Commitment-Based Decision Making for Bounded Agents

Olivier Roy[⋆]

Institute for Logic, Language and Computation
Plantage Muidergracht 24
1018TV Amsterdam
The Netherlands
oroy@science.uva.nl

Abstract. This paper explores ways to introduce commitment, as conceived in terms of plans and intentions, into decision theory. An intention-based representation theorem is proved, along with a proposal to model how plans simplify decision making for resource-bounded agents.

1 Introduction

In this paper I adapt the decision theoretical framework proposed in [Anscombe and Aumann, 1963] (AA) to represent formally rational decision-making by agents who are *committed* to certain plans of action. In Section (2) I present the decision theoretical formalism, and Section (3) I sketch the theory of planning agency I'm relying on. In Section (4), I envisage two ways of introducing plans in the formalism. First, I propose to make explicit how plans help resource-bounded agent to simplify decision-making. Second, I show that one can easily use plans to build an *intention-based* utility ranking that is independent of the usual "preference-based" ranking. To facilitate the reading, all the proofs have been moved to the Appendix.

This is not the first attempt at formalizing plan-based decision making. [Bratman et al., 1991] is probably one of the best known. Beside being much closer to decision theory, the framework I propose accounts for one function of plans (modeled in Section 4.2) that is left out in their approach. [van Hees, 2003] also proposed something similar to what I do in Section 4.3, but in a totally qualitative framework. Although I won't explore them here, I think that there are interesting connections to be made between the present work and [McClennen, 1990].

2 Decision making under uncertainty

Decision theory is often called the study of *rational* decision making under *risk* or *uncertainty*. Typically, "uncertainty" is thought of as *subjective*, that is, referring to the partiality of the decision-maker's information. Whereas "risk" is *objective*, that is, about the outcome of certain random events such as a lottery. In the decision theoretical framework I will use, both risk and uncertainty are modeled. I rely heavily on the presentation of [Myerson, 1991, chap.1], but the framework goes back at least to [Anscombe and Aumann, 1963]. Of course, this is neither the only nor the simplest decision theoretical approach available. See for example [Luce and Raiffa, 1957], where a more classical model is introduced, and further (early) references can be found. I have chosen to work within the AA framework because it models both random happenings (objective probabilities) and uncertainties (subjective probabilities).

[⋆] I am grateful to Richard Bradley, the LRBA referees, Johan van Benthem, Mikaël Cozic and the attendees at the PALMYR-2 workshop held in Paris in October 2006 for their useful comments.

Given an arbitrary set A, I will use $\Delta(A)$ to denote any probability distribution on A. A *decision problem*, $\mathcal{DP}$, is a tuple $\langle X, \Omega, L, \Xi, Pref \rangle$ such that:

- X and Ω are both finite sets of *prizes* and *states*, respectively. I will use x, y, z to range over elements of X and $t, t', t_1, t_2...$ to range over elements of Ω.
- $L = \{f : \Omega \to \Delta(X)\}$ is the set of all *lotteries* over the prizes in X. $f(x|t)$ should intuitively be read as "the probability of getting prize x given that the current state is t". $[x]$ denote the lottery that gives x for sure, in all states. That is, $[x] = f$ such that $f(x|t) = 1$ for all t. These lotteries are intended to represent random happenings *in the world*. They are "objective uncertainties" or "roulette lotteries" in AA terminology. These contrast with "horse lotteries" or "subjective uncertainties", which represent the agent's beliefs about the true state of the world. In the AA framework, the firsts are sharply distinguished from the seconds. Subjective uncertainties are captured by $\Xi = \{S | S \subseteq \Omega \ \& \ S \neq \varnothing\}$, the set of events.
- $Pref : \Xi \to \mathcal{P}(L \times L)$ is a function that gives a weak linear preference ordering on lotteries, given an event in Ξ. I will follow Myerson and write $f \preceq_S g$ to say that g is at least as good as f, given that the true state of the world is in S. $f \sim_S g$ and $f \prec_S g$ are defined as usual.

The crux of the AA decision-theoretical framework is to show that, given certain axiomatic constraints, decision problems can be represented by a quantitative utility function and a probability distributions where the preferred lotteries are exactly those that maximize expected utility. More precisely, given a decision problem $\mathcal{DP}$, a *conditional-probability function* $p : \Xi \to \Delta(\Omega)$ is a function that gives the probability of a state t given that an event $S \in \Xi$ occurs. This will be the probabilistic representative of the agent's subjective uncertainty. We constraint p as follows, for any t and S:

$$p(t|S) = 0 \text{ if } t \notin S \text{ and } \sum_{r \in S} p(r|S) = 1$$

Now, an *utility function* is any function $u : X \times \Omega \to \mathbb{R}$. This function is used to compute the *expected utility value* of a lottery f given an even S, $E_p(u(f)|S)$, as follows:

$$E_p(u(f)|S) = \sum_{t \in S} p(t|S) \sum_{x \in X} u(x,t) f(x|t)$$

With these definitions in hand, the representation theorem is proven via the method of [Savage, 1954], where the agent's probabilistic beliefs and utility are extracted from his (event-based)-preferences over objective lotteries. Below are Myerson's axioms and statement of the representation theorem. See [Myerson, 1991, p.14-17] for details of the proof.

1. (a) *(Completeness)* $f \preceq_S g$ or $g \preceq_S f$.

   (b) *(Transitivity)* If $f \preceq_S g$ and $g \preceq_S h$ then $f \preceq_S h$.
2. *(Relevance)* If, for all $t \in S$, $f(\bullet|t) = g(\bullet|t)$ then $f \sim_S g$.
3. *(Monotonicity)* If $f \preceq_S g$ and $0 \leq \beta < \alpha \leq 1$ then $(1-\beta)f + \beta g \prec_S (1-\alpha)f + \alpha g$.
4. *(Continuity)* If $h \preceq_S g$ and $g \preceq_S f$ then there exists a number γ such that $0 \leq \gamma \leq 1$ and $g \sim_S \gamma f + (1-\gamma)h$.

5. (a) *(Weak Objective substitution)* If $f \preceq_S e$, $h \preceq_S g$ and $0 \leq \alpha \leq 1$, then $\alpha f + (1-\alpha)h \preceq_S \alpha e + (1-\alpha)g$.
   (b) *(Strict Objective substitution)* If $f \prec_S e$, $h \preceq_S g$ and $0 < \alpha \leq 1$, then $\alpha f + (1-\alpha)h \prec_S \alpha e + (1-\alpha)g$.
6. (a) *(Weak Subjective substitution)* If $g \preceq_S f$, $g \preceq_T f$ and $S \cap T = \varnothing$ then $g \preceq_{S \cup T} f$.
   (b) *(Strict Subjective substitution)* If $g \prec_S f$, $g \prec_T f$ and $S \cap T = \varnothing$ then $g \prec_{S \cup T} f$.
7. *(Non-triviality)* For all $t \in \Omega$ there exist $y, z \in X$ such that $[z] \prec_{\{t\}} [y]$

Theorem 1. *Given a decision problem $\mathcal{DP}$, the following are equivalent:*

1. *$\mathcal{DP}$ satisfies the axioms enumerated above.*
2. *There exists a utility function u and a conditional probability function p such that :*
   (a) For all $t \in \Omega$, $\max_{x \in X} u(x,t) = 1$ and $\min_{x \in X} u(x,t) = 0$.
   (b) For all R, S, T such that $R \subseteq S \subseteq T \subseteq \Omega$ and $S \neq \varnothing$, $p(R/T) = p(R/S)p(S/T)$.
   (c) For all $f, g \in L$ and $S \in \Xi$, $g \preceq_S f$ iff $E_p(u(g)/S) \preceq E_p(u(f)/S)$.

Now, suppose a decision maker is somehow committed to act in a certain way. It is highly debatable whether such commitments can be straightforwardly represented in this framework, for example by changing the preferences or utility assignments. What is more, if one thinks this *can't* be done, it is not clear how to adapt the framework to capture the notion of commitment. Of course, all this depends on what is meant by "the agent is committed to x". In the next section, I will present one understanding of commitment, based on the notion of *plans of action*, and try to explain why one may want to introduce them as an independent feature in the AA framework, especially when one has resource-bounded agents in mind.

3 Plans of Action and Committed Agency

In a nutshell, the driving idea runs as follows: for agents with limited time and computational capacities, being able to commit to plans of action is useful. Of courses, there are many things an agent can be committed to. Here I will restrict to commitment in terms of previously adopted intentions and plans. Moreover, the content of these plans and intentions, what an agent is committed to, will be identified with the agent's *goals*. So, by saying that "the agent is committed to x" I will just mean "he has the intention to get x", "he aims for x" or "x is his goal". Granted, this blurs the relationship between plans and goals, as well as other forms of commitments[1]. I see the present proposal as a first step, upon which a more conceptually fine-grained analysis could be based.

The idea that plans of action are useful for bounded agents has been strongly advocated by [Bratman, 1987]. Since my goal here is not to argue for Bratman's views, but rather to propose a way to implement it into decision theory, its main insights will be uncritically reviewed here, with an outright focus on the different functions of intentions and plans.

There is a pertaining distinction in action theory between *future-directed* intentions and intentions *in action*. The latter intentions are thought as the "mental components" of actions (see e.g. [O'Shaughnessy, 1973])

[1] Although Section (4.3) is much less bound to this restriction of commitment to intentions.

that provide "on the fly" guidance toward their own accomplishment. As their name suggests, future-directed intentions are rather about middle- and long-term goals, and are generally viewed as the output of "beforehand" deliberation. Bratman's work is concerned with future-directed intentions, and so will be mine. Thus, for the remaining of this paper, "intention" should just be understood as future-directed.

According to Bratman intentions are mental states that provide a relative *stability* of goals and *commit* agent to their achievement. Like preferences or desires, intentions can be *reason* for action. But Bratman has argued that intentions and desires are irreducible to one another and thus that the belief-desire view on rational choice is at best an incomplete account of decision-making. In this paper, I will *assume* irreducibility of intentions and ask how they can be introduced in decision theory.

Plans, are viewed as hierarchically structured but typically incomplete sets of intentions. "Hierarchically structured" means that a plan contains general intentions upon which more specific intentions are subordinated. However sharp, the most precise intention of a plan needs not to specify in every detail how it shall be carried out. This is why plans are typically incomplete, a feature which is arguably crucial for agents with limited time and computational resources.

Plans are constrained by norms of rationality. They are first required to be *intentions*- as well as *beliefs*-consistent. The intentions they are made of should not contradict each other, and neither should they contradict the agent's beliefs. In Bratman's view[2], intention consistency is grounded in the fact that intentions are *agglomerative*: if one intends that x and intends that y, he must intend that x and y. Note that this distinguishes intentions from desires: we can have contradictory desires but not contradictory intentions, and desires are clearly not agglomerative. Rational plans also call for "means-end coherence": if one intend to x he should also, at some point, come to intend some necessary means to x. Given that plans are typically incomplete, this means that plans "drive" agents towards deliberating on how they will achieve them.

So, in this model, plans of action are not only outputs but also *input* of deliberation. It is as inputs, so Bratman's theory goes, that they help resource-bounded agents to simplify decision problems. Suppose an agent with a plan $\mathfrak{P}$ is facing a certain decision problem. The pressure for intention consistency will *rule out of consideration* options that preclude the achievement of $\mathfrak{P}$. On the other hand, means-end coherence will call for adopting one of the options that promote the achievement of $\mathfrak{P}$. But it does more than that, given that deliberation is itself an activity that takes time and energy. For resource-bounded agents, it is likely that pondering endlessly over small details of the available options will itself prove to be an obstacle to the achievement of $\mathfrak{P}$. The other way around: careful management of the time and energy devoted to deliberation might prove to be, in itself, an important mean to achieve one's goals. As such, it doesn't seem unreasonable to suppose that means-end coherence will *fix the level of detail up to which options are going to be considered*, thus avoiding useless mind boggling. These two norms, intention-consistency and means-end coherence, are thus likely to simplify deliberation, an advantageous prospect for limited agents.

To sum up, Bratman proposes a "build up" view of planning agency and practical reasoning. Suppose that at some point an agent form a plan $\mathfrak{P}$ made of the intention i_1 to obtain the goal X later. As some opportunities show up he will have to choose among various attainments of $\mathfrak{P}$, select one of them, say X',

[2] A view that is also endorsed by other action theorists, such [Velleman, 2006] and [Wallace, 2003]

form the new intention i_2 to get X' and add it to $\mathfrak{P}$. $\mathfrak{P}$ will be thus enriched until it is achieved[3]. In the next sections, I will consider, in turn, how to model the effects of intention-consistency and means-end coherence on rational *plan-based* decision making under uncertainty. To my knowledge, such proposal is an original contribution. As a side issue, I will also show how one can use plans and intention to, so to speak, set a decision agenda that isindependent of the preferences. As far as this paper goes, all these models will be *static*. I leave for further work the task of modeling the dynamics of plan update.

4 Decision making with commitments

4.1 Plans of action: a formal picture

In the formalization of utility theory sketched in section (2), the object of choices are *lotteries* but, ultimately, what the agents aim for are the *prizes*. Consequently, it seems reasonable to restrict the content of commitments to prizes. That is, the formal picture I am going to depict is one where agents form intentions to get some prizes, given that a certain event S occurs. Given the view of intentions endorsed here, the intended prizes are *goals* the agent will work for. Note that in the AA framework, prizes are exclusive outcomes, so to speak. In the end, the agent gets *one* of the prizes in X. When I say that the subset X' of X is a goal for the agent, I mean that he will act in order to get *one of them*.

In the following definitions, intentions are assimilated with their content and so plans are viewed simply as sets of sets of prizes. So, given an event S, a plan $\mathfrak{P}_S$ is any set of intentions, i.e. of subsets of X. Now, belief and intention-consistency already impose constraints on what is to be considered a *rational* plan. First, belief-consistency precludes impossible intentions, which are just empty sets of prizes. So we are to require that, for any event S, $\varnothing \notin \mathfrak{P}_S$. Intention-consistency requires that the intentions that constitute a plan don't exclude each other. This will boil down to require that any two intentions in a plan have a non-empty intersection. But recall that, for Bratman, intention-consistency in grounded in the agglomerativity of intention, a property that has a clear formal counterpart: $\mathfrak{P}_S$ will be required to be closed under intersection. It should be clear that intention-consistency follows from the requirements that $\varnothing \notin \mathfrak{P}_S$ and that $\mathfrak{P}_S$ is closed under intersection,

Note that, because X is finite, we automatically get that $(\bigcap_{X \in \mathfrak{P}_S} X) \in \mathfrak{P}_S$. This will correspond to the most precise intention of the plan. For most of the results presented below, we can consider without loss of generality that any plan $\mathfrak{P}_S$ contains only this minimal element, and we will use a special notation for it: $\downarrow\mathfrak{P}_S$. A plan is incomplete if $\downarrow\mathfrak{P}_S$ is not a singleton. When the context makes it obvious, we will omit the event subscript S.

Note also that these simple set-theoretical requirement gives already a kind of hierarchical structure to the plan. Indeed, rational plans are semi-lattices with respect to the inclusion relation, with $\downarrow\mathfrak{P}_S$ the smallest element.

[3] This is indeed a very coarse picture of Bratman's theory of practical reasoning. Many important issues are ignored. Among them are the crucial difference between deciding upon an attainment and adopting an intention - as illustrated by the famous "terror bomber" example - and the subtleties of plan reconsideration. These issues will be ignored in the proposed formalization also. They are thoroughly explored in [Bratman, 1987].

4.2 Consistency and coherence to simplify decision problems

As mentioned above, it seems that some decision problems are just too complex to be handled by agents with limited computational and representational capacities. In the usual decision theoretical models I introduced above, it can even be argued that it is simply impossible to do so. The object of choices being the lotteries, the actual choice set is the whole *uncountable* lottery space L. So, for limited agent, we can assume that, somehow, they don't consider all their options, and the norms of intention-consistency and means-end coherence provide an explanation of *why* it is so.

Intention consistency and cleaning of decision problems. Recall that a plan is intention-consistent if it doesn't contain intentions that exclude one another's achievement. This norm has a clear static consequence on plans viewed as sets of prizes, as we just saw. But it also shapes further deliberations by ruling out, in advance, options that would make the plan inconsistent if they were chosen.

Here I propose a way to model this effect of intentions on deliberation. The decision theoretical framework that I sketched above gives us some latitude on which options are going to be considered inconsistent. To keep the model as general as possible, I will just set that inconsistent lotteries are those that give too little chance of getting a prize in the intended set $\downarrow\mathfrak{P}_S$. In other words, the agent is going to make a preference-based decision among the lotteries that give at least a certain probability p to get a prize in $\downarrow\mathfrak{P}_S$. The value of p will stay undefined, making room for the whole spectrum of attitudes towards the prospect of getting an intended prize. For example, an agent for which $p = 1$ will restrict his choices to those who will guarantee a prize in $\downarrow\mathfrak{P}_S$. At the other extreme, an agent for which $p = 0$ will just decide as a regular utility maximizer, since no lotteries will be excluded from the original choice set. Needless to say that the size of the decision problem will decrease as p increases.

Precisely, we have that, given $0 \leq p \leq 1$, the *p-cleaned* version of a decision problem $\mathcal{DP}$ will contain the same prize (X), states (Ω) and events (Ξ) sets as before. What changes is the choice set, the lottery space. For each event S, define $L'_S = \{f : \forall t \in S, \sum_{x\in\downarrow\mathfrak{P}_S} f(x|t) \geq p\}$, that is, L'_S is the set of lotteries p-consistent with the plan adopted at S. Note that for all event S, L'_S is always non-empty because L is the set of *all* probability distribution over X. The preference relations over these cleaned lottery sets are just the restrictions of the original event-based preference orderings: $\preceq'_S = \preceq_S \upharpoonright L'_S$

With this in hand, the next step is to see whether we can still model agents that decide on p-cleaned decision problems as expected utility maximizers. But the very idea of cleaning is to *remove* lotteries from the choice set, which means that L'_s won't satisfy *continuity* anymore, except in the trivial case where $p = 0$. I won't provide a representation theorem that copes with this difficulty.

One could, of course, change the original preference relations by forming an "indifference cluster" with all lotteries that are cleaned out of the choice set. That is, define a new preference relation $\preceq''_S$ as follows: for all $f, g \in L'_S$, $f \preceq''_S g$ iff $f \preceq_S g$ and for all f, g outside L'_S, $f \sim''_S g$. Furthermore, for all f not in L'_S and g in L'_S, set $f \preceq''_S g$. This would have the immediate consequence of ruling out the lotteries outside L'_S as potential expected utility maximizers, while preserving the preference structure within L'_S. However, this "preference shifting" strategy seems to betray the very motivation of cleaning: reducing the "size" of the choice set L. Clearly, under the preference shift strategy the agent just uses the original choice set. So, I think that one should avoid such an easy way out if he genuinely aims at modeling the simplification function of plans.

As I said, I will not provide a way out of the failure of continuity for cleaned decision problems. What I'll do instead is simply to point out that certain special cases of preference orderings are representable. Say that a preference relation $\preceq'_S$ is *p-cleaned friendly* only if for all $f \in L$ and $t \in S$, if $\sum_{x\in\downarrow\mathfrak{P}_S} f(x|t) < p$ then there exists a lottery g in L such that $f \sim_S g$ and $\sum_{x\in\downarrow\mathfrak{P}_S} g(x|t) \geq p$. In such cases, we immediately get the following:

Theorem 2. *All p-cleaned friendly preference relations restricted to a p-cleaned decision problem $\mathcal{DP}$ are representable by a utility function u' and a probability distribution p', defined in that same way as in Section 2.*

Means-end coherence and the clustering of decision problems In the previous section, I've proposed a framework to reduce the size of a decision problem by "cleaning out" lotteries that give too low a prospect of getting an intended prize. The idea behind the cleaning procedure is that the chosen options shouldn't go against the achievement of the intentions the agent already has, which would violate intention-consistency. Now I turn to the norm of means-end coherence.

As we saw, this norm calls for adopting effective means to reach intended ends. Put into our framework, this means that an agent will have to choose one particular ways to achieve its plans, one *attainment*. But I also mentioned that, for resource-bounded agents, means-end coherence has a further effect. It presses for a careful management of the time devoted to deliberation, and so requires that the agent leaves irrelevant details out. Again, this can be modeled in the present decision-theoretical framework, by *clustering* lotteries that are the equivalent modulo certain attainments of the plan.

Let a pair $\langle \mathcal{A}, p\rangle$ be a set of *p-attainments* in S for a plan $\mathfrak{P}_S$ where $\mathcal{A}$ is a partition of $\downarrow\mathfrak{P}_S$ and p is such that $1/2 < p \leq 1$. The intuitive idea here is that our agent has to choose between certain mutually exclusive ways to achieve its most precise intention $\downarrow\mathfrak{P}_S$, each of them being one "cell" A_i in the partition $\mathcal{A}$. This partition can be more or less fine-grained, making room for various level of details. For now, I make no assumption regarding *how* this a particular partition of $\downarrow\mathfrak{P}$ came to be considered as *the* set of attainments for $\downarrow\mathfrak{P}_S$. I just take it as given, and see how an agent can use it to reduce the size of his decision problem. The parameter p can be seen as the bound under with the agent considers that a lottery f is giving too little chances to get a prize in one the attainment A_i. I require that p is strictly greater than $1/2$ to make sure that the pair $\langle \mathcal{A}, p\rangle$ will finitely partition L.

This is done as follows. Given a cell $A_i \in \mathcal{A}$ and a set of lotteries L, define the *cluster* C_i of L corresponding to A_i as $C_i = \{f : \forall t, \sum_{x\in A_i} f(x|t) \geq p\}$. A lottery cluster C_i puts together all the lotteries that give a price in A_i with at least probability p. Now, the set of clusters C_i will not completely partition L. We need two more clusters to complete the partition. First, the *p-spread cluster* C_p is defined as $\{f : \neg\exists A_i$ s.t. $f \in C_i$ but $\sum_{x\in\downarrow\mathfrak{P}_S} f(x|t) \geq p\}$. This cluster will contain the lotteries that, for each cell A_i taken individually, don't give high enough probability to be in the cluster C_i, but that nevertheless reach the required probability p over the *whole* set $\downarrow\mathfrak{P}_S$. Finally, the *out of attainment cluster* $C_\varnothing$ is defined as $\{f : \neg\exists A_i \ s.t. \ f \in C_i$ and $f \notin C_p\}$. That is, this cluster contains the lotteries that just don't give high enough probability to get an intended prize. Observe that this cluster will contain exactly the lotteries that would be p-cleaned, according to the procedure described in the previous section. Given all these clusters, we automatically obtain the following fact:

Fact 3 *For a decision problem $\mathcal{DP}$ and a plan $\downarrow\mathfrak{P}_S$, the set of cluster C_i corresponding to an attainment set $\langle \mathcal{A}, p\rangle$ together with the the p-spread cluster C_p and the out of attainment cluster $C_\varnothing$* finitely partition *L.*

Given this finite partition, we can construct the new *p-clustered version* decision problem $\mathcal{DP}$, relative to an attainment set $\langle \mathcal{A}, p\rangle$ is defined as follows. Take X, Ω and Ξ as before. Set, for all $S \in \Xi$, $L'_S = \{C_i\}_{A_i \in \mathcal{A}} \cup C_p \cup C_\varnothing$ and take $\preceq'_S$ as a total and transitive weak ordering on L'_S. In a nutshell, the options an agent faces now are sets of lotteries, within each he get what he considers high enough chances to attain his goals.

Here I impose no *a priori* restriction on the cluster ordering $\preceq'_S$. But one might want that it somehow reflects the underlying preference ordering $\preceq_S$. This requirement boils down to the following: The preference ordering over a clustered decision problem in an event S is said to be *preference-aligned* if for all $f, g \in L$, if $f \in C_i$ and $g \in C_j$ and $E_p(u(f)/S) \leq E_p(u(g)/S)$then $C_i \preceq'_S C_j$.

With this definition in hand, the challenge is to fine easily computable ways to order a clustered decision problem in a preference-aligned way. This will be postponed for further work. For now I will content myself with stating an easy fact, that relates clustering and cleaning. The proof of this fact is direct, given the observation made above that the outside of attainment cluster $C_\varnothing$ contains just the lotteries that would be p-cleaned.

Fact 4 *For all decision problem $\mathcal{DP}$ and $1/2 < p \leq 1$, the p-cleaned version of the p-clustered $\mathcal{DP}$ is the same as the p-clustered version obtained after a p-cleaning of $\mathcal{DP}$.*

4.3 "Intention-based" utility theory

We thus have two ways to model the effects of plans of action on deliberation, cleaning and clustering, each of which is related to one rational constraint on intention. Now I turn a somewhat orthogonal issue: modeling intentions as setting a "decision agenda" that is independent of the one set by the preferences.

There are many reasons why one might want such an independent ordering. A striking one is that intentions are arguably irreducible to desires and preferences, and thus that the "intention-based" and "preference-based" rationality can diverse. [Bratman, 1987] made a strong case in favor or distinguishing intentions from desires, mainly on the basis that the latter are *not* subject to consistency requirements. But [Sen, 2005] also argued that commitment-based rationality fall systematically out of preference-based decisions. Even if Sen's idea of commitment seems to differ in many respects from Bratman's, these two points of view call for incorporating a intention- or commitment-based ranking of options that is independent of preferences. This can be easily done.

Given a decision problem $\mathcal{DP}$ and a plan $\mathfrak{P}$, introduce a new ordering $\preceq'_S$ defined as follows: For all f, g in L and $S \in \Xi$,

$$f \preceq'_S g \text{ iff for all } t \in S, \sum_{x \in \downarrow\mathfrak{P}_S} f(x|t) \leq \sum_{x \in \downarrow\mathfrak{P}_S} g(x|t)$$

The equivalence and strict versions of $\preceq'_S$ are defined as usual, which means:

$$f \sim'_S g \text{ iff for all } t \in S \sum_{x \in \downarrow\!\mathfrak{P}_S} f(x|t) = \sum_{x \in \downarrow\!\mathfrak{P}_S} g(x|t)$$

and similarly for $\prec'_S$.

Notice the intention-based ordering is *derived* from the intentions of the agent, while the preference ordering is *primitive*. It turns out that this new ordering can be quite easily represented by a utility function, proviso a certain structure of the plan.

For a given decision problem, the set of event-relative plans $\{\mathfrak{P}_S\}_{s \in \Xi}$ is said to *reflect boolean operations on events* if for all $S, T \in \Xi$,

1. $\mathfrak{P}_{S \cap T} = \mathfrak{P}_S \cap \mathfrak{P}_T$
2. $\mathfrak{P}_{S \cup T} = \mathfrak{P}_S \cup \mathfrak{P}_T$
3. $\mathfrak{P}_{S - T} = \mathfrak{P}_S - \mathfrak{P}_T$

These conditions are needed to make sure that the plan-based preferences satisfy *Subjective Substitution*, a key property to get a representation in the AA framework. They, so to speak, constraint the inter-event behavior of the plan, and at the same time of the induced orderings $\preceq'_S$. Although they look like constraints imposed for purely technical reasons, they can be philosophically motivated. The first intuitively corresponds to the following maxim: "if you believe more, then your intentions shouldn't be less precise". The second condition has the converse reading: "if you believe less, then your intentions shouldn't be more precise". The third is, I think, the more objectionable. It says that your plan in case you believe that S occurs but T doesn't should be the same as the plan that result from removing what you planned in case of T from what you planned in case of S. This is an extremely strong condition that doubtfully applies to *all* agents, and it remains an open question to me whether there is a more plausible constraint that still allows a utility representation.

Let me now introduce formally the intention-based utility notions that will be used below. Given a decision problem $\mathcal{DP}$, a *plan-based utility function* is a function $u^{\mathfrak{P}} : X \times \Omega \to \mathbb{R}$. Assume a conditional probability p as defined in Section (2), the *expected intention-based utility value* $E^{\mathfrak{P}}_p(u(f)|S)$ of a lottery f is calculated as follows

$$E^{\mathfrak{P}}_p(u^{\mathfrak{P}}(f)|S) = \sum_{t \in S} p(t|S) \sum_{x \in X} u^{\mathfrak{P}}(x,t) f(x|t)$$

Theorem 5. (Intention-utility representation theorem) *For any decision problem $\mathcal{DP}$, plans $\{\mathfrak{P}_S\}_{s \in \Xi}$ that reflects boolean operations on events and preference relations $\preceq'_S$ based on the latter, there exists a plan-based utility function $u^{\mathfrak{P}}$ and a conditional probability function p such that :*

1. *For all $t \in \Omega$, $\max_{x \in X} u^{\mathfrak{P}}(x,t) = 1$ and $\min_{x \in X} u^{\mathfrak{P}}(x,t) = 0$.*
2. *For all R, S, T such that $R \subseteq S \subseteq T \subseteq \Omega$ and $S \neq \varnothing$, $p(R/T) = p(R/S)p(S/T)$.*
3. *For all $f, g \in L_{\downarrow\!\mathfrak{P}_s}$ and $S \in \Xi$, $g \preceq'_S f$ iff $E^{\mathfrak{P}}_p(u^{\mathfrak{P}}(g)/S) \leq E^{\mathfrak{P}}_p(u^{\mathfrak{P}}(f)/S)$.*

So we can model "plan-based" decisions as maximization of a utility scale independent of the preferences. At this point, I should mention that it is not at all consensual whether intentions and commitment should be thus integrated in decision theory (see e.g. [Pettit, 2005]). As I said before, I will not attempt to defend the present approach against such arguments. In fact, I have argued in [Roy, 2005] that such an independent scale is unnecessary as long as decision and game theory deal with *ideal* agents.

Along the same lines, note that the beliefs represented by the probability distribution p are now independent form the "intention-based" preferences. So, in this framework, intentions don't have to influence the beliefs. The only relation between these two mental states is imposed by reflection of boolean operations on events, and rather goes from beliefs to intentions. This leaves open whether having the intention to get some prizes in X implies a higher degree of belief that one will actually get some prizes in X.

5 Conclusion

In this paper, I proposed a framework for intention-committed agency built upon a decision theoretical model. I have shown how to represent an intention-based decision making which, in short, models an agent who tries to get prizes in a specific subset of the set of all possible prizes. I have also proposed a way to represent two key aspects of planning agency for resource-bounded agents: simplification by exclusion of inconsistent options and simplification by ignoring irrelevant details.

Many open questions remain for further work, among which the most urgent are:

- Compare intention- and preference-based decisions and examine their combination in interactive situations (games).
- Find a general result concerning the existence of a utility representation for cleaned decision problem.
- Explore simple and efficient algorithms to align preferences over clustered decision problems on intentions and/or original preferences.

Finally, I should mention that some authors have expressed strong skepticism regarding the very idea of introducing plans of action as an independent feature in decision theory (see for example [Strotz, 1956]). This paper can be seen as a modest attempt to rise to the challenge.

A Appendix

Proof of Theorem 2. Just use the same procedure as in [Myerson, 1991], except that everywhere a lottery outside of L'_S would have to be used to define u' or p', use the one in L'_S to which the agent is indifferent.

Proof of Fact 3. It is enough to show that $\approxeq$ is well defined and an equivalence relation, for it is clear that, by the finiteness of X, the partition corresponding to $\approxeq$ is then finite.

1. Assume $f \in C_i$ and $g \in C_j$ for $i \neq j$. We have to show that $f \neq g$. The case where one of the cluster is $C_\varnothing$ or C_p is trivial. Now $f \in C_i$ means that for all t $\sum_{x \in A_i} f(x|t) \geq p$, and similarly from g. But since $\langle \mathcal{A}, p \rangle$ partitions X, and $1/2 < p \leq 1$ we know that $\sum_{x \in A_j} f(x|t) < p$, and the same for g and A_i, which is enough to show that $f \neq g$.
2. It is easy to see that $\approxeq$ is reflexive and symmetric. Transitivity follows from the fact that $\approxeq$ is well-defined.

Proof of Theorem 5. I first establish that $\preceq'_S$ satisfies Subjective Substitution.

Lemma 1. *If $\{\mathfrak{P}_S\}_{s \in \Xi}$ reflects boolean operations on events then $\preceq'_S$ satisfies subjective substitution.*

Proof. Assume that $\{\mathfrak{P}_S\}_{s \in \Xi}$ reflects boolean operations on events. Clearly, for all events S, $\preceq'_S$ is transitive and complete. Take two lotteries g and f such that $g \preceq'_S f$ and $g \preceq'_T f$, i.e.

$$\sum_{x \in \downarrow\mathfrak{P}_S} g(x|t) \leq \sum_{x \in \downarrow\mathfrak{P}_S} f(x|t)$$

and

$$\sum_{x \in \downarrow\mathfrak{P}_T} g(x|t') \leq \sum_{x \in \downarrow\mathfrak{P}_T} f(x|t')$$

for all t and t' in S and T, respectively. Assume also that $S \cap T = \varnothing$. We now have to show that

$$\sum_{x \in \downarrow\mathfrak{P}_{S \cup T}} g(x|t) \leq \sum_{x \in \downarrow\mathfrak{P}_{S \cup T}} f(x|t)$$

for all $t \in S \cup T$. Now we know by assumption that $\downarrow\mathfrak{P}_{S \cup T} = \downarrow\mathfrak{P}_S \cup \downarrow\mathfrak{P}_T$. In other words, we have,

$$\sum_{x \in \downarrow\mathfrak{P}_{S \cup T}} g(x|t) = \sum_{x \in \downarrow\mathfrak{P}_S} g(x|t) + \sum_{x \in \downarrow\mathfrak{P}_{T-S}} g(x|t)$$

and similarly for $f(x|t)$. Now assume that this is not the case that $g \preceq'_{S \cup T} f$. That is, by completeness,

$$\sum_{x \in \downarrow\mathfrak{P}_S} f(x|t) + \sum_{x \in \downarrow\mathfrak{P}_{T-S}} f(x|t) < \sum_{x \in \downarrow\mathfrak{P}_S} g(x|t) + \sum_{x \in \downarrow\mathfrak{P}_{T-S}} g(x|t)$$

This is equal to

$$\sum_{x \in \downarrow\mathfrak{P}_S} f(x|t) - \sum_{x \in \downarrow\mathfrak{P}_S} g(x|t) < \sum_{x \in \downarrow\mathfrak{P}_{T-S}} g(x|t) - \sum_{x \in \downarrow\mathfrak{P}_{T-S}} f(x|t)$$

Now $g \preceq'_T f$ also decompose into:

$$\sum_{x \in \downarrow\mathfrak{P}_{T \cap S}} g(x|t) + \sum_{x \in \downarrow\mathfrak{P}_{T-S}} g(x|t) \leq \sum_{x \in \downarrow\mathfrak{P}_{T \cap S}} f(x|t) + \sum_{x \in \downarrow\mathfrak{P}_{T-S}} f(x|t)$$

which is the same as

$$\sum_{x \in \downarrow\mathfrak{P}_{T-S}} g(x|t) - \sum_{x \in \downarrow\mathfrak{P}_{T-S}} f(x|t) \leq \sum_{x \in \downarrow\mathfrak{P}_{T \cap S}} f(x|t) - \sum_{x \in \downarrow\mathfrak{P}_{T \cap S}} g(x|t)$$

But then, by transitivity, we have

$$\sum_{x \in \downarrow\mathfrak{P}_S} f(x|t) - \sum_{x \in \downarrow\mathfrak{P}_S} g(x|t) < \sum_{x \in \downarrow\mathfrak{P}_{T \cap S}} f(x|t) - \sum_{x \in \downarrow\mathfrak{P}_{T \cap S}} g(x|t)$$

which is impossible, given that $T \cap S \subseteq S$ and that the right side of the last inequality must be greater than 0.

Now I turn to the proof of the main theorem. Let $a_1(\bullet|t)$ and $a_0(\bullet|t)$ be the lotteries defined exactly as in [Myerson, 1991, chap.1]. They respectively give for sure the best and worst prizes at t, according to the preference ordering. These will be used to build the probability function p. Define a "bet on t", which will be used to construct the probability function p, as follows

$$b_S(x|t) = \begin{cases} a_1 \text{ if } t \in S \\ a_0 \text{ Otherwise} \end{cases}$$

Now construct the conditional probability function such that for every $t \in \Omega$, $p(t|S)$ satisfies

$$b_{\{t\}} \sim_S p(t|S)a_1 + (1 - p(t|S))a_0$$

This is essentially Savages' method to extract conditional beliefs from preferences. It is a standard argument to show that p as the required properties. Details can be found in [Myerson, 1991, chap.1].

As for the plan-based utility function, just define $u^{\mathfrak{P}}(x,t)$, for every $x \in X$ and $t \in \Omega$, as follows:

$$u^{\mathfrak{P}}(x,t) = \begin{cases} 1 & \text{if } x \in \downarrow\mathfrak{P}_S \\ 0 & \text{Otherwise} \end{cases}$$

It should now be clear that, for all t,

$$\sum_{x \in X} u^{\mathfrak{P}}(x,t) f(x|t) = \sum_{x \in \downarrow\mathfrak{P}_{\{t\}}} f(x|t)$$

But then, because $\preceq'_S$ satisfies *Subjective Substitution*, this extends to all S. That is,

$$\sum_{x \in X} u^{\mathfrak{P}}(x,t) f(x|t) = \sum_{x \in \downarrow\mathfrak{P}_S} f(x|t)$$

And from that we directly get

$$f \preceq'_S g \text{ iff } \sum_{t \in S} p(t|S) \sum_{x \in X} u^{\mathfrak{P}}(x,t) f(x|t) \leq \sum_{t \in S} p(t|S) \sum_{x \in X} u^{\mathfrak{P}}(x,t) g(x|t)$$

Which completes the proof.

Remark 1. Getting a representation of $\preceq'_S$ in terms of expected utility is much easier than for the normal preferences because I have defined this ordering using the quantitative information already contained in the lotteries. This is way I could bypass the usual "extraction" of the utility function from preferences, and directly define $u^{\mathfrak{P}}$. This reveals that, in the end, $u^{\mathfrak{P}}$ is no more than a "coarse-grained" utility function.

References

[Anscombe and Aumann, 1963] Anscombe, F. J. and Aumann, R. (1963). A definition of subjective probability. *Annals Math. Stat.*, 34:199–205.

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

[Bratman et al., 1991] Bratman, M. E., Israel, D., and Pollack, M. E. (1991). Plans and resource-bounded practical reasoning. In Pollock, J. and Cummins, R., editors, *Philosophy and AI: Essays at the Interface*, pages 7–22. MIT Press.

[Luce and Raiffa, 1957] Luce, D. R. and Raiffa, H. (1957). *Games and Decisions; Introduction and Critical Survey*. Dover Publications, Inc.

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

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

[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.

[Pettit, 2005] Pettit, P. (2005). Construing sen on commitment. *Economics and Philosophy*, 21(01):15–32.

[Roy, 2005] Roy, O. (2005). What does game theory have to do with plans? Technical report, ILLC, Prepublication Series, PP-2005-13.

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

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

[Strotz, 1956] Strotz, R. H. (1955 - 1956). Myopia and inconsistency in dynamic utility maximization. *The Review of Economic Studies*, 23(3):165–180.

[van Hees, 2003] van Hees, M. (2003). Intentions, utility and rationality. http://www.philos.rug.nl/˜vanhees/.

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

[Wallace, 2003] Wallace, R. J. (2003). Normativity, commitment, and instrumental reason. *Philosophers' Imprint*, 1(3):1–26.