Intentions, Decisions and Rationality

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

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

May 7, 2007

1 Introduction

Rational agents, as conceived in decision and game theory (see e.g. [Myerson, 1991] and [Osborne and Rubinstein, 1994]), are agents who choose the actions they believe to be the best means to achieve their preferred ends. But human beings are also *planning* agents. That is, we have the capacity to 'settle *in advance* on more or less complex plans concerning the future, and then [let] these plans *guide our later conduct*' [Bratman, 1987]. Do decision theoretical models do sufficient justice to this planning aspect of human agency? In particular, can one account *within* a decision theoretical framework for the fact that planning agents can *commit* themselves by forming intentions? In [van Hees and Roy, 2007] we addressed this question for one particular kind of intention, namely the intention to bring about some future states of affairs, or what we now call 'outcome intentions'. These intentions, e.g. the intention 'to be in France next week', can be distinguished from 'action intentions' which refer to the future performance of some actions, e.g. the intention 'to take the plane to France'. We argued that outcome intentions can be fruitfully incorporated within a decision theoretic framework, and presented conditions under which rationality with respect to intentions is compatible with the standard payoff-maximizing notion of rationality.

This paper presents a further extension of the decision-theoretic model which now also includes action intentions. In the first sections we introduce the philosophical theory of intentions that motivates our investigation (Section 2) and a fairly standard decision theoretic model (Section 3). In Section 4 we propose a way to introduce action intentions in this model. We explain why – assuming that such intentions do not have autonomous consequences – they reduce to something which is already modeled. We nevertheless show that intentions can be useful as 'tie breakers' between equally desirable plans of actions. From Section 5 on, we concentrate on intentions with autonomous consequences. We first look at a case where different intentions may lead to the same outcome, while nevertheless influencing the payoffs. We argue that the standard decision theoretic framework falls short of accounting for such cases and present a counterfactual extension that fixes this deficiency. We then

(Section 6) turn to the famous 'Toxin Puzzle' [Kavka, 1983], in which it is rational to *form* a certain intention to act but not to *perform* that very act. We argue that decision theory cannot account for this particular type of autonomous effect either, and propose a second modification to cope with it. This extension allows us to capture the tension that, in our view, forms the heart of the toxin puzzle, viz., the tension between, on the one hand, a consistency requirement between intentions and actions; and the assumption of payoff maximization on the other.

All of the analysis is performed under the umbrella of fully idealized agency. We think that the non-ideal case is also very interesting, but we do not investigate it here. The reader may consult [Sen, 1977], [McClennen, 1990], and [Gauthier, 1997], where intentions for non-ideal agents are considered more thoroughly.

2 Intentions to act

Michael Bratman famously argued for the importance of intentions and plans in rational decision making in the development of his 'planning theory of intentions'.[1] This theory describes the functions of *future-directed* intentions, that is, intentions that are formed some time before their achievement. It is common in philosophy of action to distinguish between these intentions and intentions *in action* or *tryings* (see e.g. [Searle, 1983]). While the former are acquired before the action takes place, the latter form the 'mental component' [O'Shaughnessy, 1973] of intentional actions.

In the planning theory of intentions, 'plans' are special *sets* of intentions. They have an internal *hierarchical structure*. On 'top' are general intentions, for example going to Paris, beneath which come increasingly more precise intentions, for example going by train, depart at 9.15, and so on. The intentions at the bottom of this hierarchy are the most detailed, although they need not settle every single detail. The planning theory of intentions holds that it is even counter-intuitive to suppose this, particularly for agents with limited time and memory; it conceives of plans as typically *partial*. Planning agents 'cross the bridge when [they] come to it' [Savage, 1954, p.16], by forming new intentions along the way.

The planning theory of intention sees *rational* plans as regulated by norms of consistency. First, they should not contain inconsistent intentions, as for example the intention to go to a Bob Dylan concert in Paris tomorrow evening and the intention to go to a movie the same evening in Amsterdam. We call this the requirement of *endogenous* consistency.[2] A plan to

[1] [Bratman, 1987], [Bratman, 1999] and [Bratman, 2006b]. Other key references include [Anscombe, 1957] and [Harman, 1986].

[2] Rational intentions are also supposed to be *exogenously* consistent. This kind of consistency relates the agent's intentions to what he believes. A rational plan does not have to be feasible in the world as it actually is. It suffices that the plan is feasible in the world as the agent believes it to be. In this paper we always deal with 'perfect information' situations, where the agent is completely and truthfully informed. In this context, plans and intentions will always be exogenously consistent. We thus leave this constraint aside.

achieve a certain end should also be supplemented with intentions to undertake appropriate means. A planning agent must find a way to cross the bridge when he comes to it. This pressure for plan completion is called the requirement of *means-end* consistency.

One central claim of the planning theory of intentions is that rational plans are 'stable'. Once an agent has formed an intention to act in a certain way, then if nothing unforeseen happens he will not be disposed to reconsider it. Rather, he will act in order to achieve it, and take this new intention into account when he plans other actions. As such, plans are 'all-purpose means' [Bratman, 2006b, p.275] for a human agent. They allow for a better personal coordination towards the achievement of our ends, and provide a straightforward way to avoid costly deliberations. What is more, they constitute a solid base for coordination in social contexts.

3 Extensive decision models

3.1 Basis definitions

Decision theory offers a huge collection of mathematical tools to analyze *rational* decision making in the face of *uncertainty*. Rationality is understood here as *instrumental*, i.e., choosing what is believed to be the best means to reach one's preferred ends. The notion of uncertainty refers both to the agent's partial control over the consequences of his actions and to imperfect information about his environment. For example, the consequence of an action such as buying a lottery ticket depends on the result of a random process. Similarly, the consequences of placing a bet in a horse race depends on which horse is the fastest, a fact that gamblers are typically uncertain about. 'Nature' is thus the source of uncertainty in decision theory. This should be contrasted with uncertainty that arises from the interaction with other rational agents, who form expectations about each other's behavior. This is the subject matter of *game theory*, which we do not consider in this paper.

Since the publication of [von Neumann and Morgenstern, 1944] and [Savage, 1954], a plethora of decision theoretic models have been proposed. We want to investigate the place of future-directed intentions to act in decision theory, so we choose to work with *extensive models*, because they make explicit the temporal structure of decision problems. A *finite decision tree* [Osborne and Rubinstein, 1994], like the one depicted in Figure 1, is a finite set T of finite sequences of action called *histories*. We assume that T contains the empty sequence $\varnothing$, which will be called its *root*, and that it is closed under taking sub-sequences. That is, if $(a_1, \ldots, a_n, a_{n+1})$ is a sequence in T then $(a_1, \ldots, a_n)$ should also be in T. Given a history $h = (a_1, \ldots, a_n)$, we denote $ha = (a_1, \ldots, a_n, a)$ the history h followed by the action a. A history h is *terminal in* T whenever it is the subsequence of no longer history $h' \in T$. The set of terminal histories in T is denoted Z.

Each non-terminal history h ends with either a *choice node* or a *chance node*. As a shorthand we often talk of the histories as being themselves nodes, viz., the last node in

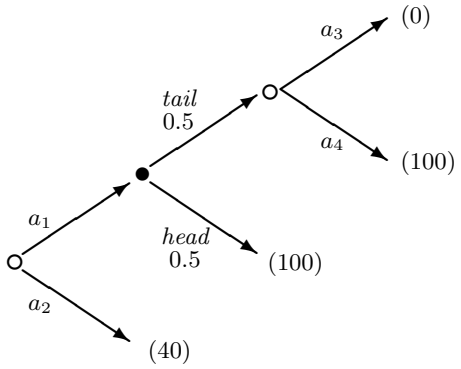

Figure 1: A simple decision problem

the sequence. Graphically, we represent the choice nodes/histories by white circles and the chance nodes/histories by black ones. In the example shown in Figure 1 there are two choice nodes and one chance node. If h is a choice node, then we call the set $A(h) = \{a : ha \in T\}$ the set of *actions available at* h. In the diagram, the actions available at a choice node are arrows labeled a_1 through a_4. If h is a chance node a function δ gives a probability value to the elements of $A(h)$, which we then call the *alternatives at* h. In figures, the number neighboring an arrow that stems from a chance node is the probability assigned by δ to that alternative. For example, in Figure 1, each alternative is assigned 0.5.

The real-valued payoff function $\pi : Z \to \mathbb{R}$ gives for each terminal history the payoff that the agent receives on reaching it. In Figure 1, the history (a_1, tail, a_4) gives a payoff of 100, for example. A *strategy* s is a function that gives, for every choice node h, an action $a \in A(h)$. When no confusion can arise, we describe a strategy – as we shall later do with plans of action – as a vector of actions. In Figure 1, for instance, there are four strategies: (a_1, a_3), (a_1, a_4), (a_2, a_3), (a_2, a_4). We say that a node h' is not excluded by the strategy s from h if the player can reach h' by choosing according to s from h on. Again, in our example, the history $(a_1, tail, a_3)$ can obviously be reached from the root by choosing according to (a_1, a_3), but not by choosing according to (a_2, a_3). Observe that a strategy tells the agent what to do even at histories that are not reachable by the strategy itself. In contrast, a *plan of action* is a function that assigns to each choice node h *that it does not itself exclude*, an action $a \in A(h)$.

A *partial* plan of action is a function p' from a *proper* subset of the set of all choice nodes to the set of actions such that there is at least one plan p that coincides with p' on the set of points belonging to p's domain.[3] The set of all partial plans in Figure 1, for instance, is

[3]As for strategies, plans of action can be described both as vectors or as partial functions, i.e. as sets of pairs (h, a) with $a \in A(h)$. When we say that (partial or non-partial) plan p coincides with partial plan p' we mean that, viewed as sets of pairs, $p' \subseteq p$. In what follow we will also talk about union of plans of action, which should always be read as the union of such pairs.

$\{a_1, a_2, a_3, a_4\}$.

The set of outcomes reachable by choosing according to a strategy s, or a plan of action p, is defined as the union of the $\pi(h)$ for all terminal histories reachable from the root by s (or p). The set of outcomes reachable by a *partial* plan of action p is the union of the outcome reachable by all the plans of actions p' that coincide with p. We shall abuse our notation slightly and denote these sets by $\pi(p)$.

The *expected value* of a strategy s at the history h, denoted $EV(h, s)$, is defined inductively.

$$EV(h,s) = \begin{cases} \pi(h) & \text{If } h \text{ is a terminal history} \\ \Sigma_{a \in A(h)} \delta(a) EV(ha, s) & \text{If } h \text{ is a chance node} \\ EV(hs(h), s) & \text{If } h \text{ is a choice node} \end{cases}$$

One can readily calculate that in the example of Figure 1 the expected values of the strategies (a_1, a_3), (a_1, a_4), (a_2, a_3), (a_2, a_4) are 50, 100, 40 and 40, respectively. The expected value of a plan of action p is defined only for pairs (h, p) such that p is defined on h and is computed similarly. Observe that, given this definition, a plan has exactly the same expected value as all the strategies that coincide with it. The expected value of (a_2), for example, is the same as (a_2, a_3), (a_2, a_4). A strategy s is *rational at* h if and only if, for all strategies s': $EV(h, s) \geq EV(h, s')$. That is, it is rational for the agent to choose according to s at h whenever s maximizes his expected value at that node. We say that a strategy s is *rational for the whole decision problem* T if it is rational at the root of T. Borrowing from the game-theoretic vocabulary, we often refer to rational strategies as the *solution* of a decision problem. The rational strategy or the solution of our simple example is thus (a_1, a_4).[4]

3.2 Assumptions about the decision maker

Computing the solution of a decision problem is not always as easy as in Figure 1. If there are numerous choice nodes, interlaced with chance nodes, representing the decision tree or calculating its solution might be very tedious. Most treatments of decision theory abstract from time and complexity constraints by making two assumptions about the agent: ideal *intelligence* and ideal *rationality*.

The first assumption concerns the agent's representational and computational capacities. An agent 'is intelligent if he knows everything that we [the modeler] know about the [problem] and he can make any inferences about the situation that we can make' [Myerson, 1991, p.4]. In

[4]Strategies are tailor-made for game-theoretic purposes. The information they carry about "off-path" behavior is a key input in the interactive reasoning that underlies solution concepts such as the sub-game perfect equilibrium [Selten, 1975]. But in decision theory, what the agent would do were he to deviate from his own plan of action usually has no consequence, at least as long as we talk about ideal agents. The reader might already have noticed that, according to our definition, a strategy can be said to be rational even though it prescribes moves at nodes that it itself excludes. In other words, our definition of what it is for a strategy to be be rational in a decision problem already ignores the "off-path" prescriptions.

other words, if a decision problem is representable at all and its solution is computable in any sensible sense of these terms, then the agent is assumed to be capable of representing it and computing its solution. The time and energy costs of these computations are usually thereby ignored. The rationality assumption splits into two components. First, the preferences of the agents over uncertain outcomes are assumed to satisfy certain conditions, such as transitivity, completeness and what has been called the 'sure-thing principle' [Savage, 1954].[5] These, together with a few others, are sufficient conditions for representing the agent's choices as a maximization of expected value (see [von Neumann and Morgenstern, 1944]). In the model above we directly represented preferences in these terms. Decision theoretic agents are also assumed to be constant and flawless maximizers, meaning that at every choice point they choose according to a rational strategy, and that they do not make mistakes, i.e. irrational choices.

Ideal decision theoretic agents are thus perfectly rational agents who can represent any decision problem, however great, and compute without effort its rational solution. These are indeed extremely strong idealizations, and most of them are explicitly made as simplifying hypotheses. Regarding the computation cost, for example, Savage says :

> [We] deliberately pretend that consideration costs the person nothing [...]. It might, on the other hand be stimulating [...] to think of consideration and calculation as itself an act on which the person decides. Though I have not explored the later possibility carefully, I suspect that any attempt to do so formally leads to fruitless and endless regression. [Savage, 1954, p.30]

Decision models for non-ideal or resources-bounded agents have been proposed (see e.g. [Simon, 1982], [Rubinstein, 1998] or [Gigerenzer and Selten, 2002]), and intentions can be shown to be important in such cases (see [Hammond, 1976] and, again, [McClennen, 1990], [Gauthier, 1997] and [van Hees and Roy, 2007]). As mentioned, we do not examine the non-ideal case in this paper.

4 Intentions as Plans of Action

In this section we introduce intentions to act in the decision-theoretic framework sketched earlier. We argue that for a large class of decision problems, the intentions of ideal agents can be viewed as plans of action, in the technical sense introduced above. This amounts to saying that intentions are already present, somehow implicitly, even though there is almost no mention of them in 'standard' decision theory. Does it therefore mean that the introduction of intentions in decision theory is a redundant enterprise? We believe not. As we show at

[5] Transitivity states that if x is preferred to y, and y is preferred to z, then x is preferred to z. Completeness states, for any x and y, that either x is preferred that y, or y to x. Finally, the sure-thing principles stipulates that if x were preferred to y upon learning that event E occurred, and also upon learning that E did not occur, then x is preferred to y regardless of E.

the end of this section, intentions may help to 'break ties' between equally desirable plans. It should be kept in mind that we restrict the analysis throughout this section to intentions that do not have so-called 'autonomous effects'. As we show in Section 5 and 6, difficulties arise when one tries to incorporate intentions that do have autonomous effects in decision theory. The conclusions of the present section thus only apply to intentions without autonomous effects.

Recall that we are interested in *future*-directed intentions, that is, intentions that are formed some time before their execution. In the decision models we use we represent this anteriority by assuming that the agent comes already equipped, so to speak, with both outcome and action intentions. We thus assign to the agent, given a decision tree T, an intention structure $\mathcal{I} = \langle M_X, M_A \rangle$ where M_X is a collection of sets of *outcomes* and M_A a collection of (maybe partial) *plans of actions*. Our strategy is to impose constraints on these sets to capture the features familiar to the planning theory of intention.

The set M_X represents the outcome-intentions of the agents. Each set A in M_X describes a state of affairs that the agent intends to realize. For instance, if A is a set of states of affairs in which the person is attending a Bob Dylan concert, then $A \in M_X$ represents the intention to attend a Bob Dylan concert. If B is a set of states of affairs in which the agent is in Paris, then $A \cap B \in M_X$ stands for the intention to watch a Bob Dylan concert in Paris. Borrowing directly from [van Hees and Roy, 2007], we assume that M_X satisfies the following constraints.

Postulate 1 (Endogenous Consistency of Outcome Intentions) *$\emptyset \notin M_X$ and $M_X \neq \emptyset$. Furthermore, if $A, B \in M_X$, then $A \cap B \in M_X$.*

This postulate prescribes that the agent does not intend to do the impossible, and that he at least intends something. It also enforces the set of outcome intentions to be closed under intersection.[6] The postulate implies that the agent has a 'smallest' outcome intention. That is, for some $A \in M_A$ we have $A \subseteq B$ for all $B \in M_A$. We often refer to this smallest set in what follows, denoting it by $\downarrow M_X$.

Inasmuch as $\downarrow M_X$ can be viewed as the most detailed outcome intention of the agent, or his 'ends', M_A can be seen as a collection of 'means' to achieve it. Each element of M_A represents a particular action intention of the agent. Clearly, it is natural to require that endogenous consistency also applies here.

Postulate 2 (Endogenous Consistency of Action Intentions) *For all $p, p' \in M_A$, $p \cup p' \in M_A$.*

The postulate precludes an agent from having two action intentions that are not executable in a single run. Suppose, for instance, that in the decision tree of Figure 1, the agent has the action intention $p \in M_A$, where p is defined only on $\varnothing$ and gives a_2. A partial plan p' which

[6] Or that intentions are 'agglomerative' in terms of [Bratman, 2006a] and [Velleman, 2003].

is defined only on (a_1, tail) cannot then be an element of M_A. Observe, finally, that given Postulate 2, M_A contains a 'largest' (but still maybe partial) plan of action, that is some p that encompasses all other plans in M_A. We denote this plan by $\uparrow M_A$.

The next postulate, means-end consistency, specifies which kind of connection should hold between action and outcome intentions.

Postulate 3 (Means-End Consistency) *There is some $p \in M_A$ such that $\pi(p) \subseteq \downarrow M_X$.*

The postulate imposes constraints both on the outcome and the action intentions. Together with Postulate 1, it implies that M_A is never empty. It also aligns the action intentions to the outcome intentions, saying that there should be at least one partial plan of action which, if enacted, ensures the agent will obtain some outcome in $\downarrow M_X$. Together with Postulate 2, Postulate 3 makes the connection between action and outcome intentions even tighter, as shown in the following:

Fact 4.1 *For any $\mathcal{I}$ that satisfies Postulate 2 and 3, $\pi(\uparrow M_A) \subseteq \downarrow M_X$.*

Proof. Take any such $\mathcal{I}$. Observe that for any (maybe partial) plan of action, $p \subseteq p'$ implies that $\pi(p') \subseteq \pi(p)$. Now, because we assume that $\mathcal{I}$ satisfies Postulate 2, we know that for all $p \in M_A$, $p \subseteq \uparrow M_A$. Also, since $\mathcal{I}$ satisfies Postulate 3, we know that there is a $p \in M_A$ such that $\pi(p) \subseteq \downarrow M_X$, and thus that $\pi(\uparrow M_A) \subseteq \downarrow M_X$. QED

This is indeed a tight connection. It could be objected that it seems to exclude the possibility of *conditional* action intentions. Consider a variation of the decision problem presented in Figure 1 in which $\pi(a_1, \text{tail}, a_4) = 200$ instead of 100. In this case the agent has a clear favourite outcome. Now observe that, in virtue of Postulate 3, the agent cannot intend to realize this outcome. That is, he cannot have $\downarrow M_X = \{200\}$. One fairly intuitive plan of action for reaching this outcome would, however, be a conditional one: *if* he makes it to $(a_1, tail)$, then he will act in order to get 200. Now the best he can do to reach $(a_1, tail)$ is indeed to choose a_1 and, if tails subsequently comes up, a_4. That is $\uparrow M_A = (a_1, a_4)$. But this plan is not means-end consistent since $\pi(\uparrow M_A) = \{100, 200\}$, which is not a subset of $\downarrow M_X$.

At first sight, this conclusion is rather unwelcome. After all, the conditional plan (a_1, a_4) seems perfectly reasonable. But one has to realize that the relation between M_A and M_X enforced by Postulate 3 goes both ways. Not only does it 'align' the means with the intended ends, it also precludes the agent from intending ends that he cannot force. In other words, Postulate 3 constrains both M_X and M_A. In the example of the previous paragraph, even though 200 is the most preferred outcome, it is not an outcome that the agent can ensure. Indeed, whereas it is perfectly reasonable to say that a person has the intention of buying a lottery ticket, it seems far less so to say that he has the intention of being the winner of the lottery. To paraphrase [Bratman, 1987, p.31], the outcome intention cannot be appropriately filled by action intentions and Postulate 3 thus imposes a feasibility constraint on the outcome intentions.

It should be noted that these postulates do not preclude the agent from forming action-intentions for their own sake. What is more, as hinted above, $\uparrow M_A$ need not be a full plan of action. It can also be partial. As the planning theory of intentions maintains, the action intentions of the agent will typically be incomplete even when they satisfy the three postulates. Means-end consistency 'only' requires that some intended outcomes be secured, and this can leave a lot of choice nodes undecided. That is, the plan can well remain silent on what choices will be made once the outcome intentions are within reach. For agents with limited capacity, there is no doubt that this feature is an asset. They have only to make up their mind in advance on a limited number of choice points, leaving the other decisions for later. They indeed 'cross the bridge when they come to it'. But what about ideal agents? From what we said in Section 3.2, it should be clear that such a policy is rather pointless. They are capable of computing *in advance*, for any decision problem, a maximally detailed plan, and they will be willing to carry it out all the way.[7] In the famous words of Savage [Savage, 1954, p.83], they are perfectly capable of pushing to its extreme the 'look before you leap' approach:

> Making an extreme idealization, [...] a person has only one decision to make in his whole life. He must, namely, decide how to live, and he might in principle do once and for all.

In that context, it seems to us that ideal agents have little to do with partial plans. This point has also been made by [von Neumann and Morgenstern, 1944, p.79], who mention that the only assumption needed for agents to be able to choose full strategies 'is the intellectual one to be prepared with a rule for behavior for all eventuality.' This, they say, 'is an innocent assumption within the confines of a mathematical analysis'. Whether or not one agrees with the 'innocent' character of this assumption, the point remains. For ideal agents, there is nothing that stands in the way of choosing beforehand a full plan of action. We thus think it is natural to assume that intentions of ideal agents reduce to plans of actions.

Postulate 4 (Actions Intentions for ideal agents) *$\uparrow M_A$ is a plan of action.*

This does not yet reduce intentions to full strategies, since strategies also provide information about the courses of action an agent would undertake in case he deviates from what he decided to do. However, as we indicated earlier, we think that this restriction is harmless in the context of decision theory with ideal agents.

But even though intentions reduce to plans of actions for the ideal agent, we do not think they are useless additions to existing decision theoretic models. In [van Hees and Roy, 2007], we showed that outcome-intentions can be used to 'focus' or 'break ties' between equally desirable strategies. This analysis can be carried into to the present framework. Consider the following postulate:

[7] The argument for this last point has been made by several authors in decision and game theory. See especially [Myerson, 1991, p.11], [McClennen, 1990] and [Gauthier, 1997] for discussions of such cases.

Postulate 5 (Intentions and Expected Payoff Compatibility) *For all plans for action p, p' such that $EV(\varnothing, p) > EV(\varnothing, p')$: if $p \notin M_A$, then $p' \notin M_A$.*

We can now easily establish the following result:[8]

Proposition 4.2 *For any decision tree T and intention structure $\mathcal{I}$ that satisfies Postulates 1 – 6, there is one and only one plan that coincides with all (partial or non-partial) plans in M_A and that also maximizes expected value.*

The proposition shows that the agent's intentions are compatible with expected value maximization. This is not at all surprising, being in essence what Postulate 6 says. But note that it is also possible that a combination of intentions and traditional expected value maximization will lead to a focus on some solutions when there is more than one plan that maximizes expected value. Proposition 4.2 is thus in line with something philosophers of action have been claiming for a long time: intentions are key anchors to one-person sequential decision making.

A limitation of the framework is, of course, its restriction to situations of *perfect information*. At any choice node the agent is aware of what happened before. Generalizing to a decision tree with imperfect information is indeed an interesting enterprise, in which the notion of exogenous consistency should play a much greater role, but we shall not pursue that route here. Instead, we argue that there are two fundamental shortcomings to the analysis presented thus far, regardless of the idealized context of a fully rational person and complete information. Both shortcomings arise from the fact that intentions can have *autonomous consequences*, that is, they may affect an agent's appraisal of the outcome *independently* of the action to which they commit.

5 Intentions and Counterfactuals

To study the distinction between intentions and expected side effects, Michael Bratman discusses the following example:

> Both Terror Bomber and Strategic Bomber have the goal of promoting the war effort against Enemy. Each intends to pursue this goal by weakening the Enemy, and each intends to do that by dropping bombs. Terror Bomber's plan is to bomb the school in Enemy's territory, thereby killing children and terrorizing Enemy's population. Strategic Bomber's plan is [...] to bomb Enemy's munitions plant. [He] also knows, however, that next to the munitions plant is a school, and that when he bombs the plant he will also destroy the school, killing the children inside.[Bratman, 1987, p.139]

[8] The proof consists only of unpacking the definitions, and is therefore omitted.

How should we model the decision problem? Note that we have assumed thus far that an agent already has the relevant outcome- and action-intentions. If we make the standard assumption – as we have thus far – that a plan of action describes the available moves, then the tree is very simple (see Figure 2). There are only two possible plans – and also only two strategies – namely 'Bomb' and 'Not Bomb'. But the consequence of bombing may be

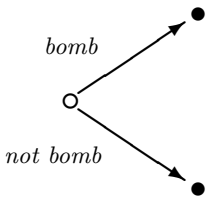

Figure 2: The Bombing Problem

different if it realizes an intention to kill the children than if it does not not. For instance, Terror Bomber might be prosecuted for war crimes if it was indeed his intention to kill the children, whereas such prosecution may be less likely for Strategic Bomber. In this scenario, the payoffs are thus not only dependent on which terminal history is reached but also on the intention with which it is reached. Be that as it may, our current model cannot distinguish between bombing with or without the intention of killing the children. For both cases we will have the same action intention set $\uparrow M_A = \{\text{bomb}\}$ and value of $\pi(\text{bomb})$, even though the intentions and the payoffs, by assumption, are different.

Bratman argues, in essence, that Strategic Bomber does not have the intention of killing the children because, contrary to Terror Bomber, he *would* not adopt a new plan of action *if* the children were moved somewhere else, far from the munitions plant. That is, information about counterfactual events reveal the intentions of agents. We define a *counterfactual extension* of a decision problem as any decision tree that starts with a chance node and in which the original decision problem is one of the subtrees following that chance node. An example is given by Figure 3. At the first node it is determined whether the school will be deserted or not. If not, one faces the original decision problem, otherwise the counterfactual scenario arises. If we now consider this extended tree, we can assume that the plans of action of Terror and Strategic Bomber will differ. Terror Bomber's $\uparrow M_A$ will be the plan 'Only bomb when children are at school' whereas for Strategic Bomber it will be 'Always bomb'. Following Bratman's suggestion, we can use the counterfactual information carried by these plans to assign the payoff to the terminal histories.

Let ρ be a *refined payoff function* that assigns a real-valued payoff to *pairs* (h, p) where p is a plan of action and h is a terminal history reachable from the root by p. An intention or a plan of action p has *autonomous consequences* when, for a given history h and another plan p' that coincides with p on h, $\rho(h, p) \neq \rho(h, p')$. Observe that in such a case the intention does indeed have autonomous consequences. The agent, in the course of reaching h, accomplishes

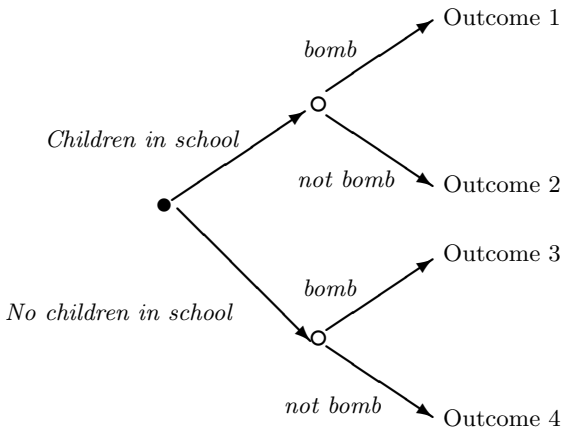

Figure 3: A Counterfactual Extension of the Bombing Problem

the same action whether he follows p or p'. He would have acted differently only outside h, that is if things had turned out differently.

This is precisely what happens in the counterfactual version of the bombing case. The payoff $\rho(\text{Outcome 1}, \text{Always bomb})$ does not equal $\rho(\text{Outcome 1}, \text{Only bomb when children are at school})$; dropping bombs with the intention of killing the children has different consequences from dropping bombs without this intention.

Of course, the autonomous consequences of acting with a certain intention should be taken into account when the agent chooses his plan of action. We thus redefine the expected value of a plan (or a strategy) in terms of the refined payoff function ρ in a straightforward way. For all pairs (h, p) such that h is reachable from $\varnothing$ by choosing according to p:

$$EV'(h,p) = \begin{cases} \rho(h,p) & \text{If } h \text{ is a terminal history} \\ \Sigma_{a \in A(h)} \delta(a) EV'(ha,p) & \text{If } h \text{ is a chance node} \\ EV'(hp(h),p) & \text{If } h \text{ is a choice node} \end{cases}$$

Refined payoff functions are indeed generalizations of the standard functions π. The latter are functions for which the value at a terminal history is not dependent on the plan with which it is reached. That is, any standard payoff function π can be emulated by a refined payoff function ρ for which $\rho(h,p) = \rho(h,p')$ for all terminal histories h and plans p and p' that are defined on h.

From a decision-theoretic point of view, the question, of course, is how to establish the conditions under which the agent's preferences can be represented by such a refined payoff function. After all, it may be rather difficult to determine what the appropriate counterfactual extension is. We do not address this question here. Our aim was merely to point out the necessity of such an extension of the standard decision theoretic model if we want to take account of the fact that intentions can have autonomous effects.

6 The Formation of Intentions

Postulate 6 establishes a relation between a person's intentions and his preferences. It states that certain intentions will not be formed given the preferences of the agent. Apart from this postulate we have not made any further assumptions concerning the formation of intentions. We have modeled *acting with* a certain intention, but not the *formation* of intentions. Yet the possibility that the formation of intentions has autonomous consequences poses additional modelling problems that necessitate a further departure from the standard decision-theoretic framework. In particular, it poses a problem with respect to Postulate 6.

Our analysis is again driven by an example from the philosophical literature. In [Kavka, 1983], the now famous 'Toxin Puzzle' was introduced:[9]

> You have just been approached by an eccentric billionaire who has offered you the following deal. He places before you a vial of toxin that, if you drink it, will make you painfully ill for a day, but will not threaten your life or have any lasting effects [...]. The billionaire will pay you one million dollars tomorrow morning if, at midnight tonight, you *intend* to drink the toxin tomorrow afternoon. He emphasizes that you need not drink the toxin to receive the money; in fact, the money will already be in your bank account hours before the time for drinking it arrives [...]. All you have to *do* is [...] to intend at midnight tonight to drink the stuff tomorrow afternoon. You are perfectly free to change your mind after receiving the money and not drink the toxin. (The presence or absence of the intention is to be determined by the latest 'mind reading' brain scanner.) [...] Arrangement of such as external incentives is ruled out, as are such gimmicks as hiring a hypnotist to implant the intention, forgetting the main relevant facts of the situation, and so forth. [Kavka, 1983, pp.33–34]

This scenario involves two choice points: some time before midnight when you have to decide whether to form the intention or not, and tomorrow morning, when you have to decide whether to drink or not. The first choice point is an 'intention-formation' point. Such choice points are notably absent from the models we have used thus far. We have described agents who come 'already equipped' with some future-directed intentions; the moment of intention formation is not part of the decision problem. Even with the appropriate refined payoff function, we can only model the choice situation of a person who already has formed the intention or who did not form the intention to drink the toxin. As a consequence, the Toxin Puzzle cannot be captured in the present model.[10]

[9] See e.g. [Mele, 1992], [Mele, 1995], and discussions in [Coleman and Morris, 1998].

[10] The reader might wonder why, in the first place, is it so important to capture the Toxin Puzzle. For us this is because it has far more general features than Kavka's far-fetched scenario suggests. As noted in [Bratman, 1999, p.67], the kind of autonomous effects displayed here are essential to multi-agent *reciprocation problems*, on which game theorists have glossed at length (see e.g. [Rosenthal, 1982], [Binmore, 1996] and

Let us therefore expand our framework by allowing for the fact that the formation of an intention to perform an action a now itself forms a possible move within the tree. In such an approach, the decision tree corresponding to the toxin puzzle will be as in Figure 4 (we assume that no drink is offered if you have not formed the intention).[11]

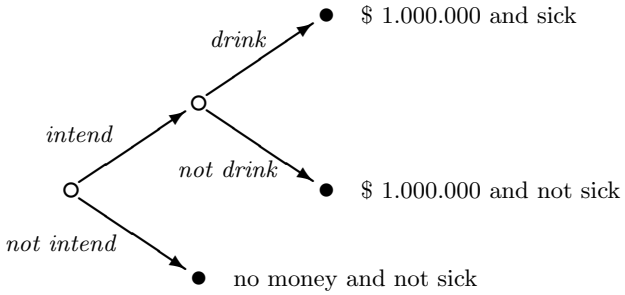

Figure 4: The Toxin Puzzle with Intentions as Moves

If the formation of intentions is itself seen as a move, an element of M_A, say partial plan of action p, need not only describe a person's intention to perform a certain action but also an intention to adopt an intention. For instance, the possible element 'drink the toxin' of M_A describes a regular action-intention, viz. the intention to drink the toxin. Having 'intend to drink' in one's M_A would stand for the second-order intention of intending to form the intention to drink. The plan of action $p =$ (form the intention to drink, drink) would stand for the combination of the first and second-order intention, etc. Given a (partial) plan, we denote a move which consists of the formation of the intention to some action a by $\vec{a}$. Since we only focus on future-directed intentions, we assume in what follows that one can form an intention to do some a only at choice nodes h preceding (either directly or indirectly) the choice node h' at which the action can be carried out. We also avoid some technical complications by assuming that for any node h there is at most one such h' at which a can be carried out.

It is reasonable to impose the following additional requirement of endogenous consistency:

Postulate 6 (Consistency Between First and Second Order Intentions) *For all (partial) plans for action $p \in M_A$, all h, h' and all a, if $\vec{a} \in p(h)$ and $a \in A(h')$, then $a \in p(h')$.*

The postulate is in line with a feature identified by the planning theory of intention: the fact that intentions are 'conduct-controlling' [Bratman, 1987, p.16]. It states that a person who

[Aumann, 1998]). Switching to game theory would go far beyond the scope of this paper. But as far as the Toxin Puzzle is, so to speak, a one person reciprocation problem, it brings into decision theory an important aspect of planning agency.

[11] This route was first explored in [Verbeek, 2002].

has a second-order intention to perform the action a also has the first-order intention to do so. In other words, if a rational person intends to take up the commitment to do a (intends to intend to do a), then he cannot intend simultaneously to break that commitment at some later point in time (intend not to do a). Of course, it frequently happens that we revise our intentions when some *unforeseen events* make them unachievable. Once again, however, we refer to the assumption of ideal agency and perfect information, which rules out these possibilities.

We can now capture explicitly the core of the tension between commitment taking and payoff maximization that is at the heart of the Toxin Puzzle. Indeed, looking again at Figure 4, we quickly see that Postulate 5 and Postulate 6 cannot be satisfied together. If we assume that $\mathcal{I}$ satisfies the former, we will have $\downarrow M_X = \{(\$1.000.000 \text{ and not sick})\}$, and thus $\uparrow M_A = (\vec{\text{drink}}, \text{not drink})$, against Postulate 6. But, in turn, the only action intention consistent with Postulate 6 where the agent goes up at the first node is $\uparrow M_A = (\vec{\text{drink}}, \text{drink})$, which clearly violates Postulate 5.

Note, however, that in our rendition of the toxin puzzle, the crux of it is not that our (first-order) intentions may clash with our preferences, but rather that consistency between our first-order and second-order intentions clashes with our preferences. Rationality requirements that concern the structure of our intentions (Postulate 6) may be at odds with the rationality requirement that we maximize our utility (Postulate 5). Finally, we note that this tension is not bounded to the eccentric scenario of the toxin puzzle. The following result shows that whenever we assume that individuals can indeed decide to form intentions, if these intentions can have autonomous effects, then one can devise a payoff function such that it is never rational to follow up on the commitments that one undertook, thus violating Postulate 6.

Proposition 6.1 *For any decision tree and any intention structure $\mathcal{I}$ such that for some nodes h, h' and some a, b ($a \neq b$), $\vec{a} \in A(h)$ and $a, b \in A(h')$: there is a refined payoff function ρ such that Postulates 2, 4, 5 and 6 cannot be satisfied simultaneously.*

Proof. Let h and h' be as defined, and take all plans p such that both h and h' are reachable from the root by choosing according to p. Let H_p be the set of all terminal histories reachable from h' by playing according to such p and in which $p(h') = b$. All we need is to fix $\pi(h', p') = 1$ for all $h' \in H_p$ and p' compatible with h', and $\pi(h', p') = 0$ for all other terminal histories h' and p' compatible with it. Clearly, in this decision problem no intention structure will satisfy all of the postulates. QED

7 Conclusion

To summarize, we have argued in this section and in the previous one that intentions with autonomous consequences pose difficulties that force us to go beyond traditional models of decision making. Of course, this does not mean that *all* intentions require such extensions.

Intentions without autonomous consequences can readily be described in terms of strategies in extensive decision problems, as we argued in Section 4. However, as we hope we have shown, even the incorporation of such 'standard' intentions may form an enrichment of the decision theoretic framework.

References

[Anscombe, 1957] Anscombe, G. (1957). *Intention*. Harvard University Press, Harvard University Press.

[Aumann, 1998] Aumann, R. J. (1998). On the centipede game. *Game and Economic Behavior*, 23(1):97–105.

[Binmore, 1996] Binmore, K. (1996). A note on backward induction. *Games and Economic Behavior*, 17(1):135–137.

[Bratman, 1987] Bratman, M. (1987). *Intentions, Plans and Practical Reasons*. Harvard UP, London.

[Bratman, 1999] Bratman, M. (1999). *Faces of Intention; Selected Essays on Intention and Agency*. Cambridge UP.

[Bratman, 2006a] Bratman, M. (2006a). Intention, belief, practical, theoretical. Unpublished Manuscript, Stanford University.

[Bratman, 2006b] Bratman, M. (2006b). *Structures of Agency: Essays*. Oxford UP.

[Coleman and Morris, 1998] Coleman, J. L. and Morris, C. W., editors (1998). *Rational commitment and social justice : essays for Gregory Kavka*. Cambridge University Press.

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

[Gigerenzer and Selten, 2002] Gigerenzer, G. and Selten, R. (2002). *Bounded Rationality: The Adaptive Toolbox*. MIT Press.

[Hammond, 1976] Hammond, P. J. (1976). Changing tastes and coherent dynamic choice. *The Review of Economic Studies*, 43(1):159–173.

[Harman, 1986] Harman, G. (1986). *Change in View*. MIT Press.

[Kavka, 1983] Kavka, G. S. (1983). The toxin puzzle. *Analysis*, 43(1):33–36.

[McClennen, 1990] McClennen, E. F. (1990). *Rationality and Dynamic Choice : Foundational Explorations*. Cambridge UP.

[Mele, 1992] Mele, A. (1992). Intentions, reasons, and beliefs: Morals of the toxin puzzle. *Philosophical Studies*, 68(2):171–194.

[Mele, 1995] Mele, A. (1995). Effective deliberation about what to intend: Or striking it rich in a toxin-free environment. *Philosophical Studies*, 79(1):85–93.

[Myerson, 1991] Myerson, R. B. (1991). *Game Theory: Analysis of Conflict*. Harvard UP, 1997 edition.

[Osborne and Rubinstein, 1994] Osborne, M. J. and Rubinstein, A. (1994). *A Course in Game Theory*. MIT Press.

[O'Shaughnessy, 1973] O'Shaughnessy, B. (1973). Trying (as the mental "pineal gland"). *The Journal of Philosophy*, 70(13, On Trying and Intending):365–386.

[Rosenthal, 1982] Rosenthal, R. (1982). Games of perfect information, predatory pricing, and the chain store paradox. *Journal of Economic Theory*, 25:92–100.

[Rubinstein, 1998] Rubinstein, A. (1998). *Modeling Bounded Rationality*. MIT Press.

[Savage, 1954] Savage, L. J. (1954). *The Foundations of Statistics*. Dover Publications, Inc., New York.

[Searle, 1983] Searle, J. (1983). *Intentionality*. Cambridge UP.

[Selten, 1975] Selten, R. (1975). Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory*, V4(1):25–55.

[Sen, 1977] Sen, A. (1977). Rational fools: A critique of the behavioral foundations of economic theory. *Philosophy & Public Affairs*, 6(4):317–344.

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

[van Hees and Roy, 2007] van Hees, M. and Roy, O. (2007). Intentions and plans in game and decision theory. In Verbeek, B., editor, *Reasons and Intentions*. Ashgate. Available on the ILLC prepublication repository (PP-2006-54).

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

[Verbeek, 2002] Verbeek, B. (2002). *Moral Philosophy and Instrumental Rationality: An Essay on the Virtues of Cooperation*. Kluwer Academic Publishers.

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