Intentions and Plans in Decision and Game Theory

Martin van Hees
University of Groningen
Martin.van.Hees@rug.nl

Olivier Roy
University of Amsterdam
oroy@science.uva.nl

November 2, 2006

1 Introduction

Given the important role that intentions play in the way we make decisions, we would expect intentions to occupy a substantial place in any theory of action. Surprisingly enough, in what is perhaps the most influential theory of action, rational choice theory, explicit reference is made to actions, strategies, information, outcomes and preferences but not to intentions.[1] This is not to say that no attention has been paid to the relation between rational choice and intentions. On the contrary, a rich philosophical literature has developed on the relation between rationality and intentions (see for example [10]). However, to our knowledge, there has been no real attempt to model the role of intentions in decision making *within* a rational choice framework.

In this paper we argue that such modelling is a worthwhile enterprise. Starting from a very simplistic rational choice model, we show that enriching it with tools to represent intentions helps to account for known phenomena such as focal points, and gives rise to new questions about intention-based strategic interactions. We build our representation of intention on the philosophical foundations laid down in [2]. Our contribution is twofold. We first show that intentions can account for focal points in decisions and games. We then show how agents can use their intentions to simplify decision problems. In neither part do we go into the question whether intentions can be defined in terms of strategies, preferences, beliefs, or in any other ingredient of the existing models; we simply introduce intentions as an extra parameter and then examine some conditions that might be imposed on them.

[1] In what follows, we take the term "rational choice theory" as encompassing both "decision theory" and "game theory".

2 Rational Choice Theory

Rational choice theory refers to a huge collection of theories, approaches and models. Here we are only concerned with two of its branches: single-agent decision making and multi-agent strategic interactions, which are subjects of *decision* and *game* theory, respectively. Our framework is totally qualitative, i.e., it leaves aside probabilities and utilities. This is indeed a major simplification, for important results in these fields rest on properties of probability spaces and real-valued utility, e.g. the existence of equilibria in games. But, as we shall see, interesting issues readily arise in this simplified environment. Of course, we do hope for an integration of planning agency into full rational choice theory, and we consider the present work a first step towards it.

2.1 Models of decision situation

Let X and I be finite sets of outcomes and agents. Each element of X is a complete description of a state of affairs. Each individual $i \in I$ is assumed to have a preference ordering R_i over the outcomes in X that is complete (for all $x, y \in X$, either xR_iy or yR_ix) and transitive (for all $x, y, z \in X$, if xR_iy and yR_iz then xR_iz). xR_iy intended to mean "x is at least as good as y". The relation of "strict preference" is defined from R_i as follows: xP_iy if and only if xR_iy but not yR_ix. An element is called a *best element* if it is at least as good as any other. The set of best elements is denoted $C(R_i)$, i.e., $C(R_i) = \{x|\ xR_iy$ for all $y \in X\}$.

In a parametric (i.e. single agent) setting the actions of an individual coincide with the outcomes: an individual picks exactly one outcome of X. In a multi-agent or strategic context, the result of one's actions also depends on what the others do. We thus assign to each individual i a set of strategies S_i, and we call a combination σ of individual strategies, one for each individual, a *strategy profile*. An *outcome function* $\pi : \Pi_{i \in I} S_i \to X$ specifies for each strategy profile which element of X will result. In our simplified decision models, single-agent decisions can be seen, somewhat artificially, as degenerate cases of multi-agent decisions. We thus use the term "strategy" in the parametric context as well. In that case, a strategy is simply an element of X or, in other words, $X = S_i$ and the outcome function π is such that $\pi(x) = x$, for all $x \in X$. In a strategic context, $\pi(s_i)$ describes the set of all states of affairs that may arise if i adopts strategy $s_i \in S_i$. Formally, $\pi(s_i) = \{x|$ there is some combination of strategies σ_{I-i} for the other players such that $\pi(s_i, \sigma_{I-i}) = x\}$.[2] In order to facilitate reading, the individual subscript will often be omitted.

A *decision problem* will thus be a tuple $DP = \langle I, X, \{S_i, R_i\}_{i \in I}, \pi \rangle$, the elements of which have just been described. In both decision and game theory, *solution concepts* stipulate what it is rational to do, given a certain decision problem. For instance, in the parametric setting, choosing elements among the

[2] It would in fact be more correct to write π_i since the definition is in terms of i's strategies, but we shall not do so.

set of preferred elements $C(R)$ can be seen as maximizing expected utility.[3] The behavioural consequences of individual utility maximization can be intricate in a strategic setting, because the outcome of the game depends on what *all* agents choose. Several solution concepts exist and whether it is rational to choose a strategy depends on which solution concept is used.

Formally, a solution concept yields a set Γ of strategy profiles for each decision problem. Our analysis is limited to solution concepts that consist of combinations of *pure* strategies, thus leaving aside individual decision making under risk and randomization over strategies in games. Throughout the analysis it is assumed that some (non-empty) Γ is given. To each strategy-profile belonging to Γ correspond a particular outcome, which will be called a *feasible outcome*. The set of feasible outcomes will be denoted by Γ_π. Hence, for all $\sigma \in \Pi_{i \in I} S_i$, $\Gamma_\pi = \{\pi(\sigma) | \sigma \in \Gamma\}$. Γ_i denotes the set of i's strategies which are part of a strategy profile belonging to Γ. In the parametric context we simply have $\Gamma_\pi = \Gamma_i = C(R_i)$.

Given a choice situation and a solution concept Γ, we say that a strategy s_i for $i \in I$ is *utility-compatible* if and only if it is an element of Γ_i. The demand of rational choice, or the demand of *utility-rationality* as it will be called, can be defined as the requirement that an individual always chooses a utility-compatible strategy. Note that this terminology deviates somewhat from the standard terminology. Take for instance the game-theoretical solution concept of a pure Nash equilibrium. If there are several equilibria in a given game, it is conceivable that all individuals choose a utility-compatible strategy but that the combination of those strategies does not form an equilibrium. The reason is, of course, that a solution concept like the Nash equilibrium singles out strategy *profiles* rather than the strategies constituting those profiles. This tension between utility rationality and equilibrium concepts, which is sometimes latent in rational choice theory, is exacerbated once one introduces intentions, as we shall see at the end of Section 4.

3 Intentions and planning agency

Our analysis is restricted to future-directed intentions. This is one of at least three forms of intentions that have been extensively studied in the philosophy of action. As early as [1], future-directed intentions (e.g. "I intend to go to Berlin tomorrow") were distinguished from intentional actions ("I intentionally take the train to Berlin") and from intentions in actions ("I go to Berlin with the intention of enjoying myself for a few days"). The analysis of future-directed intentions has been enormously influenced by the work of Michael Bratman (see [2] and [3]). He sees future-directed intentions as plans for action that have certain characteristics. First, such plans are often *partial*: they usually do not describe every aspect of a person's future behaviour. I may have an intention to go to a concert tonight and yet not have decided how to get there, whether

[3] There exist other decision-theoretical solution concepts, but in this paper we will focus on utility maximization.

I will ask someone to join me, etc. Secondly, plans typically have a *hierarchical structure*. My intention to rent a car this afternoon may be the result of my intention to go to the concert tonight, my intention to go the concert tonight may be related to a more general intention to relax more often, and so on. Finally, intentions involve a certain kind of commitment, which Bratman calls the "volitive commitment". If I have the intention to realise a certain future state of affairs, and nothing unforeseen happens after I formed this intention, I will normally try to achieve it.

Two different kinds of future-directed intentions can be distinguished: I can have the intention to perform a certain action (e.g. "I intend to fly to Berlin next week") or I can have the intention to realise a certain state of affairs or outcome (e.g. "I intend to be in Berlin next week").[4] In hierarchically structured plans the two types are often related: the intention to fly to Berlin can be seen as the consequence of the intention to realise the state of affairs of me being in Berlin next week. In this paper we focus primarily on intentions to realise certain states of affairs i.e., certain elements of X. The intentions of an individual i are taken to be given exogenously. Each individual i has a set of intentions, denoted by M_i, which consists of subsets of X. Intuitively, an intention $A \in M_i$ is an intention to realise the aspects that all the states in A have in common. To illustrate, suppose that each element of X stands for a specific holiday destination. If it is my intention to spend my vacation in France, then it is my intention to realise the set A consisting of all possible destinations in France: one element of A may, for example, describe a vacation in Paris and another a vacation in Toulouse. If B is the set of all possible beach destinations, then having the intention to realise $A \cap B$ describes the intention of spending one's vacation on a French beach, $A \cup B$ describes the possible intention of spending one's vacation in France *or* at a beach, and so on.

Bratman has argued that certain norms apply to *rational* intentions.[5] We are going to capture these norms by imposing axioms on the intention set M_i of each agent.

Intentions are first required to be feasible.[6] An individual should not intend impossible states, such as visiting the mountains of Holland. To avoid trivial cases, we also require that the agent has *some* intentions, hence the following axiom:

Axiom 1. $\emptyset \notin M_i$ *and* $M_i \neq \emptyset$

[4]Note, however, that the distinction between outcomes and actions is not very rigid. An action (drinking a glass of milk) can also be described as a state of affairs (i.e. the state of affairs in which I am drinking a glass of milk).

[5]Some arguments can already be found in [2], but a clear and up-to-date argumentation can be found in [4].

[6]There is an important proviso to this norm in [4]. Bratman's view is that intentions should be feasible "together with one's beliefs"[p.1] or, in other words, that intentions of an agent i should be feasible if i's beliefs were true. Since beliefs are not modelled in our minimal rational choice framework, we shall reduce this belief-feasibility requirement to plain feasibility. This is of course not an argument for such reduction. Quite the contrary: we believe that this simplification only shows the importance of proceeding towards an analysis of intentions in more sophisticated models of rational.

Intentions are also required to be "agglomerative", meaning that they should close under intersection.[7]

Axiom 2. *If* $A, B \in M_i$*, then* $A \cap B \in M_i$*.*

An immediate consequence of these two axioms is that the intentions in a set M_i share a non-empty intersection. This translates into another norm for which Bratman has argued: the intentions of an agent should not exclude each other. Note that the intersection $\bigcap_{A \in M_i} A$ of all the intentions in M_i will be its smallest element. We use a special notation for this, $\downarrow M_i$, for it will prove to be crucial in many of the examples below.

Finally, we impose some kind of "intention logical omniscience".[8] One will be thought as to intend all the logical consequences of his intentions. If, for instance, I intend to go to Florida, then it is also my intention to go to the United States.

Axiom 3. *If* $A \in M_i$*, and* $A \subseteq B$ *then* $B \in M_i$

Bratman has stressed repeatedly the tight connection between, on the one hand, feasibility, agglomerativity and consistency and, on the other hand, specific functions of intentions, as both *input* to and *output* of the decision making process. In what follow we examine two of these roles in turn, to see how intentions constrained by Axiom 1, 2 and 3 influence decision making. In the next section we look at what Bratman calls the "volitive commitment" of intentions, which concern mainly what happens *after* an agent has adopted certain intentions. As an illustrative example, we show that it sheds light on the existence of "focal points", something that goes beyond standard rational choice theory.[9] In Section 5 we turn to the "reason-centered commitment" of intentions, according to which previously adopted intentions influence deliberation. We do so by modelling the fact that agents can use their intentions to simplify the decision problems they face. Finally, we put these two functions together in Section 6, and analyze some examples that reveal their joint effect on decision making.

4 Intentions and focal points

By "volitive commitment", Bratman denotes the fact that intentions act as motivational force towards action. In this section we see how this "driving force" can be used to explain focal points.

As a first step, let us set down some notation connecting strategies and intentions. Given a decision problem and a set of intentions M_i for an agent i, a strategy s_i will be said to be *intention-compatible* if and only if there is a

[7] We take the term "agglomerative" from [17].

[8] This constraint is mainly imposed for technical convenience: every intention set that satisfies Axiom 1, 2 and 3 turns out to be a filter, a property that will prove to be useful below. Axiom 3 has *not* been argued for by Bratman, but it has been defended in the literature, notably in [13].

[9] For an introduction to the literature on focal points, see [11, chap.3, sec. 5].

non-empty $A \subseteq \pi(s_i)$ such that A is a subset of all $B \in M_i$. In a parametric setting, where each strategy is associated with exactly one outcome, intention-compatibility implies that the alternative a person chooses is an element of each of the person's intentions. Thus my choice to go to Paris is only intention-compatible if I do not have an intention *not* to go to Paris. The definition is a little more complicated for strategic settings since a strategy may lead to various outcomes. Suppose, for instance, that I choose to pay a visit to a very close friend of mine who lives in Paris. Even though I cannot be sure that he will indeed be in Paris, going off to Paris is an intention-compatible strategy. It seems reasonable to assume that the volitive commitment of intentions is related to these intention-compatible strategies, that is, it seems reasonable to assume that a rational individual will choose one of his intention-compatible strategies (assuming there is one). This will be called the demand of *intention-rationality*.

We thus have the notion of acting rationally because one's actions are intention-compatible, intention-rationality, and we have the usual notion of instrumental rationality, or utility-rationality, as expressed by the requirements of rational choice. We will locate focal points at the intersection of these two notions.

Given a choice situation, an intention set M_i for each i, and a solution concept Γ, intention-rationality *coincides* with utility-rationality if for all $i \in I$ the set of intention-compatible strategies is a non-empty subset of the set of utility-compatible strategies. If, furthermore, there is an $i \in I$ for which it is a proper subset, then intention-rationality is said to *focalise* utility rationality. A strategy profile is a *focal point* if it is in the set of intention-rational profiles that focalise utility-rationality.

By defining focal points this way we only consider the special case where intentions coincide with the demands of utility-rationality. Amartya Sen [14] has famously argued that intentions (or more generally commitments) are of interest mainly when they do *not* coincide with utility maximization. The present work should not be seen as an argument against Sen's idea. Quite the contrary: our purpose here is to show that intentions can be of interest even in cases where they coincide with utility maximization.

4.1 Focal points in a parametric setting

Here we show the conditions under which intentions create focal points in a parametric setting. Of course, we could have jumped right away to the multi-agent setting, and treat individual decision making as a special case. But we think that the simpler setting of single-agent decision will shed light on our method, which will later be generalized to an arbitrary (finite) number of agents.

It turns out that the existence of focal points rests on a tight connection between intention- and utility-rationality, as expressed by the following axiom.

Axiom 4. *For any $A, B \subseteq X$, if $A \cup B \in M_i$ and for all $x \in A$ and $y \in B$, xP_iy then $A \in M_i$.*

Read contrapositively, this axiom states that if an individual does not have the intention to realise some set of outcomes A, and if he finds every element of

A strictly better than any element of B, then the union of A and B does not belong to his intention set. If I prefer France to Holland as a holiday destination, but it is not my intention to go on holiday in France, then it will surely not be my intention to go on holiday in France or in Holland.

We can now establish the following fact, which states that an individual will intend to realise one or more of his best elements. Furthermore, he intends to realise any set to which these particular elements belong, and will not intend any other set.

Proposition 1. *M_i satisfies Axioms 1-4 if and only if there is a non-empty $A \subseteq C(R_i)$ such that $M_i = \{B|\ A \subseteq B\}$.*

Proof. The only interesting direction is from left to right. We are going to show that $C(R_i) \in M_i$. This will be enough because, by Axioms 1 and 2, this will mean that any smaller set in M_i has to be a subset of $C(R_i)$, and M_i is closed under supersets (Axiom 3).

Take $A = C(R_i)$ and $B = X - A$. Observe that $A \cup B = X$ which means, by Axiom 2, that $X \in M_i$. But then, by definition of $C(R_i)$, we have that xP_iy for all $x \in C(R_i)$ and $y \notin C(R_i)$, that is, for all $y \in B$. We can thus apply axiom 4 and conclude that $A \in M_i$. □

The proposition reveals that Axioms 1 to 4 impose a specific structure on the intention sets: each of the individual's intentions is a superset of a set of best elements. Hence, the following result follows.

Corollary 1. *If M_i satisfies Axioms 1-4, intention-rationality coincides with utility-rationality.*

What about focalisation? Consider the following axiom, which states that the absence of an intention to realise a certain state of affairs always entails the existence of an intention to realise the "negation" of that state of affairs.[10]

Axiom 5. *If $A \notin M_i$, then $X - A \in M_i$.*

The following proposition can now be established.

Proposition 2. *M_i satisfies Axioms 1-5 if, and only if, there is an $x \in C(R_i)$ such that $M_i = \{B|\ x \in B\}$.*

Proof. Because X is assumed to be finite, this follows directly from Proposition 1, and the fact that under Axiom 5, M_i is an ultrafilter. □

In case Axioms 1-5 are satisfied the individual always intends to realise a specific element of X. Since there is only one intention-compatible strategy, viz. the one leading to that outcome, following his intentions entails that a best element will be chosen, namely the particular best element that he intends to realise. Hence, we immediately derive the following.

[10] Stated differently, the axiom states that the external negation of an intention ("it is not the case that A is intended") is equivalent to its internal negation ("it is the case that not-A is intended").

Corollary 2. *If M_i satisfies Axioms 1-5, and if $C(R_i)$ contains more than one element, then intention-rationality focalises utility-rationality.*

In these cases intentions create a focal point. If there are several utility-compatible strategies, rational choice does not tell us which of those strategies to adopt. A person's intentions do tell us, however: among those utility-compatible strategies there is only one which is also intention-compatible. One can interpret this by invoking the well-known distinction between picking and choosing, as it has been already pointed out in [9, p.183]: "to 'pick', in the relevant sense, is to form an intention".[11]

4.2 Focal points in strategic contexts

We now generalize the results of the last section to strategic contexts. It can easily be shown that Axioms 1-4 cannot be straightforwardly applied there. For if one were to do so, Proposition 1 would entail that the only intention-compatible strategies are those that may lead to one of the individual's best elements. Stated differently, Axioms 1-4 together imply that an intention-compatible strategy is *always* a maximising strategy, that is, a strategy s_i such that $\pi(s_i)$ contains a best element.

However, in many strategic situations such an assumption about the nature of intention-compatible behaviour is not very plausible since aspects of the situation will have a bearing on a person's intentions. Consider, for instance, the following game in normal form.

	t_1	t_2
s_1	(4,1)	(2,4)
s_2	(1,2)	(3,3)

Table 1: Intention-rationality in a strategic context

The most preferred outcome of the row player, whom we call 1, is (s_1, t_1). Hence, by Propositions 1 and 2 it follows that the intention set of 1 consists of $\{(s_1, t_1)\}$ and any of its supersets. Since $\pi(s_1) = \{(s_1, t_1), (s_1, t_2)\}$ is a superset of $\{(s_1, t_1)\}$, it belongs to 1's intention set. Because $\pi(s_2) = \{(s_2, t_1), (s_2, t_2)\}$ is not a superset of $\{(s_1, t_1)\}$, 1's only intention-compatible strategy is s_1. However, the desired outcome (s_1, t_1) will only be realised if the column player, whom we call 2, adopts a dominated strategy.

Now, can an individual really intend to realise an outcome that will *only* come about if the others do *not* act in a utility-rational way?[12] It seems reasonable to assume that considerations of utility-rationality will affect a person's

[11] Thus, on this interpretation an intention can have a role in decision-making processes without it being a reason for action. For a defence of the possibility of such a view, see [5]. For the distinction between picking and choosing, see [16].

[12] It is assumed here that a strictly dominated strategy will not belong to Γ_i.

intentions. In parametric setting, Axiom 4 established a link between a person's intention- and utility-rationality. The example above shows that the axiom should be modified so as to take account of the feasibility of the outcomes that will be intended.

Axiom 6. *Let A^* and B^* denote the sets consisting of all feasible elements of A and B, respectively. If $A \cup B \in M_i$, $A^* \neq \emptyset$ and either $B^* = \emptyset$ or for all $x \in A^*$ and $y \in B^*$, xP_iy then $A \in M_i$.*

Let $C_i(R_\Gamma)$ denote the set of an individual's most preferred feasible outcomes. We now derive the following proposition.

Proposition 3. *For all i, M_i satisfies Axioms 1-3 and 6 if and only if there is a non-empty $A \subseteq C_i(R_\Gamma)$ such that $M_i = \{B|\ A \subseteq B\}$.*

Proof. The proof of Proposition 1 is readily adaptable to feasible sets. □

Whereas the axioms imposed on intentions in the parametric setting lead to the conclusion that individuals will intend to realise one or more of their best outcomes, Proposition 3 shows that in a strategic context individuals will intend to realise one or more of their best *feasible* outcomes. In terms of the two types of rationality that were distinguished, we can also say that intention-rationality coincides with utility-rationality. In fact, it focalises it if at least one person has a utility-compatible strategy that can never lead to one of the feasible outcomes he finds best:

Corollary 3. *If for some individual there is a strategy in Γ_i that can never lead to one of her best feasible outcomes, then intention-rationality focalises utility-rationality.*

To illustrate this, consider a game in which there are two equilibria, one of which is Pareto-dominated by the other:

	t_1	t_2
s_1	(3,3)	(0,0)
s_2	(0,0)	(2,2)

Table 2: An intention-rational focal point

Applying the axioms to this game, we immediately see that both individuals intend to realise the Pareto-superior outcome. Playing their first strategy is thus the only course of action that is both utility-compatible and intention-compatible. Since the familiar game-theoretic solution concepts do not narrow down the solution this way, we see that intentions do indeed form a solution to at least some co-ordination problems: intention-rationality here gives *more* information than utility-rationality.

It could be argued that this in itself is not very revealing. After all, we would be surprised if players actually playing this game were to wind up with any of

the other outcomes. The outcome forms a focal point, and focal points serve to narrow down the set of equilibria. It should be emphasised, however, that although the conclusion may not be very surprising, the analysis of intentions gives an underpinning of this expectation, something which standard game theory cannot do.[13] After all, the notion of a focal point refers to factors "beyond" the game. Our expansion of the game-theoretic framework can be seen as a way of incorporating some of these factors into the game.

The strategic counterparts of Proposition 2 and Corollary 2 are:

Proposition 4. *For all i, M_i satisfies Axioms 1-3 and 5-6 if, and only if there is an $x \in C_i(R_\Gamma)$ such that $M_i = \{B |\ x \in B\}$.*

Corollary 4. *If, for all i, M_i satisfies Axioms 1-3 and 5-6 and there is some j such that at least two strategies in Γ_j can never lead to the same feasible outcome, then intention-rationality focalises utility-rationality.*

To illustrate, consider the classic example of a pure coordination game, i.e., a game in which there are two outcomes between which both individuals are indifferent:

	t_1	t_2
s_1	(1,1)	(0,0)
s_2	(0,0)	(1,1)

Table 3: A pure coordination game

To refer to the standard example of such a coordination game, assume you have an appointment to meet a friend but have forgotten to name a specific meeting point. Let one of the outcomes be the natural focal point, say the railway station. Assuming that Axioms 1-3 and 5-6 hold, intention-rationality not only says that you will go to one of the two places, but also which one you will go to. Hence, in this case intentions give more information. This does not, of course, guarantee that the coordination problem will be successfully solved. After all, the individuals may intend to go to different places. A distinction should be made, however, between the possibility of showing how the particular structure of intentions yields a way out of a coordination problem, and the possibility that intentions as such form the locus where such a solution is to be found. That the individuals will in fact attain an equilibrium outcome is not ensured by the particular structure that is imposed on the intentions. The individuals may have different intentions, may therefore adopt different strategies, and may thus end up in a non-equilibrium. However, it could plausibly be maintained that *if* one of the equilibrium outcomes forms a focal point then, by virtue of it being a focal point, it will be the outcome that individuals intend to realise in

[13] See, however, [7] who offers a general method for the selection of a unique outcome of non-cooperative games that is based on pre-game moves, some of which are described as "self-commitment moves". It would be interesting to explore the extent to which these self-commitment moves can be construed as intentions in the sense described here.

such circumstances. Saying that the railway station is a natural meeting point in such a coordination game is then understood to entail that the players will intend to realise that particular outcome. The intention to go there is thus a result of that particular outcome being a focal point. Again, the usefulness of the approach is that it brings these outside factors within the game.

Bratman himself argued that intentions are useful for solving coordinating problems, both on the intrapersonal and on a social level, and the results established thus far underline this point. In the parametric case, intentions are helpful in solving the problem of choosing between equally attractive outcomes. On the social level, they may help to narrow down the set of possible equilibria. However, although intentions can thus be shown to solve some coordination problems, there are also decision problems that our approach does not seem to be able to solve, in which recourse to intentions may in fact have counter-intuitive consequences. Consider games in which there are multiple equilibria and in which the players have divergent preferences concerning those equilibria. Take, for instance, the familiar Battle of the Sexes. Sticking to the stereotype version of it, assume that the male player prefers to go to a boxing match and his female friend wants to go to the ballet, but they both prefer being together than being alone at one of those events.

	Boxing	Ballet
Boxing	(2,1)	(0,0)
Ballet	(0,0)	(1,2)

Table 4: The "Battle of the Sexes"

Now assume that Axioms 1-3 and 6 hold for both individuals and also assume that the solution concept is pure Nash equilibrium. It can readily be seen that in this case each individual has only one strategy that is both intention-rational and utility-rational: the row player should choose his first strategy and his friend (the column player) her second one. But this combination of utility- and intention-compatible strategies leads to an outcome that is *not* an equilibrium.[14] In other words, in this example the combination of intention-rationality and utility-rationality can *only* result in a non-equilibrium outcome.

This is due to Axiom 6, under which a person will always intend to realise one or more of *his* best feasible outcomes. In other words, this axiom makes intention rationality "egocentric". Intentions are forced into an *individualistic* maximising behavior, irrespective of what the other players' intentions are. It should thus come as no surprise that, in examples such as the "Battle of the Sexes", intention-rationality inevitably steers the players out of the equilibria. There seems to be a notion of compromise built into the two pure Nash-feasible outcomes of this example, a notion that turns out to be at odds with the present definition of intention-rationality.

[14] In fact, it not only fails to be an equilibrium, it is also Pareto-dominated.

The next section, in which we study the simplificative function of intentions, provides other occasions to study the divergence between individualistic intention-rationality and interactive solution concepts such as (pure) Nash equilibrium. But a natural question at this moment is whether Axiom 6 can be weakened so as to give a less individualistic interpretation. This could be done in many ways, notably by "furnishing" our model with an information structure (cf. [12, chap.5]) thus allowing one to model a player's knowledge about the intentions of others. Another line of attack would be to look for a weakening of Axiom 6 more in line with a satisficing approach (cf. [15]). This is not pursued here, but it should be noted that it would contribute to a theory of "bounded rationality".[15]

5 Intentions and simplification of decision problems

In the previous section we showed that intentions can account for focal points. Now we turn to the function of intentions as *input* of deliberation, the "reason-centered commitment". Bratman has repeatedly emphasised that a plan acts as a *filter* over the set of options that will be considered during practical reasoning. It rules out options that are incompatible with its own achievement. If one intends to go to France for one's holidays, one will not consider Madrid as a potential candidate for best destination. Plans are also usually partial, and as such they ask for completion. The agent who plans to go to France will, at some point, have to decide between, say, Chamonix, Marseilles and Paris. But this completion does not need to settle every detail of the trip. The agent will ponder between alternatives that differ only *up to a certain level of detail*, and will surely not bother to decide now whether it is better to go to Paris with a red or white shirt, for example.

Of course, an agent without time constraints will not lose anything by considering a few incompatible alternatives, and an agent with unlimited "memory space" and "computational power" can handle any amount of detail, however irrelevant at the moment of deliberation. Thus it is only for agents with limited time and capacities that the functions of filtering and completion become really useful. In what follows we model these two simplification functions of intentions for bounded agents.

[15]Note that in the framework presented here it is possible to distinguish satisficing behaviour with respect to one's intentions from satisficing with respect to one's utility. The approach suggested here would explore the possibility of bounded rationality giving an underpinning of utility-maximisation: by forming satisficing rather than maximising intentions, utility-maximisation can perhaps be ensured. In other words, bounded rationality with respect to intentions may help to ensure "unbounded" rationality with respect to utility.

5.1 Ruling out options

The first function we consider is the ruling out of options that are incompatible with the achievement of a plan. Given the structure we imposed on these sets, this corresponds to removing from the original set of alternatives those that are not in $\downarrow M_i$.

In the parametric setting, ruling out options boils down to equating X to $\downarrow M$. But in a strategic situation there are various ways in which an agent can judge that a strategy is incompatible with his plan, depending on how much the agent is willing to "risk" on getting an alternative with the intended features. Here we only examine two extreme forms of attitude toward such risk. One can be *risk inclined* and consider that all strategies that *might* lead to an intended outcome are compatible with one's plan. But one can also be *risk-averse* and retain only the options that lead *for sure* to an intended outcome.

Formally, the *cleaned strategy set* $cl(S_i)$ for an agent i is defined as

$$cl(S_i) = \{s_i \mid \text{for all } \sigma_{I-i}, \pi(s_i, \sigma_{I-i}) \in \downarrow M_i\}$$

for risk averse cleaning and

$$cl(S_i) = \{s_i \mid \text{there is an } \sigma_{I-i} \text{ such that } \pi(s_i, \sigma_{I-i}) \in \downarrow M_i\}$$

for risk inclined cleaning. The *cleaned* version of a decision problem DP will be defined as the tuple $cl(DP) = \langle I, X', \{cl(S_i), R'_i\}_{i \in I}, \pi' \rangle$ with:

- $X' = \{x \mid \exists \sigma \in \Pi_{i \in I} cl(S_i) \text{ s.t. } \pi(\sigma) = x\}$
- $R'_i = R_i \cap (X' \times X')$
- π' is π with the restricted domain $\Pi_{i \in I} cl(S_i)$ and image X'

Clearly, every decision problem has a unique cleaned version, which is not empty if every cleaned strategy set is not empty. Note that the strategies that remain after risk-averse cleaning are just the intention-compatible ones but that a risk-inclined agent does not always comply with the demands of intention-rationality, as defined in Section 4. It should be noted, however, that there are some decision problems for which the cleaned strategy set of an agent i is empty for risk averse cleaning. Sometimes an agent can be too fussy about his own strategies! To avoid such cases we will use risk-inclined cleaning in the examples below, unless explicitly mentioned.

5.2 Ignoring irrelevant details

We mentioned that a plan demands a completion, but not for an over-detailed one. The level of detail to be considered will depend on the modes of achievement the agent has to decide upon. To capture this, take a partition $\mathcal{A}_i$ of $\downarrow M_i$ to be a set of means of achievement for the plan, on which the agent will have to make a decision. Intuitively, $\mathcal{A}_i$ can be seen as a set of mutually exclusive ways to achieve $\downarrow M_i$. In our model we are going to ignore irrelevant details by

grouping together options that are identical up to these attainments. Formally, two outcomes x, x' are *equivalent modulo* $\mathcal{A}_i$, denoted $x \sim^i_X x'$, if and only if they belong to the same cell $A \in \mathcal{A}_i$ or they do not belong to $\downarrow M_i$ at all. Two strategies s_i and t_i of i are equivalent modulo $\mathcal{A}$, denoted $s_i \sim^i_S t_i$, if and only if for all σ_{I-i} , $\pi(s_i, \sigma_{I-i}) \sim^i_X \pi(t_i, \sigma_{I-i})$. The reader can check that both relations $\sim^i_X$ and $\sim^i_S$ are equivalence relations on X and S_i, respectively, and that they are the same in parametric settings. The strategies belonging to the same equivalence classes $[s_i]$ can be seen as *attainment-equivalent* for i, with respect to his attainment set $\mathcal{A}_i$. However the other agents play, i will achieve his intention in the same way by choosing any strategy in $[s_i]$. When confusion may arise, we use $[x]_{\sim_X}$ and $[s]_{\sim_S}$ to denote the equivalence class of x and s under $\sim_X$ and $\sim_S$, respectively. Otherwise we just omit the subscripts and write $[x]$ and $[s]$. The equivalence classes of the strategies whose outcomes never belong to $\downarrow M_i$ will be denoted $[s]_{\overline{\downarrow M_i}}$.

To build a decision problem out of these equivalence classes of strategies we need to specify how the agents are going to "evaluate" them, so to speak.[16] One can imagine different ways to do so. Here we study one in which the agent "picks" one representative per equivalence class, and considers that the outcome(s) of this equivalence class is (are) just the outcome(s) that would be secured by choosing this representative in the original decision problem.

Given some decision problem DP, we call any function $\theta_i : \{[s_i] : s_i \in S_i\} \to S_i$ such that $\theta([s_i]) \in [s_i]$ a *representative function* for an agent i. A representative function gives, for each equivalence class, one chosen representative, out of which the simplified decision problem will be constructed. Given a problem DP and a representative function θ_i defined for each i, the *means-simplified version* $DP^* = \langle I, X^*, \{S^*_i, R^*_i\}_{i \in I}, \pi^* \rangle$ is defined as follows, for all $i \in I$:

- $S^*_i = \{\theta_i([s]) : s \in S_i\}$
- $X^* = \{x \mid \exists \sigma \in \Pi_{i \in I} S^*_i \text{ s.t. } \pi(\sigma) = x\}$
- $R^*_i = R_i \cap (X^* \times X^*)$
- π^* is π with the restricted domain $\Pi_{i \in I} S^*_i$ and image X^*

Many decision problems have multiple means-simplified versions, as we shall see in the next section. Each of these versions has a corresponding, different *profile of representative functions* $\Theta = \Pi_{i \in I} \theta_i$. Given their importance, we shall sometime abuse our own notation and use the profiles Θ to designate the simplified version they generate. Note that a profile Θ will simplify a decision problem, just in case there is one agent for which one of his $\sim_S$-equivalence classes is not a singleton.

[16] "So to speak", because what the agent really values are *outcomes* and not strategies.

6 Ruling out *and* ignoring irrelevant details

We proposed two operations on decision problems to model two aspects of reason-centered commitment, and we see no reason why intention-based decisions should be confined to one of them. In other words, we would like simplified decision problems to be both "cleaned" and "means-simplified". But we cannot go on without specifying the *sequential order* in which these operation are being applied, because they do not commute in the general case, assuming that the intention set of an agent satisfies Axioms 1 and 2.

Proposition 5. *It is not the case that, for all decision problems DP,* $cl(DP^*) = (cl(DP))^*$.

Proof. The following is a counter-example.

	t_1	t_2
s_1	(0,1)	(1,1)
s_2	(1,1)	(0,1)

Table 5: Counter-example to commutativity

Assume that Γ is the set of all strategy profiles.[17] Let the intention and attainment sets of each player be as follow: $\downarrow M_1 = \{(s_2, t_1), (s_1, t_2)\}$, $A_1 = \{\{(s_2, t_1)\}, \{(s_1, t_2)\}\}$, $\downarrow M_2 = C_2(R_\Gamma)$ and $A_2 = \{C_2(R_\Gamma)\}$. It means that $[s_1] = \{s_1\}$, $[s_2] = \{s_2\}$ and $[t_1] = [t_2] = \{t_1, t_2\}$. Take $\theta_2([t_1]) = t_1$, and consider the means-simplification below.

θ_2	$\theta_2([t_1]) = t_1$
$[s_1]$	(0,1)
$[s_2]$	(1,1)

Table 6: The means-simplification of Table 5

If we perform a risk-inclined cleaning on this decision problem, $[s_1]$ will be removed. In other words, $[s_1] \notin cl(DP^*)$. But observe that $cl(DP) = DP$ and that Table 6 thus is $(cl(DP))^*$, which means that $(cl(DP))^* \neq cl(DP^*)$. □

The counter-example used in this proof crucially involves more than one agent, and shows that the order in which the two simplification operations are performed makes a difference. But the reader may have noticed that it involves *one* application of each operation. In other words, it might be possible that for some decision problems and some representative function profiles cleaning and means-simplification converge towards a "minimal" simplified version after a finite number of alternative applications of these two operations. For example, the iterative simplification of the game used in the last proof does stabilize after

[17] Γ could be the removal of weakly dominated strategies, for example.

one more cleaning: the reader can check that $cl((cl(DP))^*) = cl(DP^*)$ and that any further application of cleaning or means-simplification will not simplify the problem further. This decision problem thus has a unique *simplificative fixed point*. If every decision problem had a unique simplificative fixed point where each agent has a non-empty strategy set, we could just refer to it when we mention its "simplified version". However, it is not only possible that a decision problem has different simplicative fixed points, but it may also be the case that the fixed point leaves no strategy available for one of the players. Consider the following game:

DP	t_1	t_2
s_1	(0,1)	(0,1)
s_2	(1,1)	(0,1)

Table 7: A case where simplification "empties" the strategy set of one player.

If we assume that the intention set of each player satisfies Axiom 1-3 and 6 and we take $\downarrow M_1 = \{(s_2, t_1)\}$ and $\downarrow M_2$ as the whole outcome set, a first round of cleaning results in the following.

$cl(DP)$	t_1	t_2
s_2	(1,1)	(0,1)

Table 8: The cleaned version of Table 7

Assuming further that $A_2 = \{\downarrow M_2\}$ and, taking $\theta_2([t_1]) = t_2$, we obtain the following matrix after means-simplification of $cl(DP)$.

$(cl(DP))^*, \theta_2$	$[t_1]$
$[s_2]$	(0,1)

Table 9: The means-simplification of Table 8

But then $[s_2]$ does not survive a further step of cleaning, leaving no strategy for 1. This can be seen as a disappointing result, for it forces us to specify the order in which the operations are applied, and to keep track of them if we want to avoid ending up with no strategy for one player. But this example, as well as the one displayed in Table 5, show the importance of interaction in the simplification procedure, for the two operations commute in parametric contexts. In other words, when there is only one agent a unique simplificative fixed point is always reached after a single "round" of cleaning and means-simplification.

Proposition 6. *For any parametric decision problem DP, if M_i satisfies Axioms 1 and 2 then $(cl(DP))^* = cl(DP^*)$.*

Proof. The proposition is a direct consequence of the following lemma.

Lemma 1. *For all (parametric or strategic) decision problems DP where M_i satisfies Axioms 1 and 2, given a partition $\mathcal{A}_i$, if a strategy s_i is removed after the risk-inclined cleaning of DP then its equivalence class $[s_i]$ would also be risk-inclined cleaned from the means-simplified DP^*, and so for all simplification profiles Θ on which DP^* can be computed. Furthermore, for all parametric decision problems DP where M_i satisfies Axioms 1 and 2, given a partition $\mathcal{A}_i$, if an equivalence class $[s_i]$ is risk-inclined cleaned from the means-simplified DP^*, then the strategy s_i would be removed after the risk-inclined cleaning of DP.*

For the first part, assume that s_i would be risk-inclined cleaned from DP. It means that for all profiles $\sigma \in \Pi_{j \in I-i} S_j$, $\pi(s_i, \sigma) \notin \downarrow M_i$. In turn, it means that, in the means simplification of DP, $s_i \in [s]_{\overline{\downarrow M_i}}$, which is the same as to say that for all s' such that $s' \sim_S^i s_i$ and all profiles $\sigma \in \Pi_{j \in I-i} S_j$, $\pi(s_i, \sigma) \sim_X^i \pi(s', \sigma)$ and so $\pi(s', \sigma) \notin \downarrow M_i$. But then for all simplificative functions θ_i, $\pi^*(\theta_i([s_i])) \cap \downarrow M_i = \emptyset$, which means that in any means-simplification of DP, $\theta_i([s_i])$ would be risk-inclined cleaned out.

For the second part we can restrict ourselves to a parametric setting, the game exhibited in Table 5 being a counter-example for the strategic case. Assume $[s_i]$ would be risk-inclined cleaned from DP^* obtained from an arbitrary simplificative function θ. It means, that $\theta([s_i]) \notin \downarrow M_i$ (recall that in parametric contexts, $S_i = X$). Now there are two cases to consider. If $\theta([s_i]) = s_i$, then we automatically get that $s_i \notin \downarrow M_i$ and so that this strategy would be risk-inclined cleaned. If $\theta([s_i]) = s_j$ for $i \neq j$, this means that $s_j \notin \downarrow M_i$. But since $s_j \in [s_i]$, we know that $s_i \sim_X^i s_j$ (recall again that $\sim_X^i$ is the same relation as $\sim_S^i$ in parametric setting), which can only be the case if s_i is also not in $\downarrow M_i$. It means that s_i would be cleaned out of the original decision problem. But observe that the last step holds for all $s_j \in [s_i]$, and so for any simplificative function θ, which proves the implication. □

Cases where the two operations do not commute, or end up with an empty strategy set, are thus to be found in strategic situations. In the examples presented so far we can indeed see that each player's simplification is crucially influenced by the other's. In the example of Table 7, the fact that 1 ends up with no intention-rational strategies is a direct consequence of the fact that 2 "chooses" t_2 as a representative of $[t_1]$. Similarly, in the example of Table 5, $[s_1]$ would be cleaned out of the means-simplified version of DP only because 2 chooses t_1 as representative.

This "mutually-triggered" simplification surely recalls similar phenomena in iterated elimination of dominated strategies. Of course, the two procedures differ, as can be seen in Table 5, where there is no pure dominant strategy. But this similarity calls for further research on the connection between traditional game theoretical solution concepts and simplification behaviour, given the link between utility and intention rationality provided by Axiom 6. Under which conditions do cleaning and means-simplification commute where Γ is, say, pure

Nash equilibrium? Under which general conditions do they yield an empty strategy set? These are open questions that we are not going to answer here. Rather, we want to examine another phenomenon that parallels known game-theoretical concepts, this time at the interface between intention-rationality, simplification and decisions.

7 Optimal simplifications

So far we have only investigated the behaviour of the simplification procedure. We now turn to "intention-based" decision making, that is, decision involving both volitive and reason-centered commitments of intentions. In other words, we are interested in agents who use their intentions both to *simplify* the problem they face and to *focus* on certain solutions.

As we saw previously, the outcome of a game can be influenced by the simplification procedure. In Table 5, the final choice of 1 depends on which representative of $[t_1]$ is picked by 2: $[s_2]$ if 2 picks t_1 and $[s_1]$ if 2 picks t_2. Of course, in that case, one can point out that this difference is not really noteworthy, because 1 is indifferent between (s_2, t_1) and (s_1, t_2). But can it be that some simplifications are better than others for some players? Or, to put it the other way around, can some players make things worse by picking the "wrong" simplification representatives?

To investigate this question, let us say that a profile Θ_1 *weakly dominates for player i* the profile Θ_2, given the solution Γ, if and only for all feasible outcomes x in Θ_1 and y in Θ_2, xR_iy and there are feasible outcomes x' and y' respectively in Θ_1 and Θ_2 such that $x'P_iy'$. Note that the comparison benchmark is the original game, where all outcomes are still reachable.[18] Similarly, let us call an *optimal simplification* a representative function profile Θ (with θ_i the i^{th} component of Θ) such that, given a solution concept Γ, for all $i \in I$ and all θ_i', if x is feasible in Θ and y is feasible in $(\theta_i', \Theta_{I-\{i\}})$ then xR_iy.

To illustrate, take the game in the following matrix. Assume that Γ is the pure Nash equilibrium solution concept and that the intention set of each player satisfies Axioms 1-3 and 6.

	t_1	t_2	t_3
s_1	(2,2)	(1,2)	(0,0)
s_2	(2,2)	(0,0)	(0,2)

Table 10: A game with dominated simplification

This game has four pure Nash equilibria: $\Gamma = \{(s_1, t_1), (s_2, t_1), (s_1, t_2), (s_2, t_3)\}$. Among them, 1 has a clear preference, $C_1(R_\Gamma) = \{(s_1, t_1), (s_2, t_1)\}$, while 2 is indifferent between all feasible outcomes: $C_2(R_\Gamma) = \Gamma_\pi$. Suppose $\downarrow M_1 = C_1(R_\Gamma)$ and $\downarrow M_2 = \{(s_1, t_2), (s_2, t_3)\}$, together with $\mathcal{A}_1 = \{\downarrow M_1\}$, $\mathcal{A}_2 = \{\{(s_1, t_2)\}, \{(s_2, t_3)\}\}$.[19]

[18] But not necessarily feasible, of course.

[19] Agent 2 might intend to "harm" 1, for example.

It means that s_1, s_2, t_2 and t_3 remain after risk-inclined cleaning.[20] Moreover, we get $[s_1] = [s_2] = \{s_1, s_2\}$ but $[t_2] = \{t_2\}$ and $[t_3] = \{t_3\}$. It means that 1 has two ways to simplify the decision problem: $\theta([s_1]) = s_1$ and $\theta'([s_1]) = s_2$, while 2 has no further simplification available. The games resulting from the two simplification possibilities of 1 are displayed in Table 11.

Θ_1	t_2	t_3
$\theta([s_1])$	(1,2)	(0,0)

Θ_2	t_2	t_3
$\theta'([s_1])$	(0, 0)	(0,2)

Table 11: The two simplifications of Table 10

Since 1 strictly prefers the Nash equilibrium of Θ_1 over that of Θ_2, the former dominates the latter and is an optimal simplification of this game.

Again, one can ask whether there are connections between dominated strategies and dominated simplifications, as well as between Nash equilibria and optimal simplifications. We have not investigated the full generality of these connections, but we do hope to have shown the importance of them.

8 Conclusion

The point from which this paper departed was the question of whether intentions and plan can introduce interesting questions *within* rational choice theory. We have focused on intentions to realise states of affairs, and tried to capture some of their important features axiomatically. We then proceeded to explore the extent to which "intention-rationality" is compatible with "utility-rationality". In particular, it was shown that intentions can account for focal points. We subsequently switched focus to simplification, and showed an interesting interplay between simplification and traditional utility rationality.

It should be emphasised that we started from a rational-choice framework. Given the specification of a particular decision situation, we examined how intentions might add something to its analysis. It means, for instance, that some of the axioms imposed on intentions could be seen as decision-theoretic constraints on those intentions. It cannot be emphasised enough that the analysis is restricted to particular decision situations, viz., situations in which the individuals possess complete information. Obviously, in a setting of incomplete information the analysis may have to be modified considerably. Moreover, it was noted that the maximising stance that underlies the axioms that were imposed on intention sets may lead to counterintuitive results.

Further analysis of the relation between intentions and rational choice within a framework may perhaps contribute to the development of a 'richer' theory of choice. We have already hinted at a different notion of feasible option, in which the feasible set is no longer defined in terms of mainstream rational-choice theory, but rather in terms of what the agent believes or knows. Other alternative

[20] Note that *no* strategy would remain after risk-averse cleaning.

accounts of rational choice have been developed, of course: we should mention [6] and [8], alongside the other approaches to rational action that have already been referred to. This paper's attempt to provide a "standard" rational-choice-theoretic analysis of intentions, and the problems that it yields, may provide an underpinning for the further exploration of these alternative approaches.

References

[1] G.E.M. Anscombe. *Intention*. Harvard University Press, Harvard University Press, 1957.

[2] Michael Bratman. *Intentions, Plans and Practical Reasons*. Harvard UP, London, 1987.

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

[4] Michael Bratman. Intention, belief, practical, theoretical. Unpublished Manuscript, Stanford University, January 2006.

[5] Govert den Hartogh. The authority of intentions. *mimeo*, 2001.

[6] David Gauthier. Resolute choice and rational deliberation: A critique and a defense. *Nous*, 31(1):1–25, Mars 1997.

[7] John C. Harsanyi and Reinhard Selten. *A General Theory of Equilibrium Selection in Games*. MIT Press, Cambridge, 1988.

[8] Edward F. McClennen. *Rationality and Dynamic Choice : Foundational Explorations*. Cambridge UP, 1990.

[9] A. Mele. Intentions, reasons, and beliefs: Morals of the toxin puzzle. *Philosophical Studies*, 68(2):171, 1992.

[10] A. Mele, editor. *The Philosophy of Action*. Oxford UP, 1997.

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

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

[13] Rohit Parikh. Logical omniscience and common knowledge: What do we know and what do we know? In *TARK '05: Proceedings of the 10th conference on Theoretical aspects of rationality and knowledge*, pages 62–77, Singapore, Singapore, 2005. National University of Singapore.

[14] Amartya Sen. Why exactly is commitment important for rationality? *Economics and Philosophy*, 21(01):5–14, 2005.

[15] H. A. Simon. *Models of Bounded Rationality*, volume 1-2. MIT Press, 1982.

[16] E. Ullmann-Margalit and S. Morgenbesser. Picking and choosing. *Social Research*, 44:757–85, 1977.

[17] David Velleman. What good is a will? Downloaded from the author's website on April 5th 2006, April 2003.