, so to speak.

Is complying with a plan a sufficient condition[13] to lower the quantity of options under the maximum that a bounded agent can handle? Although the answer seems to be "no" sometimes, I think we can conjecture that most of the time it will prove to be

[10] I use "deliberative strategies" instead of "strategies for reasoning" only to put the emphasis on the practical reasoning flavor of such processes. Indeed, I could also have used "strategies for practical reasoning". I simply go for the shortest option.

[11] Suppose, for example, that he gets systematically lost when computing optimal strategies for more than three options, or that he just cannot "store" decision problem exceeding that size.

[12] It may seem that this conception of deliberative strategy suffers from a fatal flaw from the start, for it apparently supposes that the agent has a representation of his full set of options, the size of which is much larger than anything he could handle. This is a serious objection that I will not try to overcome in full respect here. I will content myself with pointing out that there is no assumption made about whether the use of a deliberative is a fully conscious and intentional action, a partly (innate?) reflex, or a completely unconscious process. Maybe that a way to bypass the cognitive overload imposed by the original set of options. Another answer would be to focus on deliberative strategies that take very small sets of options as input and return bigger ones. The connections from those strategies to Reinhard Selten's[4, chap.2] models of bounded rationality should be studied.

[13] It is surely not a necessary condition, for we can easily imagine a deliberative strategy that always returns one option, "do nothing", despite what the prior plans of the agent are.

a sufficient condition. If a plan is very general, for example getting my Ph.D., most of the simplifying job of the deliberative strategy will consist in regrouping options in equivalence classes. In other word, the number of relevant alternative will be drastically reduced because the amount of detail to be taken into account will be typically low[14]. On the other hand, if a plan is very precise, then the ruling out of options will take care of most of the simplification. Overall, my conjecture is that the equilibrium between excluding and regrouping option, on the basis of the precision level of the plan, will keep the quantity of retained options reasonably low. What about getting a comparable set of options? Again, there is no guarantee that a complying deliberative strategy will return such a set but I think that, at least, chances are higher that an agent will be able to completely order a small set of alternatives.

So it seems that intentions and plans can play an important role in keeping decision problems manageable for bounded agents. Once the set of options is reduced below the size that such agents can handle, in such a way that its elements can be totally ordered, it makes sense to think that he will be in position to make a decision. Deliberative strategy seems to me to be a handy way to model this function of plans and intentions. But more has to be said about deliberative strategies to evaluate their real theoretical import for a theory of bounded agency. Let me conclude this section by indicating some important issues that will have to be addressed.

5.4 Deliberative strategies, ”taking for granted”, rationality, and utility maximizing

I want to make three remarks here. The first one points toward another aspect of practical reasoning that should be taken into account by studying deliberative strategies. The second emphasis that we should be able to formulate rationality criteria for deliberative strategies. Finally, I would like to reassess more explicitly the conception of bounded rationality adopted thorough this talk.

It is worth noting that plans and intentions are probably not the only things that can be used to characterize deliberative strategies. Again, I think that one of Bratman's idea could be put at work here: the concept of ”taking for granted”. In [2, chap.1], he has argued that practical reasoning involves taking some fact for granted. His typical example is taking the fact that it will not rain for granted when deliberating upon going to a bookstore by bike on lunch hour. In that case, the agent will decide what to do according to what he takes for granted, but if asked to bet on whether it will rain or not, he will probably not put all he has on the ”no rain” option. Along those lines, Bratman has argued that ”taking p for granted” and ”believing that p” are different mental states, although there should be a close connection between the two. The details of the argument is not important here. What matters is that what is taken for granted surely makes a big difference in how an agent sees a given decision problem. So, a theory of deliberative strategies should have something to say about how such mental states are fixed along the filtration. For example, some deliberative strategies may lead into taking more facts for granted, therefore reducing the structural uncertainty

[14]We should be able to test this conjecture with a formal model of the update induced by deliberative strategies.

of the output decision problem and thus simplifying the computation of the optimal strategies. Along the same lines, taking more facts for granted may help to obtain a total preference ordering. Just think how important are statements like "all things being equal" in preference contexts. So, obviously, the characterization of deliberative strategies should involve ways of taking for granted.

But, I think that a good theory of deliberative strategies should not be limited to characterization; it should also say something about the rationality of deliberative strategies. Take, for instance, an agent that use an "over-zooming" strategy of the kind I have mentioned earlier. For any decision problem, this agent would, say, always retain two options and no more than that. In many situations, one can imagine that this is not quite a fruitful strategy. Many valuable options would be left behind or blended in overly general equivalence classes. In other words, we would not say that this strategy is always a rational one. In general, I think that we should be able to formulate rationality criteria for deliberative strategies. How? One way that, first, comes to mind is to compare the optimal solutions in decisions problem that result from applying different deliberative strategies over the same set of outcome. If two deliberative strategies s_1 and s_2 give comparably complex decision problems d_1 and d_2, but in d_1 the optimal decision is far better then in d_2, I would be inclined to say that it is more rational for an agent to use s_1 then s_2. This criterion would surely have to be refined. The point I want to make here is only that rationality is a concept that applies to deliberative strategies as well as it does to strategies in GT[15].

The last point I want to make regards the conception of bounded rationality underlying this whole idea of deliberative strategies. The fact that I did not lift the utility maximizing assumption is of prime importance here. The view I take is that, once they have simplified decision problems according to a deliberative strategy, bounded agents do act as utility maximizer, generally in accordance with game and decision theory principles[16]. I am thus not arguing that shifting attention to deliberative strategies involves a rejection of GT results, quite the contrary. But, the last remark on rationality of deliberative strategies should also make clear that I don't want to commit myself to the claim that bounded agents are always rational. Maximizing utility in a decision problem obtained after the application of a bad deliberative strategy would typically lead to bad decisions. Rationality of a decision is indeed function of the beliefs and desire of the players, but it is also strongly influenced by how they conceive of the game they are playing.

6 Conclusion

The question driving this paper was: does game theory have something to do with plans? I have tried to provide a double answer. First, I have argued that the GT concept

[15] Bratman[1, p.45-49] has discussed the question of rationality of actions in a way that seems to be close to what I suggest here. He distinguishes *internal* and *external* rationality, the first but not the second one being plan-relative, so to speak. However, I think that there is an important difference: the criterion that I envisage is intended to assess the rationality of deliberative strategies and not only of the action that result from decision problem where they have been applied.

[16] This is indeed an empirical claim that should be tested.

of a strategy is functionally equivalent to the one of a plan at the abstract level of ideal agents. In other words that GT already models, although rather implicitly, plans of ideal agents. Second, I have tried to show that plans and intentions can play an important role in a game-theoretical analysis of bounded rationality, by helping to characterize how agents make complex decision problems manageable. The key word was deliberative strategies, a process that operates on the set of actions opened to an agent.

All this has been done informally, although GT involves quite a lot of mathematical machinery. I think that the idea about deliberative strategies sketched above should be formalized in extension to existing GT tools. In fact, the concluding remarks about rationality of deliberative strategies points toward the use of GT tools. This means that, not only GT has something to do with plans, but that philosophical theory of planning probably has also something to do with GT.

References

[1] Michael Bratman. *Intentions, Plans and Practical Reasons*. Harvard UP, Londre, 1987.

[2] Michael Bratman. *Faces of Intention; Selected Essays on Intention and Agency*. Cambridge UP, 1999.

[3] Donald Davidson. *Essays on Actions and Events*. Clarendon Press, Oxford, 1980.

[4] Gerd Gigerenzer and Reinhard Selten. *Bounded Rationality: The Adaptive Toolbox*. MIT Press, 2002.

[5] Roger B. Myerson. *Game Theory: Analysis of Conflict*. Harvard UP, 1997 edition, 1991.

[6] J. Von Neumann and O. Morgenstern. *A Theory of Games and Economic Behaviour*. Princeton University Press: Princeton, NJ, 1944.

[7] Martin J. Osborne and Ariel Rubinstein. *A Course in Game Theory*. MIT Press, 1994.

[8] Ariel Rubinstein. *Modeling Bounded Rationality*. MIT Press, 1998.

[9] John Searle. *Intentionality*. Cambridge UP, 1983.