On using arguments for reasoning about actions and values

John Fox
Advanced Computation Laboratory
Imperial Cancer Research Fund
Lincoln's Inn Fields
London WC2 3PX
United Kingdom

Simon Parsons
Department of Electronic Engineering
Queen Mary and Westfield College
Mile End Road
London E1 4NS
United Kingdom

Abstract

Systems of argumentation for handling beliefs about the world have been reported in earlier papers. It seems possible that these systems may also be applicable to reasoning about the effects of actions. However there are substantial differences reasoning about beliefs and reasoning about actions, so a new system of argumentation is required for the latter. This paper makes some preliminary remarks about how the argumentation framework we have introduced elsewhere might be extended to making decisions about the expected value of actions.

Introduction

Standard decision theory (Raiffa 1970) builds on the probabilistic view of uncertainty in reasoning about actions. The costs and benefits of possible outcomes of actions are weighted with their probabilities, yielding a preference ordering on the "expected utility" of alternative actions. However, as Tan and Pearl (1994), amongst others, have pointed out, the specification of the complete sets of probabilities and utilities required by standard decision theory make the theory impractical in complex tasks which involve common sense knowledge. This realisation has prompted work on qualitative approaches to decision making which attempt to reduce the amount of numerical information required.

Work on such qualitative decision making techniques has been an established topic of research at the Imperial Cancer Research Fund since the late seventies (see (Parsons & Fox 1996) for a review). Our early work was partly concerned with the description of human decision processes (Fox 1980) and partly with the practical development of decision systems for use in medicine (Fox, Barber, & Bardhan 1980). Whilst the qualitative decision procedures we developed proved to have considerable descriptive value and practical promise, our desire to build decision support systems for safety-critical fields such as medicine raised the concern that our early applications were *ad hoc*. In particular we were concerned that they, in common with all other expert systems being built at the time, were not based on a rigorously defined decision theory. As a result we have put considerable effort into developing a theoretical framework from qualitative decision making.

The best developed part of this is an approach to uncertainty and belief based on the idea of *argumentation* This approach emphasises the *construction* and *aggregation* of symbolic arguments based on the non-standard logic LA (Fox, Krause, & Elvang-Gøransson 1993; Krause *et al.* 1995). This provides rules for constructing reasons to believe in and doubt hypotheses, and reasons to believe or doubt arguments.

The generality of the everyday idea of argumentation suggests that a similar approach could be taken to reasoning about actions, for instance in deciding on medical treatments or investigations. We might hope to construct arguments for and against alternative actions in the usual way, avoiding issues about the elicitation and use of numerical utilities by representing the desirability and undesirability of actions symbolically. This suggestion immediately raises two questions. Firstly, how well does our formalisation of support and opposition transfer to reasoning about action? Secondly, is the LA directly applicable to arguments about action or will different logics be required? This paper attempts to provide some answers to these questions.

While there are similarities between arguments for and against beliefs and arguments for and against actions this discussion suggests that there are also significant differences, amounting to a requirement for additional rules for assigning values to the outcomes of actions, and for arguing the expected benefits of alternative actions.

This paper argues that the idea of argumentation is applicable to reasoning about actions and values, but that logics of argumentation other than LA will be required. It then makes some preliminary remarks about what these logics should look like. First, however, we present a brief description of LA for those unfamiliar with it.

Arguments about beliefs

In classical logic, an argument is a sequence of inferences leading to a conclusion. If the argument

$$\text{Ax}\frac{(St : G : Sg) \in \Delta}{\Delta \vdash_{ACR} (St : G : Sg)} \qquad \wedge\text{-I}\frac{\Delta \vdash_{ACR} (St : G : Sg)\ \Delta \vdash_{ACR} (St' : G' : Sg')}{\Delta \vdash_{ACR} (St \wedge St' : G \cup G' : \mathsf{comb}_{\mathsf{conj}}(Sg, Sg'))}$$

$$\wedge\text{-E1}\frac{\Delta \vdash_{ACR} (St \wedge St' : G : Sg)}{\Delta \vdash_{ACR} (St : G : Sg)} \qquad \rightarrow\text{-I}\frac{\Delta, (St : \emptyset : Sg) \vdash_{ACR} (St' : G : Sg')}{\Delta \vdash_{ACR} (St \rightarrow St' : G : \mathsf{comb}'_{\mathsf{imp}}(Sg, Sg'))}$$

$$\wedge\text{-E2}\frac{\Delta \vdash_{ACR} (St \wedge St' : G : Sg)}{\Delta \vdash_{ACR} (St' : G : Sg)} \qquad \rightarrow\text{-E}\frac{\Delta \vdash_{ACR} (St : G : Sg)\ \Delta \vdash_{ACR} (St \rightarrow St' : G' : Sg'))}{\Delta \vdash_{ACR} (St' : G \cup G : \mathsf{comb}_{\mathsf{imp}}(Sg, Sg'))}$$

$$\vee\text{-I1}\frac{\Delta \vdash_{ACR} (St : G : Sg)}{\Delta \vdash_{ACR} (St \vee St' : G : Sg)} \qquad \neg\text{-I}\frac{\Delta, (St : \emptyset : Sg) \vdash_{ACR} (\bot : G : Sg')}{\Delta \vdash_{ACR} (\neg St : G : \mathsf{comb}'_{\mathsf{neg}}(Sg, Sg')}$$

$$\vee\text{-I2}\frac{\Delta \vdash_{ACR} (St : G : Sg)}{\Delta \vdash_{ACR} (St' \vee St : G : Sg)} \qquad \neg\text{-E1}\frac{\Delta \vdash_{ACR} (St : G : Sg)\ \Delta \vdash_{ACR} (\neg St : G' : Sg')}{\Delta \vdash_{ACR} (\bot : G \cup G' : \mathsf{comb}_{\mathsf{neg}}(Sg, Sg'))}$$

$$\vee\text{-E}\frac{\Delta \vdash_{ACR} (St \vee St' : G : Sg)\ \Delta, (St : G : \top) \vdash_{ACR} (St'' : G' : Sg')\ \Delta, (St' : G : \top) \vdash_{ACR} (St'' : G'' : Sg'')}{\Delta \vdash_{ACR} (St'' : G' \cup G'' : \mathsf{comb}_{\mathsf{disj}}(Sg, Sg', Sg''))}$$

Figure 1: Argumentation Consequence Relation

is correct, then the conclusion is true. In the system of argumentation proposed by Fox, Krause and colleagues (Fox, Krause, & Elvang-Gøransson 1993; Krause *et al.* 1995) this traditional form of reasoning is extended to allow arguments to indicate support for and opposition to propositions, as well as proving them, by assigning a label to arguments which denote the confidence that the arguments warrant in their conclusions and a set of labels to indicate the formulae used in the deduction. This form of argumentation may be summarised by the following schema:

$$\text{Database} \vdash_{ACR} (\text{Sentence} : \text{Grounds} : \text{Sign})$$

where $\vdash_{ACR}$ is a suitable consequence relation. Informally, Grounds (G) are the formulae used to infer Sentence (St), and Sign (Sg) is a number or a symbol which indicates the confidence warranted in the conclusion. The system discussed here has exactly this basis. We start with a set of atomic propositions $\mathcal{L}$ including $\top$ and $\bot$, the ever true and ever false propositions. We also have the usual set of connectives $\{\rightarrow, \vee, \wedge, \neg\}$, and the following set of rules for building the well-formed formulae (*wff*s) of the language.

- If $l \in \mathcal{L}$ then l is a well-formed formula (*wff*).
- If l is a *wff*, then $\neg l$ is a *wff*.
- If l and m are *wff*s then $l \rightarrow m$, $l \vee m$ and $l \wedge m$ are *wff*s.
- Nothing else is a *wff*.

The set of all *wff*s that may be defined using $\mathcal{L}$, may then be used to build up a database Δ where every item $d \in \Delta$ is a triple $(l : G : Sg)$ in which l is a *wff*, Sg represents confidence in l, and G are the grounds on which the assertion is made. With this formal system, we can take a database and use the argumentation consequence relation $\vdash_{ACR}$ defined in Figure 1 to build arguments for propositions that we are interested in. This consequence relation is defined in terms of rules for building new arguments from old. The rule $\rightarrow$-E, for instance, says that from an argument for St and an argument for $St \rightarrow St'$ one can build an argument for St'. Typically we will be able to build several arguments for a given proposition, and so to find out something about the overall validity of the proposition, we will *flatten* the different arguments to get a single sign. Thus we have a function $\mathsf{Flat}(\cdot)$ from a set of arguments $\mathbf{A}$ for a proposition p from a particular database Δ to the pair of that proposition and some overall measure of validity:

$$\mathsf{Flat} : \mathbf{A} \mapsto \langle l, v \rangle$$

where $\mathbf{A} = \{(l : G : Sg) \mid \Delta \vdash_{ACR} (l : G : Sg)\}$, and v is the result of a suitable combination of the Sg that takes into account the structure of the arguments:

$$v = \mathsf{flat}\Big(\{\langle G_i, Sg_i\rangle \mid (l : G_i : Sg_i) \in \mathbf{A}\}\Big)$$

Together $\mathcal{L}$, the rules for building the formulae, the connectives, and $\vdash_{ACR}$ define a formal system of argumentation LA. In fact, LA is really the basis of a family of systems of argumentation, because one can define a number of variants of LA by using different sets of signs. Each set will have its own functions for handling conjunction $\mathsf{comb}_{\mathsf{conj}}$, implication $\mathsf{comb}_{\mathsf{imp}}$ and $\mathsf{comb}'_{\mathsf{imp}}$, negation $\mathsf{comb}_{\mathsf{neg}}$ and $\mathsf{comb}'_{\mathsf{neg}}$, and disjunction $\mathsf{comb}_{\mathsf{disj}}$ and each set will have its own means of flattening arguments, flat. The meanings of the signs, flattening functions, and combination functions delineate the semantics of the system of argumentation. Thus it is possible to define systems of argumentation based on LA with both probabilistic and possibilistic semantics (Krause *et al.* 1995; Parsons 1996).

Many of the systems built using LA (Fox & Das 1996) use the set of signs (or "dictionary")

$\{++, +, -, --\}$ where $+$ indicates that there is reason to support a proposition, $-$ indicates that there is reason to doubt a proposition, $++$ indicates that there is a reason to think a proposition is true and $--$ indicates that there is a reason to think a proposition is false.

Arguments about actions

At an informal level there appears to be a clear isomorphism between arguments for beliefs and arguments for actions. Suppose we wish to construct an argument in favour of treating a patient with cancer by means of chemotherapy. This might run as follows:

> Cancer is an intolerable condition and should be eradicated if it occurs. It is a disease consisting of uncontrolled cell proliferation. Certain chemical agents kill cancer cells and/or reduce proliferation. Therefore we should treat cancer patients with such agents.

The steps in this argument are *warranted* by some generalised (and probably complex) *theory* of the pathophysiological processes involved in cancer, and theories about what kinds of things are tolerable, desirable and so on. The argument is not conclusive, however, since the conclusion might be rebutted by counter-arguments, as when chemotherapy is contra-indicated if a patient is frail or pregnant.

Such arguments appear compatible with LA and consequently we might consider using LA to construct such arguments. Suppose we summarise the above example in the notation of LA:

$$A : G : +$$

where A is the sentence "the patient should be treated with chemotherapy", G denotes the grounds of the argument (the sequence of steps given), and $+$ indicates that the grounds support action A. However this conceals some significant complexities. The notion of "support" seems somewhat different from the interpretation we have previously assigned to it. For LA we have adopted the interpretation that an argument is a conventional proof, albeit one which it is acknowledged cannot in practice be guaranteed to be correct. An argument in support of some proposition is, in other words, a proof of the proposition which we accept could be wrong. This analysis of "support" does not seem to be entirely satisfactory when reasoning about what we *ought to do* as opposed to what *is the case*. Consider the following simple argument, which is embedded in the above example:

> cancer is an intolerable condition, therefore it should be eradicated

There is a possibility that this argument is mistaken, which would justify signing it with $+$ (a "supporting" argument in LA) but the sense of support seems to be different from that which is intended when we say that the intolerable character of cancer gives support to any action that will eradicate it. In other words when we say "these symptoms support a diagnosis of cancer", and "these conditions support use of chemotherapy" we are using the term "supports" in quite distinct ways. The latter case involves no uncertainty, but depends only upon some sort of statement that intolerable states of affairs ought not to be allowed to continue. If this is correct then it implies that arguing from "value axioms" is not the same thing as arguing under uncertainty and so is inappropriate for constructing such arguments.

How might we accomodate such arguments within our existing framework? One possibility might be to keep the standard form and elaborate the sentence we are arguing about to include a "value coefficient", eg:

$$(A : +) : G : +$$

Which might be glossed as "there is reason to believe that action A will have a positively valued outcome". This may allow us to take advantage of standard LA for reasoning with sentences about the value of actions, but it does not, of course, solve our problem since it says nothing about the way in which we should assign or manipulate the value coefficients.

As a result, we currently prefer another approach, which is analogous to the decision theoretic notion of expected value. In this approach we construct compound arguments based on distinct steps of constructing and combining belief arguments and value arguments. For example, consider the following argument:

> A will lead to the condition C
> C has positive value
> A has positive expected value

which could be represented as:

$$\begin{array}{ll} A \rightarrow C : G : + & e1 \\ C : G' : + & v1 \\ A : (e1, v1) : + & ev1 \end{array}$$

We can think of this as being composed of three completely separate stages as well as having three steps. The first stage, $e1$, is an argument in LA that C will occur if action A is taken, which could be glossed as "G is grounds for arguing in support of C resulting from action A". The second stage, $v1$, says nothing about uncertainty; it simply requires some mechanism for assigning a value to C, call this AV. The final stage concludes that A has positive expected value; to make this step we shall have to give some mechanism for deriving arguments over sentences in LA and AV, call this LEV.

The attraction of this scheme is that it appears to make explicit some inferences which are hidden in the other argument forms. However, it has the additional requirements that we define two new systems—AV and LEV. It seems to us that this is a price worth paying since making the assignment of values and the calculation of expected value explicit gives much more flexibility and so makes it possible to represent quite complex

The patient has colonic polyps	$cp : G1 : ++$	$e1$
polyps may lead to cancer	$cp \rightarrow ca : G2 : +$	$e2$
cancer may lead to loss of life	$ca \rightarrow ll : G3 : +$	$e3$
loss of life is intolerable	$\neg(ll) : av : ++$	$v1$
surgery preempts malignancy	$su \rightarrow \neg(cp \rightarrow ca) : G4 : ++$	$e4$
argument for surgery	$su : (e1, e2, e3, e4, v1) : +$	$ev1$
surgery has side-effect se	$su \rightarrow se : G5 : ++$	$e5$
$\neg(se)$ is desirable	$\neg(se) : av : +$	$v2$
argument against surgery	$\neg(su) : (e5, v2) : +$	$ev2$
se is preferable to loss of life	$pref(se, ll) : (v1, v2) : ++$	$p1$
no arguments to veto surgery	$safe(su) : cir : ++$	$c1$
surgery is preferable to $\neg(surgery)$	$pref(su, not(su)) : (ev1, ev2, p1) : ++$	$p2$
commit to surgery	$do(su) : (p2, c1) : ++$	$a1$

Figure 2: An example argument

patterns of reasoning. As an example of the kind of reasoning that should be possible consider the following:

> (1) The patient is believed to have colonic polyps which, while presently benign, could become cancerous.
>
> (2) Since cancer is life-threatening we ought to take some action to preempt this threat.
>
> (3) Surgical excision is an effective procedure for removing polyps and therefore this is an argument for carrying out surgery.
>
> (4) Although surgery is unpleasant and has significant morbidity this is preferable to loss of life, so surgery ought to be carried out.

Informally we can represent this argument as in Figure 2.

There are six different forms of argument in this example which has a similar scope to the examples considered by Tan and Pearl (1994). The first are those labeled $e1$,...,$e5$ which are standard arguments in LA. The second are value assignments $v1$ and $v2$ which represent information about what states are desirable and undesirable. The third are expected value arguments $ev1$ and $ev2$ which combine the information in standard and value arguments. The fourth are preference arguments $p1$ and $p2$ which express preferences between different decision options on the basis of their expected values making this explicit. The fifth type of argument is the closure argument $c1$ which explicitly states that all possible arguments have been considered, and this leads to the final type of argument, the commitment argument $a1$ which explicitly records the taking of the decision. The following sections discuss some features of these arguments.

Arguments about values

We require some language for representing values. Notwithstanding the common-sense simplicity of the idea of value its formalisation is not likely to be easy. Value assignments are commonly held to be fundamentally subjective—they are based on the preferences of a decision maker rather than being grounded in some observable state of affairs.

There are a number of possible formalisms we might consider. We might, for instance, adopt some set of modal operators, as in $desirable(P)$ or $undesirable(P)$, where P is some sentence such as "the patient is free of disease". This is the approach adopted by Bell and Huang (Bell & Huang 1996; Huang & Bell 1996). Alternatively we might attach numerical coefficients, as in the use of quantitative utilities in traditional decision theory. We propose representing the value of a state or condition C by labelling a proposition describing C with a sign drawn from some dictionary D. For example if we adopt the dictionary $\{+, -\}$ we can represent a positively valued state by the formula $C : +$ and a negatively valued state by $C : -$. Alternatively we can use a dictionary of numbers representing the possible value of states, eg $[0, \infty]$, using these, say, to represent their monetary value.

Some simple value arguments

These simple value dictionaries are analogous to qualitative and quantitative dictionaries for representing uncertainty used by LA. In this discussion we shall only consider qualitative value dictionaries because, as with uncertainty, we can invariably judge whether some state has positive or negative value, or is valueless, though we may not be able to determine a precise point value or precise upper and lower bounds on the value.

Another similarity with our view of uncertainty is that we can frequently assign different values on states from different points of view. For example the use of opiates is bad since they lead to addiction, but good if they are being used as an analgesic. We therefore propose to label value assignment expressions with the grounds for the assignment ie $C : G : V$, giving us a "value argument" analogous to the argument expressions of LA. This is not a new idea of course. For

example, multi-attribute utility theory also assumes the possibility of multiple dimensions over which values can be assigned. However, the benefits of this sort of formalisation is that it may allow us cope with situations where we cannot precisely quantify the value of a situation, and it permits explicit representation of the justifications for particular value assignments making it possible to take them into account when reasoning.

The simplest useful dictionary of values allows us to talk about states that are good or desirable and states which are bad or undesirable.

$$dict(cost\_benefit) =_{def} \{+, -\}$$

As discussed above, there is some ambiguity about the meaning of these signs. For example + could mean simply that the state has some absolute (point) positive value, but the precise value is unknown, or it could mean that we have an argument for the overall value of our goods being increased. In both cases, however, it would seem that good and bad states can be related through a complementation rule.

$$\frac{C : G : +}{\neg C : G : -}$$

There also seems to be some benefit in extending this dictionary to allow us to talk about maximal amounts of goodness (badness):

$$dict(bounded\_cost\_benefit) =_{def} \{++, --, +, -\}$$

However, there seems to be a complication here. It seems straightforward to claim that there is a lower bound on badness—we might gloss this by saying certain conditions are "intolerable" such as death for instance, but an upper bound on "goodness" (eg of a bank balance) seems hard to conceive of. However if we accept:

$$\frac{C : G : ++}{\neg C : G : --}$$

then we can obtain a reasonable interpretation for the idea of a condition which is maximally desirable as the complement of any condition that is intolerable. Furthermore sentences like "human life is priceless" are held, by their users at least, to have some meaning. From a pragmatic point of view such statements can seem merely romantic, but if we accept the above constraint it is a direct consequence of asserting that loss of life is intolerable.

The rest of the discussion will concentrate on the sign subset $\{+, ++\}$ of this dictionary but some remarks will also be made about the whole dictionary.

Basic value assignments

The basic schema of value assignment is analogous to the standard argumentation schema, viz:

$$\text{Database} \vdash_{\text{VCR}} (\text{Condition} : \text{Grounds} : \text{Value}) \quad (1)$$

A basic value argument (BVA) is a triple defining some state, the value assigned to it, and a justification for this particular assignment. The assertions "health is good" or "illness is undesirable" might be represented in grounds-labelled form by:

$$health : va : +$$

where va is a label representing the justification for the BVA. Traditionally there has been considerable discussion of the justifications for value assignments. Any discussion has to face the difficulty that values seem to be fundamentally subjective. In discussion of beliefs there is an analogous idea of subjective probability but frequency theory has provided an objective basis which has led to a formal calculus of probability. There has been a similar attempt to identify an objective framework for values, in consensual values (social mores, legal systems etc), but it seems inescapable that values are grounded in opinion rather than some sort of objective estimation of the chances of events. We therefore accept that a value assignment may in the end be warranted by sentences like "because I say so", "because the law says so", "because the church says so" etc.

In other words we have nothing new to say about the nature of the "value theories" invoked in (1). We shall simply assume that the theory provides a set of universal value assignments. Our task here is not to give or justify some universal set of value assignment sentences (any more than probability theorists are required to provide particular collections of prior or conditional probabilities) but to identify ways in which collections of such value sentences might be manipulated, aiming to take some steps towards the definition of a system AV which is analogous to LA but deals with values rather than beliefs. The assumption is that the assignment of values in sentences like "health is good" depends upon a derivation $(l_1, ..., l_n)$ which bottoms out in some set of BVAs.

Combining arguments about values

We start by considering how to calculate the value of the conjunction of two values. As an example, suppose we have the BVAs:

$$illness : va : -$$
$$expense : va : -$$

then we will require some rules for aggregating the values of the component states to yield a value for the conjunction $(illness \wedge expense)$ and a label representing the justification of this assignment. A reasonable position for these qualitative values seems to be that the overall value of two independent conditions C1 and C2 can be no less than the value of the most valuable individual condition, giving:

$$\frac{C1 : G1 : V1 \quad C2 : G2 : V2}{C1 \wedge C2 : G1 \cup G2 : V3}$$

where $V3$ is $\max(V1, V2)$. In general, of course, values are cumulative and, for example monetary value would normally be viewed as linearly or logarithmically additive. Note that we require that the two conditions must be independent (in some sense to be clarified) or we are exposed to various counter-examples based on interactions.

We can also propose rules for conjunction elimination:

$$\frac{C1 \wedge C2 : G : V1}{C2 : G : V2}$$

where $V2$ denotes an interval whose upper bound is $V1$, and for disjunction introduction:

$$\frac{C1 : G1 : V1}{C1 \vee C2 : G1 : V1}$$

Since we have already given a mechanism for handling negation and it is not currently clear what implication means for value sentences, this is as far as we can go in defining the construction of arguments in AV.

Flattening value arguments

Since values are derived with respect to some value theory we can contemplate different value arguments grounded in BVAs based on different theories. In common with LA value arguments with the same value can be aggregated. A simple summation rule may be acceptable for this but any aggregation rule we might consider should presumably honour the following constraint:

> let $Args$ be some set of arguments that a state S has positive value, then
>
> $$|Args| \leq |Args \cup S : av : +|$$

where $|Set|$ means the aggregate value of the set of arguments that S has positive value. Following previous usage we might refer to the set of arguments as the *case* for S being positively valued, and $|Args|$ as the *force* of these arguments.

Now, a condition may be desirable (undesirable) or absolutely required (intolerable) on some grounds, whereas on other grounds the condition may be valued differently so that there may be conflict between arguments, for instance:

$$C : G1 : +$$
$$C : G2 : -$$

One way to handle this is to have complementary value arguments, $C : G1 : +$ and $C : G2 : -$, cancel out in aggregation, making the flattening function:

> let Args be some set of arguments that a state S has positive value, then
>
> $$|Args| \geq |Args \cup S : av : -|$$

Another alternative, which is more in agreement with qualitative versions of classical decision theory (Wellman 1990; Agogino & Michelena 1993) is to have complementary value arguments lead to indeterminacy.

If we have an argument that a condition has absolute value (its value is one of $\{++, --\}$) then this valuation determines the overall value whatever other value arguments can be constructed unless the opposing value argument also has an absolute value. If value arguments $C : G1 : ++$ and $C : G2 : --$ hold then an overall value for the condition is undefined. The intuition here is that we cannot simply cancel an argument that a condition is absolutely desirable with an argument that it is absolutely undesirable. For example, in discussions of euthanasia we may have an absolute prohibition on killing; this cannot simply be cancelled out by arguing that a loved one's pain is intolerable. There are, of course, no simple decision rules for such situations and we do not want our system to introduce one. We therefore anticipate the need to identify such conflicts:

$$\frac{C1 : G1 : ++ \quad C2 : G2 : --}{C1 \wedge C2 : G1 \cup G2 : \perp}$$

Resolving such conflicts will require some form of metalogical reasoning, something like the opposite of circumscription, in which we introduce new assumptions or theories whose specific role is to overcome such deadlocks. In the euthanasia example, we may appeal to societal "thin end of the wedge" theories for instance in which "society's needs" were not included in the framing of the original decision.

Arguments about expected values

The previous section dealt with the problem of aggregation of value arguments. It remains to provide rules for deriving sentences from combinations of belief arguments and value arguments (ie arguments in LA and AV respectively). Call these expected value arguments. As an example of this kind of derivation, consider the argument "diseases are undesirable, cancer is a disease so cancer is undesirable", which we might represent as:

$$\frac{disease : v1 : - \quad cancer \rightarrow disease : e1 : ++}{cancer : (v1, e1) : -}$$

Conditionals like that in the second premise are concerned with belief (in this case a belief based on an *a priori* definition) which is of course the province of LA. Now, assume we have the following argument in LA:

$$C : e1 : S$$

meaning that we can argue for C with sign S and let us call this argument $la1$. Assume further that we also have the following argument in AV:

$$C : v1 : V$$

which means that the value of C is V, and let us call this argument $av1$. From these two arguments we wish to derive an argument in LEV:

$$ev(C) : (e1, v1) : E$$

meaning that the expected value of C is E.

The important question then becomes, how do we obtain E from the labels V and S? For the set of values $\{+, ++\}$ the following rules seems to apply:

$$\frac{C : la : S \quad C : av : V}{ev(C) : (la, av) : E}$$

where the value of E is given by the following table:

	++	+
++	++	+
+	+	+

When we have an argument in LA to the effect that C definitely holds, the expected value of C can be no less than the value that it is assigned by the argument in AV. When the argument that C holds is not certain, the expected value of C cannot be maximal: therefore since we have only two symbols in the dictionary $ev(C)$ must take the value $+$.

Expected value of actions

From a decision making point of view arguments about expected value of states are of little interest, except in the situation where they are the *outcomes* of actions that we can choose to take or not take. As an example, we will want to reason about sentences concerning action such as:

$$\frac{not(cancer) : v1 : + \quad surgery \rightarrow \neg(cancer) : e1 : ++}{ought\_to\_use(surgery) : (v1, e1) : +}$$

However we eschew derivations of value statements from arguments entirely in LA, such as "the patient has cancer, and surgery prevents cancer so we should carry out surgery":

$$\frac{cancer : e1 : + \quad surgery \rightarrow \neg(cancer) : e2 : ++}{ought\_to\_use(surgery) : (e1, e2) : +}$$

in other words value assignments must eventually be grounded in at least one BVA. In order to reason about the expected value of actions we have to extend the mechanism discussed above. Consider the sentence

action A will give rise to state C

Representing this action as $A \rightarrow C$ we can express this as an atomic argument:

$$A \rightarrow C : la : S$$

What can we conclude from this? Intuitively we want to be able to derive the expected value of an action from the value of its expected consequences:

$$ev(A) : (la, av) : E$$

meaning that the expected value of action A is E. If S has the value $++$ then we are saying that if we carry out action A then C will definitely occur, and if S has the value $+$ then we are saying that if we carry out A then there is a reason to believe that C will occur. In other words we have an identical pattern of reasoning to that just suggested:

$$\frac{A \rightarrow C : la : S \quad C : va : V}{ev(A) : (la, av) : E}$$

where the value of E, as before, is given by the following table:

	++	+
++	++	+
+	+	+

If we allow V to range over the extended dictionary $\{++, +, -, --\}$ we may extend the table by:

	++	+
−	−	−
−−	−−	−

However, we propose no rules for reasoning about the expected value of actions when S is one of $\{-, --\}$.

Flattening expected value arguments

In many cases a collection of qualitative expected value arguments can be aggregated using rules similar to those suggested for AV. In other words flattening could be taken to obey the following constraints:

let $Args$ be some set of arguments that a state S has positive value, then

$$|Args| \leq |Args \cup S : av : +|$$

and

let Args be some set of arguments that a state S has positive value, then

$$|Args| \geq |Args \cup S : av : -|$$

Alternatively flattening could be by having arguments with opposing values give an indeterminate result. It also seems sensible to allow $++$ and $--$ value arguments dominate. However, some qualifications are in order.

Firstly, if we have expected value arguments based on conflicting values, for instance:

$$ev(A) : G1, ++$$
$$ev(A) : G2, --$$

then, as before, such conflicts cannot be resolved within the system.

Secondly, it is not clear how far it is possible to go in handling such conflicts even stepping outside the system. Whereas it seems reasonable to perform a certain amount of reasoning about such conflicts in LA (see

(Elvang-Gøransson & Hunter 1995) for example), this is based upon the fact that what LA is dealing with is in some sense "objective". That is LA is dealing with verifiable facts about the world, and so, since the world is consistent (in the sense that any proposition x cannot both be true and false at the same time), any inconsistency encountered by LA must be the result of a mistake and so can be resolved. Since value arguments are grounded in "subjective" BVAs, rather than objective states of affairs, then there seems little scope for resolving conflicts between arguments in the way we can resolve them in LA. The conflicts are the result of two or more different opinions, none of which need be correct. One might show that one or more set of value assignments violates transitivity of preferences, but there seems to be little more that one can hope to achieve.

Finally, an action may have consequences other than those in which we are primarily interested. In other words actions have side-effects. Certain side-effects can defeat the assumptions on which expected value arguments are constructed. For example, suppose we argue for an increase in income tax, on the grounds that this will generate additional revenue for increased public spending, which is held to be desirable. If we also argue that the tax increase will reduce the incentive to work hard then total income is reduced and hence total revenue will not necessarily increase, which at least weakens and may nullify the original argument. This can be overcome if we can quantify the amounts of revenue involved, but in the present system this kind of logical deadlock can occur.

Preferences and commitments

A complete decision theory is generally held to require some means of choosing between alternative actions. Despite the work outlined above the combined system LA/AV/LEV does not have such a mechanism. However, it is possible to extend the idea of arguments about values and expected values to provide such a mechanism. In particular, we could use expected values to construct a preference ordering over a set of alternative actions as follows:

> Condition $C1$ is *preferred* to condition $C2$, $pref(C1, C2)$, if:
>
> $$|C1 : G1 : +| > |C2 : G2 : +|$$

Transitivity of preferences is implicit in this inequality, and it is also possible to base preferences on the number of opposing arguments. However we have a problem of potential instability analogous to that which arises with uncertainty orderings. We could choose to act on a preference, but the preference could be transitory; wait a little longer and we might find that we can construct an argument to the effect that taking that action could be disastrous. In classical decision theory something like this, the "stopping rule" is discussed but we are not aware of any proposals that really address the stability problem. It is likely that this is inevitable because, as with beliefs, the solution requires a system of meta-level reasoning and circumscription. These concepts are not to be found in classical decision theory.

What is needed is some stronger condition than simply a preference for such and such an action. We would like, for example, to be able to prove that the ordering is, in fact, stable or that the benefits of achieving greater stability are outweighed by the costs. We need some closure condition that says, essentially, there are no further arguments that could alter our main preference, a condition which parallels Pollock's (1992) idea of a practical warrant for taking an action. Abstractly we can think of this as a "safety argument" of the form:

$$\frac{best(A) : G : ++ \qquad safe(A) : cir : ++}{commit(A) : (G, cir) : ++}$$

where $best(A)$ means that aggregation of the arguments for a action A has greater force than the arguments for any alternative action, and $commit(A)$ represents a non-reversible commitment for executing action A, for example by executing it. Informally such safety arguments might include:

- Demonstrating that there are no sources of information that could lead to arguments which would result in a different best action.
- Demonstrating that the expected costs of not committing to A exceed the expected costs of seeking further information.

However, it is clear, as Pollock points out, that any system which is intended to have practical uses should take seriously the computational problems inherent in checking that "no sources ... could lead to arguments". It should also be noted that an idea of commitment similar to that required here has been implemented within the RED system (Das *et al.* 1997).

Conclusions and discussion

We identified a number of different types of argument that can participate in making decisions by reasoning about the outcome of possible actions and have suggested some ways in which these arguments may be built and combined. We believe that the framework we have outlined has the potential to integrate the best parts of traditional planning mechanisms and decision theory in the way suggested by Pollock (1992) and Wellman and Doyle (1991).

While recognising that much remains to be done to provide a secure foundation for this approach to reasoning about action it appears to have potential merit for extending the scope of argumentation to cover a comparable range of decisions to that addressed by classical decision theory. If this holds up then the

complete theory will provide a basis for implementing sound methods for decision making in the absence of quantitative information and the dynamic construction of the structure of the decision. Furthermore, the theory seems to be capable of allowing meta-level reasoning about the structure of the decision topology as well as providing some means for coping with contradictory beliefs and conflicting values and for explicitly including stopping rules and commitment to particular courses of action

In addition to the obvious task of continuing the development of the foundations of this approach, there are a number of areas in which we are working. The first is to refine the set of values and expected values which may be used in order to make the system as expressive as, say, the systems proposed by Pearl (1993)) and Wilson (1995). The second is to investigate alternative semantics for values and expected values as, for instance, Dubois and Prade (Dubois & Prade 1995) have done. The third is to investigate the connections between the model we are proposing and existing means of combining plans and beliefs including the BDI framework (Rao & Georgeff 1991) and the Domino model (Das *et al.* 1997).

References

Agogino, A. M., and Michelena, N. F. 1993. Qualitative decision analysis. In Piera Carreté, N., and Singh, M. G., eds., *Qualitative Reasoning and Decision Technologies*. Barcelona, Spain: CIMNE. 285–293.

Bell, J., and Huang, Z. 1996. Safety logics II: Normative safety. In *Proceedings of the 12th European Conference on Artificial Intelligence*, 293–297. Chichester, UK: John Wiley & Sons.

Das, S.; Fox, J.; Elsdon, D.; and Hammond, P. 1997. Decision making and plan management by autonomous agents: theory, implementation and applications. In *Proceedings of the 1st International Conference on Autonomous Agents*.

Dubois, D., and Prade, H. 1995. Possibility theory as a basis for qualitative decision theory. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1924–1930. San Mateo, CA: Morgan Kaufmann.

Elvang-Gøransson, M., and Hunter, A. 1995. Argumentative logics: reasoning with classically inconsistent information. *Data and Knowledge Engineering* 16:125–145.

Fox, J., and Das, S. 1996. A unified framework for hypothetical and practical reasoning (2): lessons from medical applications. In *Formal and Applied Practical Reasoning*, 73–92. Berlin, Germany: Springer Verlag.

Fox, J.; Barber, D.; and Bardhan, K. D. 1980. Alternatives to Bayes? A quantitative comparison with rule-based diagnostic inference. *Methods of Information in Medicine* 19:210–215.

Fox, J.; Krause, P.; and Elvang-Gøransson, M. 1993. Argumentation as a general framework for uncertain reasoning. In *Proceedings of the 9th Conference on Uncertainty in Artificial Intelligence*, 428–434. San Mateo, CA.: Morgan Kaufmann.

Fox, J. 1980. Making decisions under the influence of memory. *Psychological Review* 87:190–211.

Huang, Z., and Bell, J. 1996. Safety logics I: Absolute safety. In *Proceedings of Commonsense '96*, 59–66.

Krause, P.; Ambler, S.; Elvang-Gøransson, M.; and Fox, J. 1995. A logic of argumentation for reasoning under uncertainty. *Computational Intelligence* 11:113–131.

Parsons, S., and Fox, J. 1996. Argumentation and decision making: a position paper. In *Formal and Applied Practical Reasoning*, 705–709. Berlin, Germany: Springer Verlag.

Parsons, S. 1996. Defining normative systems for qualitative argumentation. In *Formal and Applied Practical Reasoning*, 449–465. Berlin, Germany: Springer Verlag.

Pearl, J. 1993. From conditional oughts to qualitative decision theory. In *Proceedings of the 9th Conference on Uncertainty in Artificial Intelligence*, 12–20. San Mateo, CA.: Morgan Kaufmann.

Pollock, J. L. 1992. New foundations for practical reasoning. *Minds and Machines* 2:113–144.

Raiffa, H. 1970. *Decision Analysis: Introductory Lectures on Choices under Uncertainty*. Reading, MA: Addison-Wesley.

Rao, A., and Georgeff, M. P. 1991. Modelling rational agents within a BDI-architecture. In *Proceedings of the 2nd International Conference on Knowledge Representation and Reasoning*, 473–484. San Mateo, CA: Morgan Kaufmann.

Tan, S.-W., and Pearl, J. 1994. Qualitative decision theory. In *Proceedings of the 12th National Conference on Artificial Intelligence*, 928–933. Menlo Park, CA: AAAI Press/MIT Press.

Wellman, M. P., and Doyle, J. 1991. Preferential semantics for goals. In *Proceedings of the 10th National Conference on Artificial Intelligence*, 698–703. Menlo Park, CA: AAAI Press/MIT Press.

Wellman, M. P. 1990. *Formulation of tradeoffs in planning under uncertainty*. London, UK: Pitman.

Wilson, N. 1995. An order of magnitude calculus. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, 548–555. San Francisco, CA.: Morgan Kaufman.