Towards qualitative approaches to multi-stage decision making

Hélène Fargier, Jérôme Lang and Régis Sabbadin
IRIT - Université Paul Sabatier - 31062 Toulouse Cedex (France)
e-mail: {fargier, lang, sabbadin} @irit.fr

1 Introduction

For a few years, there has been a growing interest in the Artificial Intelligence community towards the foundations and computational methods of decision making under uncertainty. This is especially relevant for applications to planning, where a suitable sequence of decisions is to be found, starting from a description of the initial world, of the available decisions and their effects, and of the goals to reach. Several authors have thus proposed to integrate some parts of decision theory into the planning paradigm; but up to now, they have focussed on "classical" models for decision making, based on *Markov decision processes* (where actions are stochastic and the satisfaction of agents expressed by a numerical, additive utility function), and its computational counterpart, *dynamic programming*. However, transition probabilities for representing the effects of actions are not always available, especially in AI applications where uncertainty is often ordinal, qualitative. The same remark applies to utilities: it is often more adequate to represent preference over states simply with an ordering relation rather than with additive utilities. Recently, several authors have advocated this qualitative view of decision making and have proposed qualitative versions of decision theory, together with suitable logical languages for expressing preferences, namely, Boutilier [5], Tan and Pearl [12], Dubois and Prade [7] [8]. The latter propose a qualitative utility theory based on possibility theory, where preferences and uncertainty are both qualitative. Our purpose is to extend Dubois and Prade's possibilistic framework for qualitative decision theory so as to enable multiple-stage decision making.

In order to have a synthetic view on problems and approaches pertaining to decision making under uncertainty, we may consider the following taxonomy, which gives various classes of problems (from elementary to more complex ones) when the different criteria vary. To the very general class of problems we consider, we give the generic name of "generalized Markov decision processes" (GMDP for short), since we always make the Markovian assumption that the

past of the system cannot influence the choice of the policy at a given time point.

- *Temporal structure of the decision stages.* There may be only one decision stage, or there may be an ordered set of time points (stages) where decisions are to be made; this set may be either finite (*finite horizon decision making*) or discrete infinite (*infinite horizon decision making*).

- *Available knowledge of the initial state.* This knowledge may be precise (thus described by one possible initial state), probabilistic (probability distribution on the set of possible states), or possibilistic (possibility distribution on the set of possible states). At this point we recall that a possibility distribution on a set of states S is a mapping $\pi : S \rightarrow [0, 1]$, where $\pi(s)$ measures to what extent s is likely to be the actual state, ranging from 1 (completely possible) to 0 (impossible). A possibility distribution is generally assumed to be normalized (and we will make this assumption throughout the paper), *i.e.* $\exists s$ such that $\pi(s) = 1$. When π takes its value in $\{0, 1\}$, the possibility distribution is said to be *crisp* and is equivalent to a classical set. Thus, a description of the initial state by a crisp possibility distribution comes down to specifying a set of possible initial states.

- *Knowledge on the effects of actions.* Actions may be *deterministic*, meaning that for a given state s and an action a allowable in s, there is only one possible subsequent state $Result(s, a)$; they are *nondeterministic* iff for each state and each action there is a *set* of possible subsequent states; they are *stochastic* (resp. *possibilistic*) when their effects are described by a probability (resp. possibility) distribution; from the possibilistic case we recover the nondeterministic case by allowing only crisp possibility distributions.

- *Description of the goals.* In the case of finite-horizon decision making, the final state reached is often of primary importance for the global satisfaction of the agent (it may even be the only criterion). The agent may have to achieve a crisp, non-flexible goal, *i.e.*, to reach one of the *goal states*; the notion of goal is sometimes defined in a more flexible way, by a utility function or more generally a function on an arbitrary ordered satisfaction scale.

- *Role of intermediate states and actions in the global satisfaction of the agent.* Intermediate states may also receive a utility degree, or more generally a satisfaction degree, which is taken into account when computing the global satisfaction attached to a path (a succession of states). The same remark applies to actions which may have a cost which also has to be taken into account. The satisfaction degrees and costs of the different

states and actions may be aggregated additively as in classical utility theory, or qualitatively by the minimum, or by other operators. The utilities of the different states reached may be weighted by a discounting factor, especially in the case of infinite horizon decision making [10].

- *Choice criterion for the policy.* An optimal policy consists in attaching to each reachable state the *best* action, following a criterion which has to be defined. In classical decision theory, this criterion consists in maximizing the expected utility. As to qualitative approaches, a possible criterion (which is used by Dubois and Prade [7] [8]) consists in making an assumption of commensurability between the uncertainty and the satisfaction scales and then maximizing a "pessimistic" qualitative utility (see Section 3).

- *Observability.* A GMDP is *fully observable* if the state of the world is known at each step of the process, *partially observable* if the agent may only have an incomplete knowlegde of the state of the world. In the latter case, some tests (or knowledge-gathering actions) may be available to the agent, who uses their results to maintain his/her beliefs about the current state (represented by a set of states, or a probability distribution, or a possibility distribution...).

We name the different models we are going to discuss according to four of the abovementioned criteria, namely, the temporal structure of the decision stages (respectively 1, N and ∞ for one / a finite number / an infinite discrete number of decision stages), the uncertainty model for the results of actions (D for deterministic, ND for nondeterministic, pr for stochastic, π for possibilistic), the description of the goals (G for crispness, $+$ for additive utilities, min for qualitative utilities combined by min), and the choice criterion (for instance $\bar{u}$ for maximum expected utility; others criteria are considered later on). Full observability is assumed everywhere (and therefore omitted). With these notations, the "standard" approach to multi-stage decision under uncertainty, namely, fully observable Markov decision processes, correspond to the 4-uple $\langle N, pr, +, \bar{u} \rangle$.

On the other hand, Dubois and Prade's possibilistic decision theory [7] [8] provides a framework for one-stage decision making with possibilistic uncertainty on the results of actions, a precise initial state s_0, a qualitative satisfaction function on the same scale as possibility distributions. Observability considerations are not relevant since the initial state is precisely known and there is only one decision stage. Finally, the choice of the best action is based on a commensurability assumption between satisfaction and uncertainty, that we are going to describe more formally. Let S and A the set of possible states and available actions, respectively. The possible subsequent states of action a from s_0 is described by the possibility distribution $\pi(.|a, s_0)$. The qualitative utility is a mapping u from S to a completely ordered lattice L (typically, $[0, 1]$) where

$u(s) = 1_L$ and $u(s) = 0_L$ respectively mean complete satisfaction and dissatisfaction. Then the qualitative value of a is measured by a qualitative utility function (which plays the same role as expected utility in standard decision theory), defined by

$$u^{PES}(s_0, a) = min_{s \in S} max(n(\pi(s|s_0, a)), u(s))$$

where n is an order reversing function from L to L satisfying $n(0) = 1$ and $n(1) = 0$ (when $L = [0, 1]$ the prototypical order reversing function is $n(x) = 1 - x$); the best action in s_0 is then the action a maximizing $u^{PES}(s_0, a)$. The quantity $u^{PES}(s_0, a)$ is a pessimistic criterion [8]; indeed, it is the necessity of a fuzzy event, namely, it can be viewed as a degree of inclusion of the fuzzy set of more or less possible situations in the fuzzy set of preferred outcomes [8]. Using our notations, Dubois and Prade's framework corresponds to the 4-uple $\langle 1, \pi, min, u^{PES} \rangle$ where u^{PES} is the pessimistic utility with a commensurability assumption. Note that a similar pessimistic criterion has been proposed by Whalen [13], in terms of "disutility". Lastly, in the case where $\pi(.|s_0, a)$ is a crisp possibility distribution, $u^{PES}(s_0, a)$ is the utility of the worst possible outcome when performing a (this criterion is known as the Wald index).

In this paper we aim to generalize Dubois and Prade's framework to multistage decision, first (Section 2) assuming full observability, possibilistic uncertainty and crisp goals, and then (Section 3) assuming full observability, possibilistic uncertainty and qualitative utility with the abovementioned commensurability assumption. We will relate our work with some other approaches to qualitative multi-stage decision making, and we will briefly give some hints on how to further generalize our framework in several ways.

2 Multi-stage decision making with possibilistic uncertainty and crisp goals

2.1 Notations

Following [10], let T denote the finite set of time points (or stages) at which decisions are to be made : $T = \{1, 2, ..., N\}$

The set of possible states at stage $t \in T$ is denoted by S_t. This is defined for $t = 1, 2, ..., N + 1$ although decisions are made only at stages $t = 1, 2, ..., N$. This is because the decision at stage N will lead to a final state at stage $N + 1$ on which the goals are expressed.

Under our assumption of full observability, at stage t, the decision maker will observe the system in state $s \in S_t$ and choose an action a from the set of allowable actions at t in state s, denoted by $A_{s,t}$.

Since the effects of the actions are ill-known, our knowledge about the subsequent state is described by a transition possibility function $\pi_t : \pi_t(s'|(s, a))$

is the possibility that the state at stage $t+1$ will be $s' \in S_{t+1}$ knowing that action $a \in A_{s,t}$ has been chosen in state s at stage t.

The question is then to define a policy, i.e. a decision rule d_t for each stage t : d_t is a function which associates to each state s in S_t an action $d_t(s) \in A_{s,t}$. A decision rule of this type is called Markovian since it depends only on the state of the system.

The set of allowable decision rules at stage t is denoted by D_t and called the decision set at t. It is the set of all the mappings from S_t to $A_{s,t}$.

A partial policy $d_{t \to N}$ specifies the sequence of decision rules to be used by the decision maker from the stage $t \in T$ over the course of the planning horizon. That is, $d_{t \to N} = \{d_t, d_{t+1}, ..., d_N\}$ where $t \in T$ and $d_i \in D_i$. The set of all the allowable partial policies from t to N will be denoted $D_{t \to N}$

A complete policy specifies the sequence of decision rules to be used by the decision maker from the initial state . That is, any $d_{1 \to N}$ is a complete policy.

2.2 Problem formulation

In this first model, we assume *non-flexible goals*, posted only for stage $N+1$. The description of the goals consists then in partitioning S_{N+1} into a set of *goal states* $Goal \subseteq S_{N+1}$ and a set of *non-goal states* $(S_{N+1} \setminus Goal)$. Hence, the prpblem is completly specified by the set of stages $T = 1, ..., N+1$, the set $S_t, t = 1, ..., N+1$ of possible states at each stage, the sets $A_{s,t}$ of allowed actions for each s in each S_t, the transition possibility functions $\pi(.|(s,a))$ and the set $Goal \subseteq S_{N+1}$ of goal states at stage $N+1$.

We are now going to define the notion of optimal policy, and then we will give a generic algorithm (in the style of dynamic programming) to compute it. This notion of optimal policy is based on the computation, for each t, each $s \in S_t$, and each $a \in A_{s,t}$, of the two quantities $\Pi(Good_t(s,a))$ and $\mathsf{N}(Good_t(s,a))$ which are respectively the possibility and the necessity that performing a in state s will eventually lead to a goal state *provided that the optimal strategy is followed.*

Let us first consider a state s at stage N, and let us suppose now that the action $a \in A_{s,N}$ is performed in s. Since the subsequent state is described by a possibilistic transition function, it is possible to compute the possibility and the necessity of the event "the subsequent state will be a goal state", denoted by $Good_N(s,a)$:

Definition 1
$\Pi(Good_N(s,a)) = \max_{s' \in Goal} \pi_N(s'|(s,a))$
$\mathsf{N}(Good_N(s,a)) = \min_{s' \notin Goal}(1 - \pi_N(s'|(s,a)))$

Proposition 1
$\Pi(Good_N(s,a))$ and $\mathsf{N}(Good_N(s,a))$ are standard possibility and necessity degrees, and thus, verify:
$\mathsf{N}(Good_N(s,a)) > 0 \Rightarrow \Pi(Good_N(s,a)) = 1$ (1)

Namely, $\Pi(Good_N(s,a))$ and $\mathsf{N}(Good_N(s,a))$ are respectively the possibility and the necessity of the event "the subsequent state of s when action a is performed is in $Goals$". A cautious, pessimistic approach consists in preferring an action which maximizes the necessity to reach a goal state. While this criterion seems rather natural, it is however often not discriminating enough since there may often be no action at all leading to a goal state with some strictly positive certainty. The idea is then to discriminate further the actions using $\Pi(Good_N(s,a))$, which is a more optimistic index. Formally, the following criterion will be used, where $a \geq_s a'$ reads "a is at least as good as a' in state s " :

Definition 2
$a \geq_s a'$ iff ($\mathsf{N}(Good_N(s,a)) > 0$ and $\mathsf{N}(Good_N(s,a)) \geq \mathsf{N}(Good_N(s,a'))$
or $\mathsf{N}(Good_N(s,a)) = \mathsf{N}(Good_N(s,a'))$ and $Pi(Good_N(s,a)) \geq \Pi(Good_N(s,a'))$

Since $\Pi(Good_N(s,a))$ and $\mathsf{N}(Good_N(s,a))$ are possibility and necessity degrees, we have:

Proposition 2

$a \geq_s a'$ iff $\mathsf{N}(Good_N(s,a)) \geq \mathsf{N}(Good_N(s,a'))$ and $\Pi(Good_N(s,a)) \geq \Pi(Good_N(s,a'))$

Obviously, $\geq_s$ is a complete preorder. We define in a classical way the strict order $>_s$ induced by $\geq_s$ by $a >_s a'$ iff it is not the case that $a' \geq_s a$. From Proposition 1 it can be checked easily that

$a >_s a'$ iff $\mathsf{N}(Good_N(s,a)) > \mathsf{N}(Good_N(s,a'))$ or $\Pi(Good_N(s,a)) > \Pi(Good_N(s,a'))$

An action a will then be *optimal for* s iff there is no action a' such that $a' >_s a$. The set of optimal actions for s (at stage N) will be denoted by $A^*_{s,N}$. Due to (1), optimal actions can be characterized more intuitively by the following property: in case there is an action a leading from s to a goal state with some certainty, *i.e.*, $\mathsf{N}(Good_N(s,a)) > 0$, then $A^*_{s,t}$ is the set of actions for which $\mathsf{N}(Good_N(s,a))$ is maximal; otherwise, if it is the case that $\mathsf{N}(Good_N(s,a)) = 0$ for each action a, then $A^*_{s,t}$ is the set of actions for which $\Pi(Good_N(s,a))$ is maximal

Now, for a given state s at stage N, an optimal policy will assign an arbitrary action in $A^*_{s,N}$. Then, we can define the possibility and the necessity that a goal state can be reached from s, being the corresponding possibility and necessity degrees obtained for an optimal action for s.

Definition 3 *Let a^* be an arbitrary action in $A^*_{s,N}$; then*
$\Pi(Good_N(s)) = \Pi(Good_N(s,a^*))$
$\mathsf{N}(Good_N(s)) = \mathsf{N}(Good_N(s,a^*))$

This definition is well-formulated since $\Pi(Good_N(s,a^*))$ and $\mathsf{N}(Good_N(s,a^*))$ are constant for all actions a^* in $A^*_{s,N}$. Due to (1), it can be checked easily that the definition is equivalent to $\Pi(Good_N(s)) = \max_{a \in A_{s,N}} \Pi(Good_N(s,a))$ and $\mathsf{N}(Good_N(s)) = \max_{a \in A_{s,N}} \mathsf{N}(Good_N(s,a))$. In other terms, $\Pi(Good_N(s))$ (resp. $\mathsf{N}(Good_N(s))$) is the possibility (resp. the necessity) of the event "there is a policy which leads from s to a goal state", or equivalently, "performing an optimal action in s leads to a goal state". It follows that

Proposition 3 *$\Pi(Good_N(s))$ and $\mathsf{N}(Good_N(s))$ are standard possibility and necessity degrees*

This is straightforward since, for any s, $\Pi(Good_N(s))$ and $\mathsf{N}(Good_N(s))$ are equal to $\Pi(Good_N(s,a))$ and $\mathsf{N}(Good_N(s,a))$ for a given a (namely, an optimal action for s). By backwards induction, we are now going to define, for every stage t, for every $s \in S_t$ and every $a \in A_{s,t}$: (i) the possibility and the necessity degrees that performing a in s, followed by an optimal policy for stages $t+1$ to N, will eventually lead to a goal state; this event will be denoted by $Good_t(s,a)$; and (ii) the possibility and necessity degrees that a goal state can be reached from s following an optimal policy for stages t to N. This event will be denoted $Good_t(s)$. More formally:

Definition 4
$\Pi(Good_t(s,a)) = \max_{s' \in S_{t+1}} \min(\pi_t(s'|(s,a)), \Pi(Good_{t+1}(s')))$
$\mathsf{N}(Good_t(s,a)) = \min_{s' \in S_{t+1}} \max(1 - \pi_t(s'|(s,a)), \mathsf{N}(Good_{t+1}(s')))$

Hence we can compare actions with respect to a given state $s \in S_t$ in the same way as done for states in S_N. Namely,

Definition 5
$a \geq_s a'$ iff $[\mathsf{N}(Good_N(s,a)) \geq \mathsf{N}(Good_N(s,a'))$ and $(\mathsf{N}(Good_N(s,a)) > 0]$
or $[\mathsf{N}(Good_N(s,a)) = \mathsf{N}(Good_N(s,a'))$ and $\Pi(Good_N(s,a)) \geq \Pi(Good_N(s,a'))]$

An action a is optimal for s iff there is no action a' such that $a' >_s a$. The set of optimal actions for s (at stage t) will be denoted by $A^*_{s,t}$. *An optimal policy assigns to each stage s an optimal action .*

Definition 6 *Let a^* be an arbitrary action in $A^*_{s,N}$. Then,*
$\Pi(Good_N(s)) = \Pi(Good_N(s,a^*))$
$\mathsf{N}(Good_N(s)) = \mathsf{N}(Good_N(s,a^*))$

$\Pi(Good_t(s))$ and $\mathsf{N}(Good_t(s))$, are respectively measuring to what extent *applying an optimal policy* from s will lead possibly (resp. certainly) to a goal state at stage $N+1$. They are actually the possibility and necessity measures of the event "there is a policy which leads from s to a goal state". This is the meaning of the following result which can be proved by induction, given that the probem is Markovian (where $D_{t \rightarrow N}$ stands for the set of all policies for stages $t, t+1, ..., N$).

Proposition 4
$\Pi(Good_t(s)) =$
$\max_{d \in D_{t \to N}} \max_{\substack{(s_{t+1},\ldots,s_{N+1}) \in S_{t+1} \times \ldots \times S_{N+1} \\ s_{N+1} \in Goal}} \min_{j=t\ldots N} \pi(s_{j+1}|(s_j, d(s_j)))$

$\mathsf{N}(Good_t(s)) =$
$\max_{d \in D_{t \to N}} \min_{\substack{(s_{t+1},\ldots,s_{N+1}) \in S_{t+1} \times \ldots \times S_{N+1} \\ s_{N+1} \notin Goal}} \max_{j=t\ldots N}(1 - \pi(s_{j+1}|(s_j, d(s_j))))$

In other terms, $\Pi(Good_t(s)) = \max_{a \in A_{s,t}} \Pi(Good_t(s,a))$ and $N(Good_t(s)) = \max_{a \in A_{s,t}} \mathsf{N}(Good_N(s,a))$. Hence, policies maximizing the necessity and the possibility to eventually lead to a goal state may be computed backwards, assigning to each state s in stage t, an action in $A^*_{s,t}$.

2.3 Computing an optimal policy

An optimal policy can be computed by a possibilistic variant of dynamic programming [2]; it computes the policy backwards (from later stages to earlier ones). The correctness of the algorithm comes straightforwardly from the recursive definition of an optimal policy.

Algorithm 1:

$d = \emptyset$; {the policy}
for $s \in S_{N+1}$ **loop** { initialisation}
  if $s \in Goal$ **then** {good final state}
    $\mathsf{N}(Good_N(s)) = 1$; $\Pi(Good_N(s)) = 1$;
  else {bad final state}
    $\mathsf{N}(Good_N(s)) = 0$; $\Pi(Good_N(s)) = 0$;
  endif;
end for;
for $t := N$ downto 1 **loop1** {backward computing}
  for $s \in S_t$ **loop2**
    for $a \in A_{s,t}$ **loop3**
      compute $\mathsf{N}(Good_t(s,a)) = \min_{s' \in S_{t+1}} \max(1 - \pi_t(s'|(s,a)), \mathsf{N}(Good_{t+1}(s')))$
      compute $\Pi(Good_t(s,a)) = sup_{s' \in S_{t+1}} min(\pi_t(s'|(s,a)), \Pi(Good_{t+1}(s')))$
    { choose the best action }
    end loop3;
    $d_t(s) = a^*$ { a^* maximizes $[\mathsf{N}(Good_t(s,a)), \Pi(Good_t(s,a))]$}
  end loop2;
  add the decision rule d_t to the policy d.
end loop1;

Interestingly, the use of qualitative operators (min and max) – instead of + and product for classical MDP – gives us more opportunities to avoid

unnecessary computations; thus, a given action may be detected to be sub-optimal w.r.t. a given state before the whole computation of $\Pi(Good_t(s,a))$ and $\mathsf{N}(Good_t(s,a))$ is completed (*). This leads to the following improved version of the algorithm:

Algorithm 1':

$d = \emptyset$; {the policy}
for $s \in S_{N+1}$ **loop** {initialisation}
  if $s \in Goal$ **then** {good final state}
    $\mathsf{N}(Good_N(s)) = 1$;
    $\Pi(Good_N(s)) = 1$;
  else {bad final state}
    $\mathsf{N}(Good_N(s)) = 0$;
    $\Pi(Good_N(s)) = 0$;
  endif;
end for;
for $t := N$ downto 1 **loop1** {backward computing}
  for $s \in S_t$ **loop2**
    $n_{opt} = 0$; $pi_{opt} = 0$;
    for $a \in A_{s,t}$ **loop3** {compute N and Π }
      $n = 1$; $pi = 0$;
      for $s' \in S_{t+1}$ **loop4**
        if $pi_{opt} < 1$ **then**
          $pi = \max(pi, \min(\pi_t(s'|(s,a)), \Pi(Good_{t+1}(s'))))$;
          if $pi > pi_{opt}$ **then**
            $pi_{opt} = pi$; $a_{opt} = a$;
          { else, pi_{opt} cannot be improved, and $\Pi(Good_t(s,a))$ is useless}
          end if;
        end if;
        $n = \min(n, \max(1 - \pi_t(s'|(s,a)), \mathsf{N}(Good_{t+1}(s'))))$;
        if $n < n_{opt}$ **then** EXIT loop4 (*) **endif** ;
        {in this case a is sub-optimal, since n can only decrease}
      end loop4;
      if $n > n_{opt}$ **then**
        $n_{opt} = n$; $a_{opt} = a$;
      end if;
    end loop3;
    $\Pi(Good_t(s)) = pi_{opt}$; $\mathsf{N}(Good_t(s)) = n_{opt}$; $d(s) = a_{opt}$;
  end loop2;
end loop1;

3 Taking flexible goals into account

Up to now, we have taken into account only two types of final states : those which satisfy the goals, and those which do not. This dichotomy is inadequate to model many real decision problems, where the actor only expresses preference or indifference between states of the world: goals are flexible.

Let us now assume flexible goals posted only for stage $N+1$ (no intermediate rewards for actions or non-final states).This order over the final states in S_{N+1} is expressed by assigning to each of them a "qualitative utility index" u taken from an ordinal scale : "a state of the world s will be prefered to a state s', if and only if $u(s) > u(s')$".This utility index $u(s)$ can be seen as the degree to what extent the final state s is satisfying, according to the actor's point of view.

In other terms, taken flexibility into account leads to have a *fuzzy set* of goal states, instead of a crisp one as in the previous section : $\mu_{Goal}(s) = u(s)$ for $s \in S_{N+1}$

3.1 Problem formulation

A possibilistic flexible Markov decision problem will be specified by the set of stages $T = 1, ..., N$, the sets $S_t, t = 1, ..., N+1$ of possible states at each stage point, the sets $A_{s,t}$ of allowed actions for each s in each S_t, the transition possibility functions $\pi(.|(s,a))$ and the utility function u over the final states. We shall now see how this utility index on final states can be used to allow to compare decisions. As previously, we can base this comparison on two different indices, depending on whether we take an optimistic or a pessimistic point of view.

3.2 Pessimistic utility

Consider that in stage N, we observe state $s \in S_N$ and we choose action $a \in A_{s,N}$. As before, we know the possibility distribution over S_{N+1} : $\pi_N(s'|(s,a))$. But for each state $s' \in S_{N+1}$ we also know its utility $u(s')$.

According to a *pessimistic* point of view, we can say that action a in state s is all the better as the *worst possible outcome* s' has a *high utility*. This leads us to an assumption of commensurability : we assume that $1 - \pi_N(s'|(s,a))$, and $u(s')$ are expressed on the same scale. As in [8] we define then the pessimistic utility index by:

Definition 7 $u_N^{PES}(s,a) = \min_{s' \in S_{N+1}} \max(1 - \pi_N(s'|(s,a)), u(s'))$

Note that [8] $u_N^{PES}(s,a)$ is the degree of inclusion of the possible successors of s by a in the fuzzy set of goal states: it is the necessity of the fuzzy event : "the subsequent state of s when performing action a will be a goal state". Following a cautious approach, we will prefer actions maximizing $u_N^{PES}(s,a)$:

Definition 8 $\forall (a, a') \in A_{s,N}$ $a \geq_s^{PES} a'$ *iff* $u_N^{PES}(a, s) \geq u_N^{PES}(a', s)$

We define in a classical way the strict order $>_s^{PES}$. The set of optimal actions for s will be denoted by $A_{s,N}^{*PES}$.

Two particular cases are noticeable. First, if $\pi_N(s'|(s, a))$ only involves 0 or 1 possibility levels, we recognize here the usual Maximin Criterion of decision theory (also known as Wald criterion), which also choose the actions that maximize the minimal utility of the possible consequences. Secondly, if u only uses 0 or 1 possibility levels (crisp goal set), $u_N^{PES}(s, a)$ is nothing but $N(Good_N(s, a))$ and $u_N^{PES}(s)$ is $N(Good_N(s))$.

We can then establish the necessity of the fuzzy event "Choosing an optimal action in s leads to a goal state" :

Definition 9 $u_N^{PES}(s) = u_N^{PES}(s, a^*)$ *for a* $a \in A_{s,N}^{*PES}$

which is also the necessity of the event "there is a actions which leads from s to a goal state". Indeed, we have :

Proposition 5 $u_N^{PES}(s) = Sup_{a \in A_{s,N}} u_N^{PES}(s, a)$

Considering the multi-stage decision problem, the pessimistic indices over states or actions can then be defined by induction:

Definition 10

- $u_t^{PES}(a, s) = \min_{s' \in S_{t+1}} \max(1 - \pi_t(s'|(s, a)), u_{t+1}^{PES}(s'))$
- $\forall (a, a') \in A_{s,t}$ $a \geq_s^{PES}$ a' *iff* $u_t^{PES}(a, s) \geq u_t^{PES}(a', s)$
- $u_t^{PES}(s) = u_t^{PES}(s, a^*)$ *for a* $a \in A_{s,t}^{*PES}$

where $A_{s,t}^{*PES}$ *is the set of optimal actions according to* $\geq_s^{PES}$.

In case of non flexible goals, we obviously get back $N(Good_t(a, s))$ and $N(Good_t(s))$. It can also be shown by induction that $u_t^{PES}(s)$ is the *necessity* of the fuzzy event :"there is a policy which leads from s to a goal state". In other terms :

Proposition 6 $u_t^{PES}(s) = Sup_{a \in A_{s,t}} u_t^{PES}(a, s)$

Hence, as in the case of binary utilities on the final states, policies maximizing the necessity to eventually lead to a goal state may be computed backwards, assigning to each state s in stage t, an action in $A_{s,t}^{*PES}$.

3.3 Optimistic utility

According to an *optimistic* point of view, we could say that action a in state s is all the better as the *best possible outcome* s' has a *high utility*. This point of view leads us to the assumption that utilities on final states and transition possibilities are commensurable.

If we assume that $\pi_t(s'|(s,a))$, and $u(s')$ are expressed on the same scale, we can define the optimistic utility functions and the choice criterion at stage N as follows :

Definition 11

- $u_N^{OPT}(s,a) = Sup_{s' \in S_{N+1}} min(\pi_N(s'|(s,a)), u(s'))$
- $\forall (a,a') \in A_{s,N}$ $a \geq_s^{OPT}$ a' *iff* $u_t^{OPT}(a,s) \geq u_t^{OPT}(a',s)$
- $u_t^{OPT}(s) = u_t^{OPT}(s,a^*)$ *for a* $a \in A_{s,t}^{*OPT}$

This criterion is a generalisation to multi-stage decision of the index $E(\pi) = max_{x \in X} min(\pi(x), u(x))$ first proposed by Yager [14] and later used by Kacprzyk [9] (see also Conclusion). $u_N^{OPT}(s,a)$ is the *degree of intersection* of the fuzzy set of possible successors of s when a is applied with the fuzzy set of goal states : it is the possibility degree of the fuzzy event "a in state s will lead to a goal state". Following the same reasoning as in the previous sections, it can be shown that :

$$u_N^{OPT}(s) = Sup_{a \in As,N} u_N^{OPT}(s,a)$$

. The optimal actions according to $\geq_s^{OPT}$ maximize the possibility to reach a goal state. If $\pi_N(s'|(s,a))$ only uses 0 or 1 possibility levels, we get :

$$u_N^{OPT}(s,a) = Sup_{(s'/\pi_N(s'|(s,a))=1)} u(s')$$

In this case, prefering the actions maximizing $u_N^{OPT}(s,a)$ leads to use the usual Maximax criterion of decision theory (which is to choose the action that maximizes the maximal utility of the possible consequences).

Let us now define the utility functions and the choice criterion at stage t as follows :

Definition 12

- $u_t^{OPT}(a,s) = Sup_{s' \in S_{t+1}} min(\pi_t(s'|(s,a)), u_{t+1}^{OPT}(s'))$
- $\forall (a,a') \in A_{s,t}$ $a \geq_s^{OPT}$ a' *iff* $u_t^{OPT}(a,s) \geq u_t^{OPT}(a',s)$
- $u_t^{OPT}(s) = u_t^{OPT}(s,a^*)$ *for a* $a^* \in A_{s,t}^{*OPT}$

where $A^{*OPT}_{s,t}$ is the set of optimal actions according to $>^{OPT}_s$, the strict order induced by $\geq^{OPT}_s$.

The measures generalize $\Pi(Good_t(a,s))$ and $\Pi(Good_t(s))$. Once again, can be shown that :

Proposition 7 $u^{OPT}_t(s) = Sup_{a \in As,t} u^{OPT}_t(s,a)$

This means that the possibility to reach eventually a goal state applying an optimal policy is also the possibility of the fuzzy event :"there is a policy which leads from s to a goal state". Hence, it is enough to choose at each stage an action maximizing $u^{OPT}_t(a,s)$ in order to get a policy that maximize the possibility to lead to a goal state a stage $N+1$.

3.4 Refining the decision ordering

The pessimistic criterion allowed us to define sets of pessimistic optimal policies (*i.e* cautious). Unfortunately, in many real problems, such a criterion may prove to be insufficiently discriminating : there may be too many optimal policies. As in section 2, the idea is to discriminate among actions which have the same pessimistic utility using the optimistic one.

Consider first that we have to choose an action at stage N. The following order will be used, which first prefers actions maximizing the pessimistic criterion and then discriminates using the optimisitic one :

Definition 13 $a \geq_s a'$ iff $u^{PES}_N(s,a) \geq u^{PES}_N(s,a)$ or $u^{PES}_N(s,a) = u^{PES}_N(s,a)$ *and* $u^{OPT}_N(s,a) \geq u^{OPT}_N(s,a)$

Since u^{OPT}_N and u^{PES}_N nare possibilities and necessities of a fuzzy event, they do NOT verify $u^{PES}_N(s,a) > 0 \Rightarrow u^{OPT}_N(s,a) = 1$

An important consequence is that the optimal actions with respect to u^{PES} are not necessarily optimal with respect to u^{OPT}. Hence, choosing an optimal action according to $\geq_s$ does not maximize the possibility to reach a goal state. In other terms, we have :

$u^{PES}_t(s) = u^{PES}_t(s,a^*)$ for a $a^* \in A^*_{s,t} = Sup_{a \in A_{s,t}} u^{PES}_t(a,s)$
but :

$u^{*OPT}_N(s) = u^{OPT}_N(s,a^*)$ for a $a^* \in A^*_{s,N} \leq Sup_{a \in A_{s,N}} u^{OPT}_N(a,s)$
where $u^{*OPT}_t(s)$ (resp $u^{PES}_t(s)$) stands for the possibility (resp. the necessity) to get a goal state applying an optimal action. We have choosen here the notation $u^{*OPT}_t(s)$ for the possibility of this (fuzzy) event, in order to avoid the confusion with $u^{OPT}_t(s)$ which refers to the pure optimistic ordering (in this latter case, we have seen that choosing a optimal action according to $\geq^{OPT}_s$ is equivalent to maximize the possibility to reach a goal state).

More generally, if the set $A_{s,t}$ of possible actions in state s in stage t is finite we can always find an action a^*_{OPT} such that $u^{OPT}(s) = u^{OPT}(a^*_{OPT}, s)$, and an action a^*_{PES} such that $u^{PES}(s) = u^{PES}(a^*_{PES}, s)$. Still in the general case, a^*_{OPT} and a^*_{PES} will be different. Since our option of ordering actions is cautious, we will use the optimistic utility to refine the pessimistic utility ordering. Hence the possibilistic utility of an action at stage t given that an optimal policy is chosen for stages $t+1$ to N is :

Definition 14

$$u_t^{*OPT}(a,s) = Sup_{s' \in S_{t+1}} min(\pi_t(s'|(s,a)), u_{t+1}^{*OPT}(s'))$$

where $u_{t+1}^{*OPT}(s')$ is the possibility of the fuzzy event "applying an optimal policy from stage $t+1$ leads eventually to a goal state". The preorder among actions at stage t refines the pessimistic criterion by the optimistic one:

Definition 15 *$a \geq_s a'$ iff $u_t^{PES}(s,a) > u_t^{PES}(s,a')$ or $(u_t^{PES}(s,a) = u_t^{PES}(s,a')$ and $u_t^{*OPT}(s,a) \geq u_t^{*OPT}(s,a'))$*

We define in a classical way the strict order $>_s$ induced by $\geq_s$ by $a >_s a'$ iff it is not the case that $a' \geq_s a$.An action a will then be *optimal for s* iff there is no action a' such that $a' >_s a$. The set of optimal actions for s will be denoted by $A^*_{s,t}$.

The necessity to reach a goal state applying an optimal policy observing stage s is still $u^{PES}(s)$ and is equivalent to the necessity to have a policy that leads to a goal state (as in section 3.1, we choose actions maximizing this necessity) but the possibility of this event is now :

Definition 16 *$u_{t+1}^{*OPT}(s) = u_t^{*OPT}(s,a)$ for a $a \in A^*_{s,t} \leq Sup_{a \in A_{s,t}} u_t^{*OPT}(s,a)$*

Following this approach, the optimal policy will be computed backward according to $>_s$. This will lead to get policies maximizing the necessity to get a final goal state (see section 3.1) ; among the policies maximizing the pessimistic utility index, the ones maximizing the possibility of this event will be considered as optimal. But once again, there may be a policy which is better according to u^{OPT} but having a lower u^{PES} : policies with higher pessimistic utilities will be prefered to these. Of course, if there is a policy maximizing both utility indices, it will be optimal.

3.5 Computing an optimal policy

The following algorithm is again a variant of dynamic programming. As in Section 2, it is possible to avoid unnecessary computations (we omit the details for space considerations).

algorithm 2;

{ Initialisation }

$d = \emptyset$ {the policy}
for s $\in S_{N+1}$ **loop**
    $u^{*OPT}_{N+1}(s) = u(s)$;
    $u^{PES}_{N+1}(s) = u(s)$;
end loop;

{ Backward computing }

for$t := N$ downto 1 **loop1**
    for $s \in S_t$ **loop2**
        for $a \in A_{s,t}$ **loop3**
            $Compute(u^{PES}_t(s,a))$
            $Compute(u^{*OPT}_t(s,a))$
        end loop3;
        $d_t(s) = a^*$ { a^* maximizes $u^{PES}_t(s,a)$ and then $u^{*OPT}_t(s,a)$ }
        $u^{PES}_t(s) = u^{PES}_t(s,a^*)$
        $u^{*OPT}_t(s) = u^{*OPT}_t(s,a^*)$
    end loop2;
    Add d_t to the policy d
end loop1;

4 Related work and conclusion

The main contribution of this article consisted in extending Dubois and Prade's possibilistic decision theory from the single-stage to the multiple-stage case. We have proposed two successive extensions (the latter generalising the former), corresponding, in our notations, to the 4-uples $\langle N, \pi, G, \text{maximise}(\mathsf{N}, \Pi)\rangle$ and $\langle N, \pi, min, \text{maximise}u^{PES}\ \text{andthen}\ u^{*OPT})\rangle$. We have sketched algorithms for these extensions, in the spirit of dynamic programming.

Apart from Dubois and Prade's possibilistic approach to one-stage decision making, some authors have considered more or less qualitative approaches to decision, either one-stage or multi-stage. Yager [14] proposed to use the optimistic criterion defined by

$$u^{OPT}(s_0, a) = max_{s \in S} min(\pi_{a,s_0}(s), u(s))$$

which is the optimistic counterpart of Dubois and Prade's u^{PES}. As observed by Dubois and Prade [8], this criterion can be overoptimistic. More general criteria for one-stage decision making, recovering optimistic and pessimistic criteria as particular cases, have been proposed by Bolaños, Lamata and Moral [4] and

more recently by Yager [15], Yager and Lamata [16]. Extending these general frameworks to multi-stage decision is worth considering for further research.

Bellman and Zadeh's fuzzy set-based approaches to multistage decision making [3] assume a qualitative utility function (where intermediate utilities of different states of a both are aggregated by min) and transition functions which are either deterministic, either stochastic – these two approaches corresponding thus to $\langle N, D, min, u\rangle$ and to $\langle N, pr, min, \bar{u}\rangle$, respectively. This last approach can be seen as a semi-qualitative, semi-quantitative approach to decision making – the selection criterion for the optimal policy consists in maximizing the expected qualitative utility. On the contrary, our framework is fully qualitative (both on the uncertainty and the utility side). Other significant work about using possibility theory for multi-stage decision making has been done by Kacprzyk [9] and also Baldwin and Pilsworth [1]. Kacprzyk [9] defines an optimal policy by maximising the aforementioned optimistic criterion u^{OPT} at each stage – his approach corresponds thus to the 4-uple $\langle N, \pi, min, u^{OPT}\rangle$; he computes it with a branch and bound algorithm; however, as noticed by Dubois and Prade [8], maximizing u^{OPT} is practically not reasonable; to make the argument simpler, suppose that all transition possibilities are binary ($\forall s, s, s', \pi(s'|(s,a)) \in \{0,1\}$), then maximizing u^{OPT} comes down to assume at each stage that the best possible outcome occurs (in other words, the agent takes its desires for reality). The author also proposes extensions of his approach to the cases where termination time is fuzzy, or infinite. The alternative approach of Baldwin and Pilsworth [1], similar to dynamic programming, also consists in maximising an optimistic criterion, but the search is performed over a set of fuzzy states and the maximisation over a set of fuzzy decisions, which, as noticed by Kacprzyk [9], makes the approach practically unreasonable due to its computational complexity.

Further work would consist first in implementing the proposed algorithms, taking account of the suggested techniques (and possibly others) for avoiding unnecessary computations. Next, we think of generalising our framework in several directions:

- Considering the contribution of intermediate states and actions performed to the global utility function: the most natural way to do it in a purely qualitative way consists in combining these qualitative utilities by the minimum, for instance for u^{PES} we get

$$u_t^{PES}(s,a) = \min(u(s), u(a), \min_{s' \in S_{t+1}} \max(1 - \pi(s'|(s,a)), u_{t+1}^{PES}(s')$$

  and $u_t^{PES}(s) = \max_{a \in A_{s,t}} u_t^{PES}(s,a)$. Integrating these intermediate utilities in the framework described in Section 3 does not present any particular difficulty.

- Retracting the commensurability assumption between possibility degrees and utility degrees. This would enable us to consider mixed qualitative/quantitative frameworks, with, for instance, additive utilities and

possibilistic transition functions – the latter would probably lead us to rank fuzzy numbers in order to define optimal policies.

- Retracting the assumption of full observability. To this purpose we may adapt some methodologies and results from partially observable Markov decision processes. In this case we would have to consider the presence of information-gathering actions in policies.

Lastly, we think that developing qualitative approaches to multiple-stage decision making would provide alternative frameworks for decision-theoretic planning. Namely, some work is currently being done in the direction of applying possibilistic GMDP to an agricultural planning problem [6].

References

[1] J.F. Baldwin and B.W. Pilsworth, *Dynamic programming for fuzzy systems with fuzzy environment*, J. Math. Anal. & Appl. 85, 1-23, 1982.

[2] R.A. Bellman, *Dynamic Programming*, Princeton University Press, 1957.

[3] R.E. Bellman and L.A. Zadeh, *Decision making in a fuzzy environment*, Management Science, 17, B-141-B-164, 1970.

[4] M.J. Bolaños, M.T. Lamata and S. Moral, *Decision making problems in a general environment*, Fuzzy Sets and Systems 25, 135-144, 1988.

[5] Craig Boutilier, *Towards a logic for qualitative decision theory*, Proc. of KR'94.

[6] Celia Da Costa Pereira, Hélène Ducrocq, Michel Duru, Frédéric Garcia, Jérôme Lang and Roger Martin-Clouaire, *Aide à la planification de la gestion d'un atelier de production agricole : cas du pâturage tournant*, manuscript.

[7] Didier Dubois and Henri Prade, *Possibility theory as a basis for qualitative decision theory*, Proc. of IJCAI'95.

[8] Didier Dubois and Henri Prade, *Towards possibilistic decision theory*, Preprints of the 3rd Fuzzy Logic in AI Workshop, Montreal, August 95. Also in Tech. Report IRIT, Toulouse (1995).

[9] Janucz Kacprzyk, *Multi-stage decision making under fuzziness*, Verlag TÜV Rheinland, 1983.

[10] Martin L. Puterman, *Markov Decision Processes*, John Wiley and Sons, New York, 1994.

[11] Martin L. Puterman, *Dynamic programming*, Encyclopedia of Physical Science and Technology, Vol. 4 (1987), 438-463.

[12] Sek-Wah Tan and Judea Pearl, *Qualitative decision theory*, Proc. AAAI'94.

[13] Thomas Whalen, *Decision making under uncertainty with various assumptions about available information*, IEEE TRans. on Systems, Man and Cybernetics, 14, 888-900 (1984).

[14] Ronald R. Yager, *Possibilistic decision making*, IEEE Trans. on Systems, Man and Cybernetics, 9 (1979), 388-392.

[15] Ronald R. Yager, *An approach to ordinal decision making*, International Journal of Approximate Reasoning 12, 237-261, 1995.

[16] Ronald R. Yager and Maria T. Lamata, *Decision making under uncertainty with nonnumeric payoffs*, Tech. Report MII-1525, Iona College, New Rochelle, New York, 1995.