Artificial Intelligence and Law **7:** 17–35, 1999.

Norms in artificial decision making

MAGNUS BOMAN
The DECIDE Research Group, Department of Computer and Systems Sciences, Stockholm University and the Royal Institute of Technology, Electrum 230, SE-164 40 Kista, Sweden
Phone: +46 8 16 1678; Fax: +46 8 703 9025; E-mail: mab@dsv.su.se
WWW: http://www.dsv.su.se/DECIDE[1]

Abstract. A method for forcing norms onto individual agents in a multi-agent system is presented. The agents under study are supersoft agents: autonomous artificial agents programmed to represent and evaluate vague and imprecise information. Agents are further assumed to act in accordance with advice obtained from a normative decision module, with which they can communicate. Norms act as global constraints on the evaluations performed in the decision module and hence no action that violates a norm will be suggested to any agent. Further constraints on action may then be added locally. The method strives to characterise real-time decision making in agents, in the presence of risk and uncertainty.

Key words: norm, constraint, real-time decision making, decisions with risk, decisions under uncertainty, vague information, policy, social space.

1. Introduction

Artificial agents have been making decisions for some years now, the fact that questions of responsibility and other judicial matters are still open notwithstanding. Their guiding principle has usually been that of maximising their expected utility (PMEU, for short). Wherever there is room for cooperation and coordination, social issues must also come into play. In our immediate future lie artificial decision makers facing sequences of heterogeneous and complex decision situations, perhaps even in real-time. It is reasonable to believe that utterances like "Our anti-sniffer agent decided to classify your agent as harmful" will have meaning in the near future. Just as professional decision makers today often make use of the computational power in the PC sitting on their desk, an option unavailable to them only ten years ago, artificial decision making agents part of a multi-agent system (MAS) will make use of the computational power and tools made available to them. Their rationality might be explicated by their capacity for synthesising results from evaluations that employ different evaluation functions (Boman and Ekenberg, 1995), and not merely by their use of PMEU. The extent to which their

[1] The author gratefully acknowledges the ISES Project (in particular, Rune Gustavsson) at the University of Karlskrona/Ronneby, within which this research was in part carried out.

analyses govern their behaviour will vary, but the representation and modelling of their social context (in terms of their place in the MAS, in their particular coalition, and their alignments) is central.

This article deals with autonomous agents which at least partly adhere to norms. We make the following two provisos, more concise motivations for which are available in (Boman, 1997) and (Ekenberg, 1996) respectively.

Proviso 1: Agents act in accordance with advice obtained from their individual decision module, with which they can communicate.

Proviso 2: The decision module contains algorithms for efficiently evaluating supersoft decision data concerning probability, utility, credibility, and reliability.

The first proviso makes our presentation clearer, because every change of preference (or belief revision, or assessment adjustment) of the agent is thought of as adequately represented in the so-called decision module. This gives us freedom from analysing the entire spectrum of reasoning capabilities that an agent might have, and its importance to the use of the decision module. The communication requirement presents only a lower bound on the level of sophistication of agent reasoning, by stating that the agent must be able to present its decision situation to the decision module, and that the agent can represent this information in the form of an ordinary decision tree.[2] The proviso also lets us separate the important problem of agents failing to obey social norms from the other problems discussed in this paper. Finally, the proviso makes explicit that no nonconsequentialist decision bias affects the choice (Baron, 1994). In the eyes of a consequentialist, artificial agents are closer to the perfect decision maker than human agents can ever hope to be.

The second proviso also requires some explanations. *Supersoft decision theory* is a variant of classical decision theory in which assessments are represented by vague and imprecise statements, such as “The outcome o is quite probable” and “The outcome o is most undesirable” (Malmnäs, 1995). Supersoft agents need not know the true state of affairs, but can describe their uncertainty by a set of probability distributions. In such decisions with risk, the agent typically wants a formal evaluation to result in a presentation of the action (in some sense) optimal with respect to its assessments, together with measures indicating whether the optimal action is *much* better than any other action, using a distance measure. The basic requirement for normative use of such measures is that (at least) probability and utility assessments have been made, and that these can be evaluated by an evaluation function, e.g., PMEU, cf. Boman (1997).

[2] That the decision module is seen as customised is inessential: it is a metaphor for a situation where all agents utilise an oracle, but where the computations of the oracle depend on the individual agent input. In other words, the oracle acts as a decision tree evaluator, and the size of the oracle (in the case of the decision module discussed here, about 6000 lines of C code) is the only thing that might make it inconvenient for the agent to carry it around.

Section 2 discusses problems with representing and evaluating policy. These ideas are then used in Section 3, in which the basic ways of norms acting as normative constraints are introduced. Section 4 consists of a spacious but simple example. The final section offers conclusions and further research.

2. Implementing Policy

Herbert Simon led what could be called a crusade against the implementation of ideal rationality, see, e.g., (Simon, 1982), and that prompted Savage to suggest (Savage, 1954) that decision analytic tools should be applied to *small worlds*, today often referred to as *decision frames*, only. How such models can be made congruent with observation has been studied, e.g, in (Laskey and Lehner, 1994). Computational procedures for artificial decision making can be sketched in the form of the list below, which hopefully gives an intuitive feel for a plausible work cycle more useful for agent programming than similar listings made within the economics literature; see, e.g., Lavoie (1992, p. 56).

1. Formulate exclusive and exhaustive action strategies.
2. Select decision frame.
3. Select agents to trust as information sources.
4. Assess a credibility to each reporting agent.
5. Import selected utility assessments from trusted agents.
6. Evaluate the decision frame using a decision module.
7. Act on, or analyse, the output of the evaluation.
8. Perform sensitivity analyses.

Figure 1. A work cycle for artificial agent decision making.

Working with human decision makers managing departments or entire companies, we have found that constraining utility assessments is a convenient way of enforcing company (or government) norms, often called *policy*, on decisions. There are two ways to proceed, both of which are relevant to norms in MAS. The assessment of credibilities to the reports of agents (item 4 in the list in Figure 1) is pivotal.

One way to proceed is to eliminate actions with disastrous consequences, see, e.g., (Ekenberg, Boman, and Linnerooth-Bayer, 1997). An artificial agent, e.g., a robot should not choose to turn left at a corner inside a mine, for instance, if that could lead to it being destroyed as a result of falling down a shaft, even if the probability for it actually falling is very small. If information is obtained from another agent, the credibility of that agent affects the probability. With numerically precise assessments, action elimination is easily achieved by specifying security levels using two parameters – one for the maximal credibility and one for the minimal utility. An action could then be eliminated if, for instance, an agent with a credibility above 0.75 (relative to this action) reports that the utility of the strategy is below 0.1. The more difficult numerically imprecise case is formalised in (Eken-

berg, Danielson, and Boman, 1997), in which credibility is a functional mapping onto the reals, normalised to 1 if the function is total, and thus interpreted similarly to a relative weight.

A much more controversial way of issuing policy restrictions is to manipulate the utilities on the lowest level, i.e. to adjust the assessments by having an overly positive (or negative) attitude towards some consequences. This can be viewed as a deliberate revision of one's risk attitude. For human decision makers, there are different reasons for such adjustments, ranging from pure fraud to letting invisible consequences come into play. An example of the former would be to accept a bribe for overestimating the gain from making a deal with a particular competitor. An example of the latter could be a manager that wants a deal to be made with a particular competitor but does not want to expose this desire to his employees. The reason could be, for instance, that the manager is in possession of secret information about a future merge with that particular competitor. Sometimes the decision maker is even unaware of this kind of skewness. This is not a problem for artificial agents. For instance, for modelling the actions of market-based agents (Wellman, 1996), analytical game theory (Rasmusen, 1994) with only equally distributed probabilities seems to be sufficient.

Note that PMEU alone cannot model all kinds of risk attitude, see Ekenberg, Boman, and Linnerooth-Bayer (1997). To present a theory that is neither too strong, nor too weak for modelling uncertainty and risk is very difficult even in small worlds, see Boman and Ekenberg (1995). Problems with implementing company or government policy in tools for normative decision analysis are in many ways similar to the problem of constraining agent action in an MAS in order to comply with norms. We now begin our exploitation of these similarities.

3. A Model for Action Constraints

What is missing from the work cycle in Figure 1 is norms. In this section, we will combine the two established and well-known ways of implementing policy of the previous section with an idea developed for MAS. The model in Figure 2 is general in that it makes relatively few assumptions about the agent architecture, language, sensors, and communication protocols. This is not to say that these matters are unimportant. Some choices of agent language would be incompatible with the model, for example. The ambition is to let the model admit a unifying view, and we therefore encourage that it be complemented by additional components as appropriate. Not even the concept of goal is necessary.

Bootstrapping does not present a problem, since no restrictions apply to neither sense data, nor communication data. The only requirement is that the contents of the four bases conform to Proviso 2 above. The concept of agent credibility as used here in imprecise and uncertain domains, cf. (Baron, 1993) was defined in (Danielson and Ekenberg, 1997) and that of reliability in Ekenberg, Danielson, and Boman (1997), in which the four different bases were formalised.

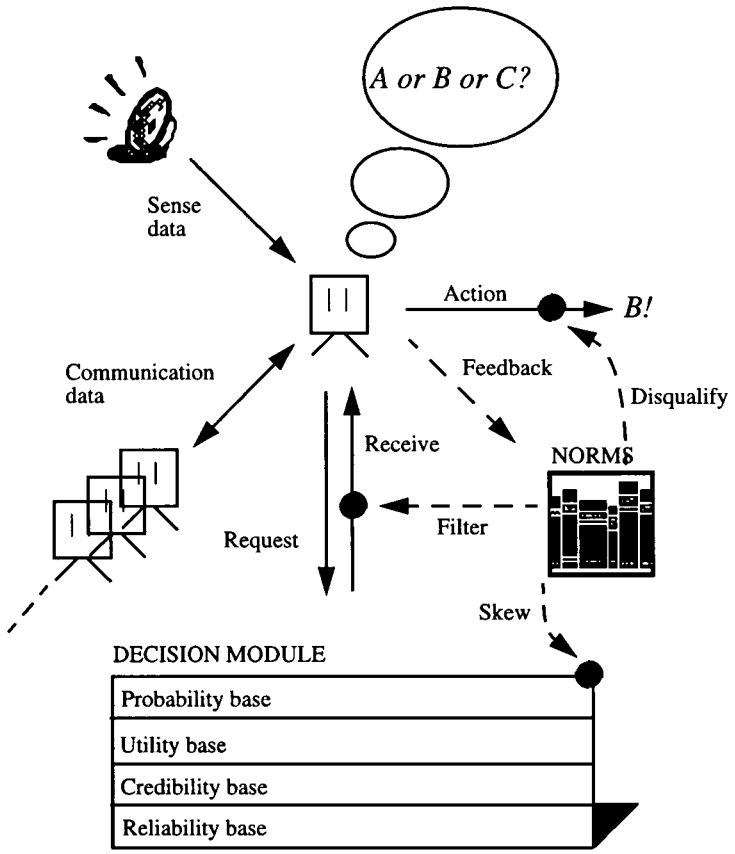

Figure 2. Constraining autonomous agent action.

Before taking an action, an agent might have used its means of communicating with other agents in the MAS, its sensors, as well as the computational power of the decision module. If there are no norms present in the MAS, the four bases in the decision module are non-linear systems of equations representing (typically subjectively assessed) supersoft data about

- probabilities of the occurrence of different consequences of actions
- utilities of outcomes of different consequences of actions
- credibilities of the reports of other agents on different assessments
- reliabilities of other agents on different agent ability assessments

The preferences of the agents can be stated as intervals of quantitative measures or by partial orderings. Credibility values are used for weighting the importance of relevant assessments made by other agents in the MAS. Reliability values mirror the reliability of another agent as it in turn assesses the credibility of a third agent, see Ekenberg (1996). All bases except the utility base are normalised. Note that an

MAS without norms is treated is this paper as a social structure where group utility is irrelevant to the individual agent. The presence of norms can manifest itself in at least three ways, each representing a different level of abstraction:

1. Through skewing of the equations in the four bases.
2. By filtering normative advice before it is received by the agent.
3. By disqualifying certain actions by referring to their negative impact on the global utility in the MAS.

The first way of constraining action through norms is by manipulation of the utilities on the lowest level, i.e. to skew assessments to have an overly positive (or negative) attitude towards some consequences (in comparison to the actual beliefs of the agent). The agent might first make assessments true to its own beliefs, and then revise these assessments to better cope with the given situation.

The second way proceeds via the elimination of actions with disastrous consequences. Both ways were discussed in the previous section, and will be thoroughly examined in the following example.

The third item in the above list requires that a global utility measure is available on the local level. It is quite possible that the agent has first skewed some of its assessments to comply with a set of norms, second that it has rejected the recommended action because it was deemed too risky (and there might be a norm stating that agents should avoid actions that violate certain preset security levels), and third that it, when considering the second best alternative, finds that even this alternative cannot be chosen. The reason is that although this action is the rational choice of the agent in view of the norms present, another kind of threshold value is violated, viz. the extent to which the global utility of the MAS is reduced should the action be taken. This form of social norm adoption was implicitly suggested in Ekenberg (1996), in which global utility among autonomous agents is formalised. This problem has emerged as one of the critical points of MAS research, see, e.g., Conte and Castelfranchi (1995), Ekenberg, Danielson, and Boman (1996), Jennings and Campos (1997), Kalenka and Jennings (1997).

Our model is highly individualistic. In open systems, this position is easily defended by noting that agents are less inclined to work toward group goals. We claim, however, that the model applies equally to cooperating agents, thanks to the three ways of affecting agent action just explained. Jennings and Campos (1997) identify the third form of social norm adoption above as a basic requirement for autonomous agent systems that they call the *individual-community balance*. It is a constructive and computationally efficient way of attaining socially responsible behaviour (Jennings, 1992) and an attempt at describing the *social level* of agents.

The first two kinds of norm adoption have already been implemented, see Danielson and Ekenberg (1997), but their functionalities relative to the third kind of norm adoption (the group utility constraint) is yet to be examined. A technical but important note is that all norms cannot be described as linear constraints, see Ekenberg, Boman, and Linnerooth-Bayer (1997). As we have seen, however, such

norms can play the same role as risk attitudes in that they affect security level settings, sensitivity analyses, and other ways of extending PMEU.

4. Example

This example is intended to ground our abstract claims in a hopefully realistic scenario of a real-time MAS in which decision modules as well as social norms play a role.

Assume a very sophisticated two-legged robot R. The robot is moving around in a room, looking for radioactive material. There are other robots in the room with the same ultimate goal as R. Robot R holds partial, vague, imprecise, and uncertain (possibly false) data about the room, about the abilities of the other robots, and about its own abilities. This information can be partitioned into:

(i) statistical information,
(ii) sensor information, and
(iii) normative advice.

Assume further that R experiences a malfunction in one of its legs. The fault is pre-classified as serious, and R has been programmed not to take any physical action before a serious error has been rectified. Instead, R will seek help from the environment and its inhabitants. In its environment, R has a decision module ever ready to assist, given that R can formulate its problem as a decision tree. This is extremely difficult, but R can be thought of as following a template developed for serious errors, something which makes the procedure viable. Aside from being recognised by R as serious, the fault leads to a repair type template, i.e. a decision situation in which R must decide upon a repair action that ultimately leads to the fault being corrected.[3] The very least way in which cooperation with other agents will affect R is through its subjective assessments. If one is less defensive in describing the skills of R, one may easily imagine a vast range of more interesting ways of other agents influencing R, but we will try in this example to make as few assumptions about R as possible.

The first thing R has to do in any repair situation is to formulate a number of alternative actions. The template yields two alternatives: for R to repair the leg itself, or for R to seek assistance from another agent in the room. For the latter alternative, a prerequisite is that there is another agent capable of carrying out the repair. Whether this condition is fulfilled or not is up to R. For each agent that R thinks is in the room, R must be the judge of the appropriateness of that particular agent carrying out the repair. If this value of appropriateness is larger than zero for an agent $R2$, choosing $R2$ represents an alternative. Note that the

[3] Note that in order for the example to be realistic, one should imagine an inheritance hierarchy of faults and further that R (perhaps in conjunction with other agents) can determine the exact place of the leg fault in the hierarchy, and thus the proper procedure for rectification.

reflexive case where $R = R2$ need not be treated separately. The value results from a combination of values in the subjective assessments of R. Those assessments can be made using probabilities, utilities, credibilities, and reliabilities. To keep this example moderate, we take only the first two kinds into consideration here. A simpler example that does deal with credibility assessments can be found in Boman (1996). We also leave aside the important knowledge acquisition problem. After considering the ability of all the agents, R formulates the following three alternatives.

> A1: R will do the repair on its own leg.
> It has the ability to carry out the repair, but not to replace the entire leg with a better one. Instead, R is forced to use spare parts customised for this older type of leg. The customised parts are more expensive and less reliable than those parts that would repair a new kind of leg, should it be necessary. Moreover, after R has repaired the broken parts of its old leg, other parts of the leg will probably break in the near future, a process speeded up by the partial mismatch (with respect to age and material) caused by the customised parts.
>
> A2: T will do the repair on R's leg.
> Agent T, a robot in the room, will carry out the repair by replacing the entire leg with a leg of the same type and quality as the old one. R has limited experience of T, but R believes that T is unfamiliar with leg replacements and also that it is unlikely that T could be committed to returning later to make minor adjustments to the leg, should such become necessary.
>
> A3: V will do the repair on R's leg.
> Agent V, a robot in the room, will carry out the repair by replacing the entire leg with a new kind of leg. This leg has several extra features that R as of now is not capable of using, but should R modernise itself further in the future, it could make full use of the new leg. R has plentiful experience of V, and believes that V is highly capable of replacing the leg. R also believes that V could be committed to return later to make minor adjustments to the leg, should such become necessary.

The consequences of each alternative must be phrased as exhaustive and exclusive. To think that an artificial agent would completely master the kind of modelling necessary for using this form of representation, and to do it in real-time, seems extremely optimistic. Again, one should picture streamlined decision situations. Robots such as R could perhaps master fault diagnosis and rectification situations, but little else. The templates mentioned earlier could make the transition from sense data to subjective assessments as smooth as possible. Moreover, subjectively assessed probabilities could in fact be little more than typed (to the agent, that is) facts from a statistical database and hence close to objectively true. If we allow ourselves the abstract view in which such steps have been taken, we can perhaps imagine the identification of the following consequences in our example.

A1C1: The leg will be inoperable for a significant portion of the following three months, and will not function well when in operation. *R* will be slower and less adequate in handing in its reports of radioactive findings than before the leg failure.

A1C2: The leg will be inoperable for a small portion of the following three months, and will not function well when in operation. *R* will be as fast as before, but less adequate in handing in its reports of radioactive findings than before the leg failure.

A1C3: The leg will be inoperable for a small portion of the following three months, and will function almost perfectly when in operation. *R* will be as fast and as adequate in handing in its reports of radioactive findings as before the leg failure.

A2C1: The leg will be inoperable for a small portion of the following three months, but will not function well when in operation. *R* will be as fast as before, but less adequate in handing in its reports of radioactive findings than before the leg failure.

A2C2: The leg will be inoperable for a small portion of the following three months, and will function almost perfectly when in operation. *R* will be as fast and as adequate in handing in its reports of radioactive findings as before the leg failure.

A3C1: The leg will be inoperable for a small portion of the following three months, and will function perfectly (better than before the replacement of the leg) when in operation. *R* will be as fast and as adequate in handing in its reports of radioactive findings as before the leg failure.

A3C2: The leg will be inoperable for a small portion of the following three months, and will function perfectly (better than before the replacement of the leg) when in operation. *R* will be as fast and as adequate in handing in its reports of radioactive findings as before the leg failure. In addition, *R* is (thanks to its state-of-the-art leg) able to walk through a part of the room which was not available to it before. Due to *R*'s inexperience in that part of its environment, however, *R* will not function as well there as in the rest of the room.

A3C3: The leg will be inoperable for a small portion of the following three months, and will function perfectly (better than before the replacement of the leg) when in operation. *R* will be as fast and as adequate in handing in its reports of radioactive findings as before the leg failure. In addition, *R* is able to walk through a part of the room which was not available to it before, in which *R* will function as well as in the rest of the room.

In the domain at hand, utility is measured by the number of hazardous pieces of metal found. The estimated numbers of such pieces for the respective consequences are the following.

A1C1	20–40	A1C2	35–50	A1C3	50–60
A2C1	35–50	A2C2	50–60		
A3C1	50–60	A3C2	60–80	A3C3	70–100

Note that the representation allows for partial and imprecise data. If the utilities involved are based on precise sense data, they are linearly transformed to real values in the interval [0, 1].[4]

Each alternative action takes time to carry out, and this time varies with the alternative. During this time, *R* must be idle. The negative effects of this idle time can be measured in terms of pieces missed, and are estimated as follows.

Alternative 1	5–15
Alternative 2	10–12
Alternative 3	25–30

This leaves us with the following set of intervals for the utility of the different consequences.

A1C1	5–35	A1C2	20–45	A1C3	35–55
A2C1	23–40	A2C2	38–50		
A3C1	20–45	A3C2	30–55	A3C3	40–75

The above numbers reflect utilities, with 0 representing the lowest utility.[5] The utilities above will be mapped onto intervals represented by linear inequalities. The interval representation is very powerful. Note, for instance, how the degree of uncertainty affects the size of the intervals. Naturally, precise numbers can be used if such information is available: The utility value 37, for instance, is represented by the interval [37, 37].

For the probabilities, we will again neglect the important problems related to knowledge acquisition. The probability base is harder to model than the utility base, since the assessments are normalised to a [0, 100] scale (hence the requirement of exhaustive and exclusive consequences). Still, the representation is extremely powerful, not least because of the possibility to state consequences of the kind

[4] For simplicity, we will stick to one decision frame. One could otherwise imagine, e.g., one frame that concentrated on energy aspects of robot movements in a room, another frame on temperature in the room, and a third frame on environmental aspects chiefly concerning increased radioactivity induced from the activities in the room. Details of multi-frame evaluation can be found in Danielson and Ekenberg (1997).

[5] The underlying method does not require that the user of a decision module has expressed opinions on every consequence, as is the case here. Sometimes, this is even undesirable from a computational viewpoint.

"none of the other consequences of this alternative". Such catch-all consequences can be assigned a utility and a probability as any other consequence, and this in a sense allows the agent to sidestep the frame problem. Just as for the utilities, we imagine a set of assessed intervals out of the blue, and we again stress that values could be left out – this would just be interpreted as a "0–100%" assessment.

A1C1	5–25%	A1C2	10–30%	A1C3	45–70%
A2C1	25–45%	A2C2	50–80%		
A3C1	60–80%	A3C2	5–15%	A3C3	5–10%

Before we turn to the evaluation of these assessments, we should mention that R is free to induce any partial ordering on the utility or probability values, as a complement or as a replacement for the interval statements. A sophisticated form of comparison between different values without stating their numerical range are so-called *comparative statements*, or *links*. To first illustrate utility links, consider the following estimated differences in utility between consequences.

A1C2–A1C1	5–15
A1C3–A1C2	10–25
A2C1–A2C2	10–25
A3C2–A3C1	10–20
A3C3–A3C2	10–30

To illustrate probability links, consider the two identical consequences A1C3 and A2C2 above. Now, assume that the probability of R functioning that good is at least 15 per cent higher if T carries out the repair than if R carries out the repair itself. That would result in a link of the kind $p(\text{A2C2}) - p(\text{A1C3}) \geq 15\%$.[6]

Evaluating this decision tree, we get the output shown in Figure 3. In this output, the first alternative (that had R repair its own leg) is treated as the base case for comparison to the other two alternatives. Put simply, A1 lies on the axis of abscissa, and is the second best alternative. The best alternative is A2: to let T carry out the repair. The worst alternative is A3: To let V carry out the repair. The difference between A1 and A3 is very small. The numbers on the axis of abscissa denote the contraction of the original intervals by an increasing percentage. On the value of 100, we have full contraction and at this stage each interval has been reduced to a single point. By contrast, on the axis of ordinate, we have more or less the original intervals. Here we can see that the difference A2–A1 is about four times bigger than A1–A3, but even A2–A1 is not a big difference, since the difference of 4 should be read on a [0, 100] scale. Hence, if humans were carrying out the analysis, they should carry out a stability analysis of the result: Could there be sensitive variables,

[6] The tree is presented in Figure 4 as it appears in the DELTA Decision Tool (Walter, 1997), the GUI to DELTA (Danielson, 1997). The upper bound added is called a *sanity limit* in the terminology of DELTA.

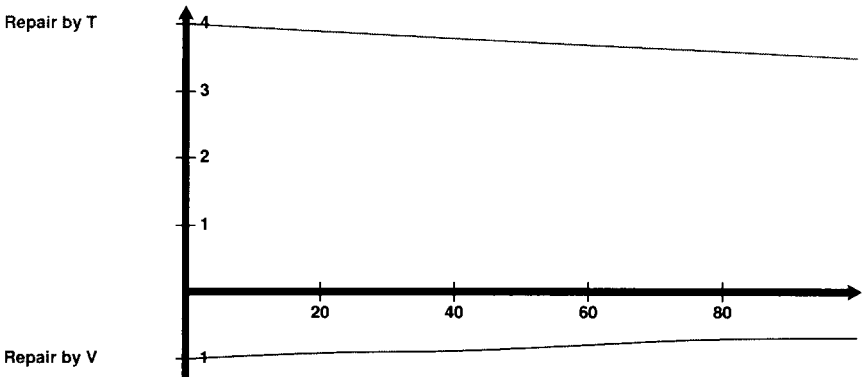

Figure 3. Evaluation output, as shown in the Delta Decision Tool.

small changes to which could have dramatic effects on the outcome?[7] Presumably, some humans would find this result counter-intuitive. However, *R* is not human and would merely be sent the output consisting of the message: "Choose A2!". We now turn to the illustration of how the other agents in the room could affect this situation through norms.

In Figure 2, norms are pictured as being separate from the decision module. In fact, DELTA contains features for revision of assessments, and for the setting of, and calculating with, security levels. In other words, the first two kinds of norm described in Section 3 can be represented within the decision module. Conceptually, however, it might be convenient to think of the decision module as containing only procedures that implement PMEU. At any rate, we will illustrate these two kinds of norm here, as a continuation of our example. The third kind of norm has not yet been implemented, so let us comment on that first.

The pure PMEU evaluation recommended *R* to let *T* carry out the repair. Now, assume that there are six robots in the room, and that several of them have experienced leg failures of a similar kind to the one that *R* is currently having. Robot *T* has for some time carried out almost all the repairs and hence has a fine reputation within the MAS: *T* is a responsible and reliable agent. A problem is that *T* is very much in demand. Robot *V* has just recently announced its capability of repairing legs and does not have a good reputation. In addition, *R* is the only agent capable of repairing its own legs. It has no reputation because it has not announced its capability, as it is unaware of its potential use to others. Currently, one of the other robots, call it *W*, is idle because of a leg failure. Robot *W* has so far been the most efficient robot in the room. Therefore, its idle time considerably affects the group utility negatively. It could therefore be the case that from the point of view of the

[7] As it turns out, these results are quite stable.

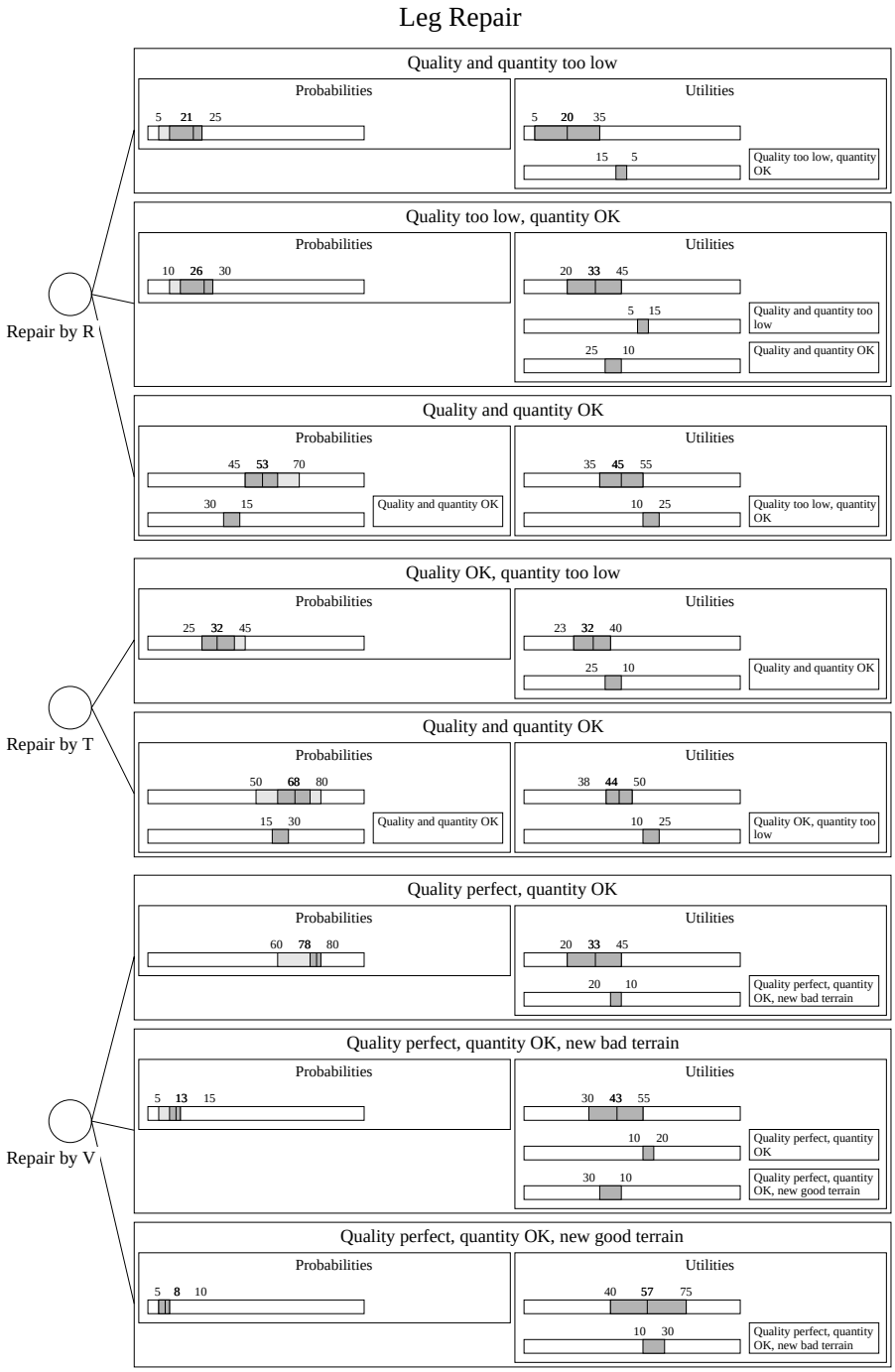

Figure 4. The decision tree with links for the example.

group, *R* should not make *T* commit to helping *R* with the repair at this stage. Instead, *R* should repair itself and allow *T* to help *W* instead.[8]

It is easier to be explicit regarding the first and second kind of norm. Let us assume first that *R* has a reason (that in mentalistic terms could be called a *desire*) to repair its own leg. It could then use DELTA to look for critical variables in the decision situation. As was noted earlier, such variables are scarce in this example: the example is stable. So *R* will either have to skew several variable values, or to skew one (and then the best choice is A2C2) considerably. It could, for example, change the probability link value $p(\text{A2C2}) - p(\text{A1C3}) = [0.15, 0.30]$ to $p(\text{A2C2}) - p(\text{A1C3}) = [-0.15, 0]$. If this admittedly highly visible change in input (see Figure 5) is complemented by a relatively minor change to the upper bound of the utility of consequence A2C2, the output of Figure 4 would be substituted by the one in Figure 6. Since the output consists of a non-qualified recommendation of action alternative, the slight superiority of A1 to A2 shown in Figure 6 suffices for *R* to fulfil its concealed goal. Its actual possibility of manipulating the result in this way will depend on its own reasoning capabilities, as well as the elements of control available to the group. We cannot pursue these important matters here.

The second kind of norm is related to the risk profile of *R*. Let us assume that *R* is sensitive to the level of adequacy of finding material that is actually radioactive. In particular, one wishes to avoid cases where *R* passes radioactive material by without reporting it. The consequences that include expectations of diminished adequacy and their probabilities and utilities are:

A1C1	5–25%	5–35
A1C2	10–30%	20–45
A2C1	25–45%	23–40
A2C2	50–80%	38–50

The interval representation makes these assessments complicated to use in calculations by hand, but DELTA offers assistance in the form of its *security level setting* function. This function allows the user to specify a maximum deficiency in utility tolerated for an alternative. In the example, the agent could have set a level that would filter away A1, A2, or both on the grounds that the alternative that got weeded out had a risky consequence. Risky means that the alternative has a consequence the utility of which is unacceptably low.[9]

Since A2 was the best alternative and does contain consequences that could be deemed risky, it is interesting to investigate under which circumstances A2 would be filtered out. An attempt to explain the semantics of Delta output begins with noting that in Figure 7, the security level has been set to 30 in the utility base.

[8] To actually carry out the evaluation using the explicit rule given in Ekenberg (1996), we would need to say more about the situation; to describe the ternary relationship between pairs of agents and the time required for a repair, to describe the binary repair ability relation, etc.

[9] What is unacceptable is set by the user, but depending on the level of sophistication of the agents involved, it is perhaps more natural to think of these border values as preset.

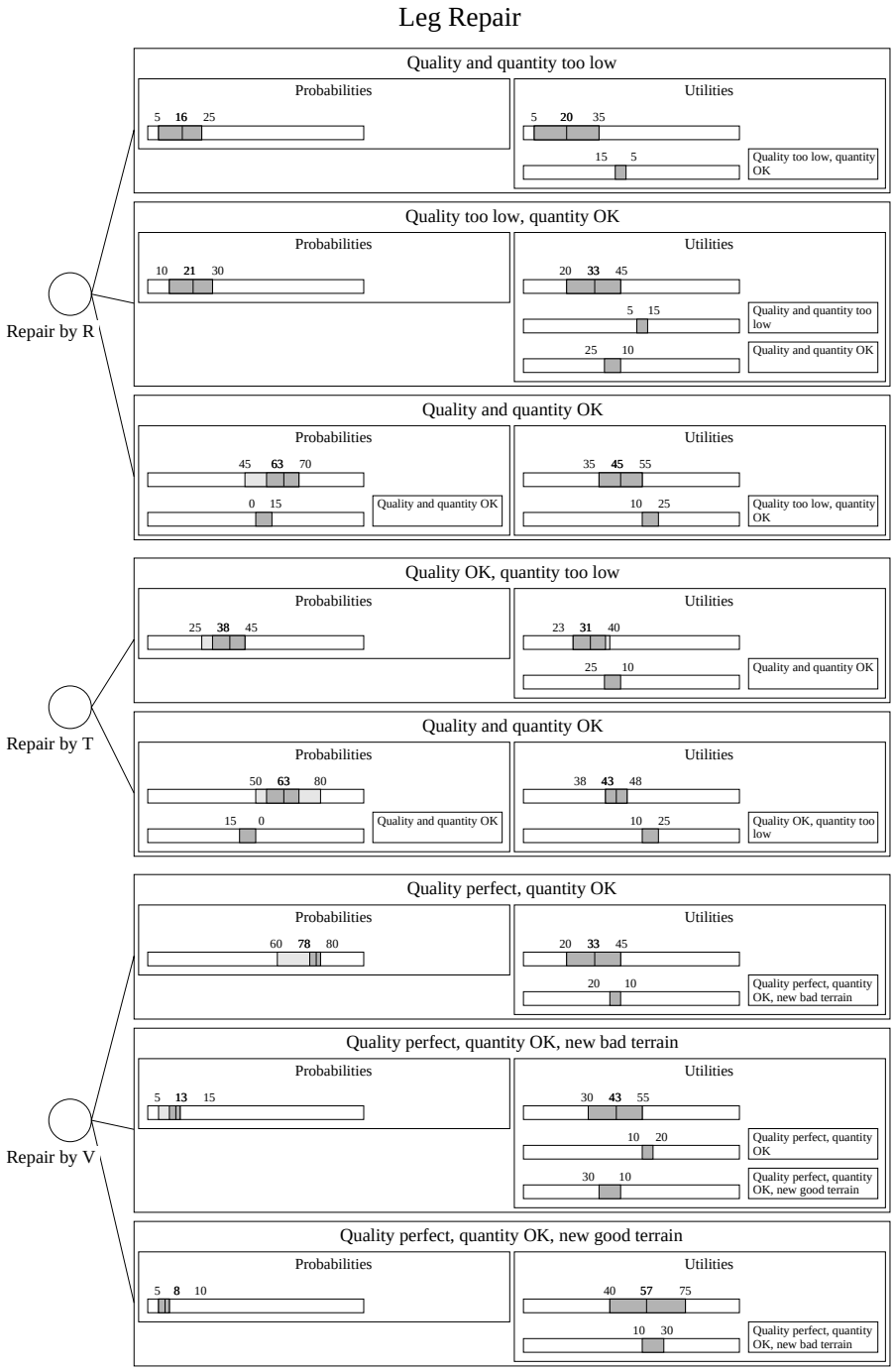

Figure 5. The decision tree with consequence A2C2 skewed.

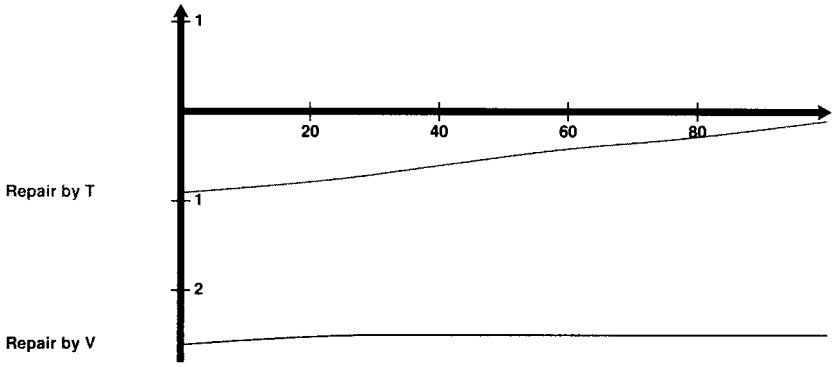

Figure 6. Skewed evaluation output of the DELTA Decision Tool.

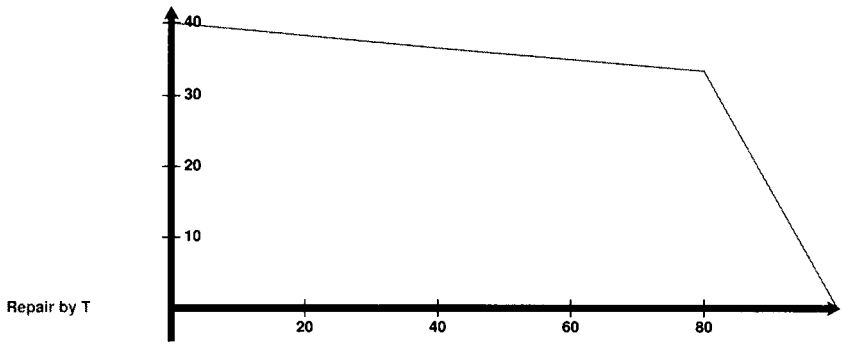

Figure 7. The sensitivity of A2 to a security level set to 30.

Analysing how risky A2 is relative to this level, we can see from the diagram that there is a 40 per cent risk of an unacceptably bad consequence occurring: We have just defined unacceptably bad as having a utility less than 30. The risk drops slightly with the percentage of contraction of the original intervals.

What is shown in Figure 7 is a worst case analysis. The interval representation makes many consistent pointwise assignments possible. Such situations are often studied using Monte-Carlo simulations, in which the interval ranges are bombarded with consistent points. Our method instead uses a deductive technique. The diagram actually has two more graphs plotted, but these are invisible since they lie on the axis of abscissa. These two graphs represent the average case and the best case. In the average case, A2 is thus risk-free with the security level set to 30. Whether it should still be filtered out is again a matter of how risk-prone the agent is, or put differently: a matter of how conservative the set of norms is, as preset by the designer of the MAS. Similarly, a more conservative view on adopting new strategies could be manifested by raising the security level to 40. This results in the diagram shown in Figure 8, which is easier to comprehend than Figure 7. It shows that up to about 20 per cent contraction, A2 will (with absolute certainty) not pass the grade

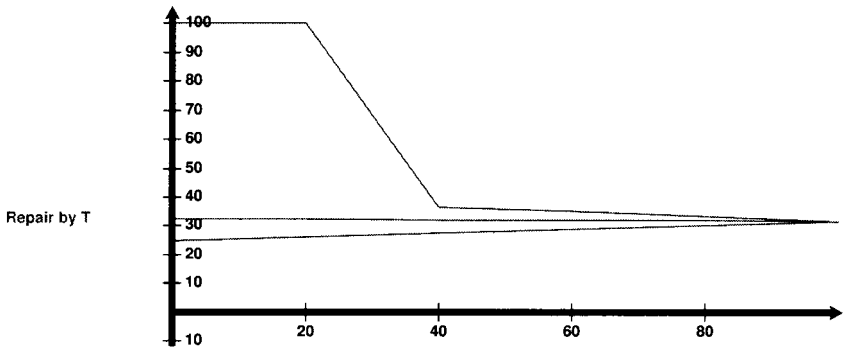

Figure 8. The sensitivity of A2 to a security level set to 40.

of being risk-free in the worst case. Even on average, the risk of A2 resulting in a consequence with a utility lower than 40 is more than 30 per cent. The important point is not the figures particular to this example, of course, but the potential for control that a set of norms might offer for autonomous consequentialistic agents.

5. Conclusions and Further Research

We have presented a model for constraining action using norms. The model constrains normative advice provided by a decision module and operates on three levels of abstraction. The lowest level deals with manipulation by non-benevolent (even malicious) agents, with modifications of assessments as a result of sensitivity analyses, and with more or less *ad hoc* adoption to social norms by means of very delicate belief revision. The middle level deals with the filtering of certain actions in accordance with the risk profile of the agent. This is the natural level for a coordinator of a group of agents to use if the coordinator wishes to implement policy. An agent that strives to become part of a coalition may also let the norms adhered to within the coalition act as filters on this level, but if the agent has reasons not to make its desires public, it should revise its assessments on the lowest level instead. Once it is part of the coalition it will probably have to adhere to the social norms of the group. The highest level of abstraction deals with the acceptance of social norms. Since the basis of our evaluation is PMEU, we do not allow agents to diminish the utility of the group that they belong to by their choice of action. This is a constructive interpretation of the principle of social rationality. For systems that utilise other rules for generating rational behaviour, this rule could be modified, but the important idea is that social norms usually affect rational choice only at the highest (social) level.

Given that the advice received by the agent is actually followed, i.e. if the agents are consequentialistic, the three levels together constitute a *metanorm* (Axelrod, 1986), a formalised procedure for enforcing norms. This is interesting in itself, since many researchers have claimed that such models are of philosophical interest

only because of their high computational complexity. We have given here the first pieces of evidence that this is not the case. The procedures already implemented help in the difficult transition from precise and unrealistic quantitative decision making by artificial agents to qualitative analyses.

Acknowledgments

The author would like to thank Love Ekenberg, and also Mats Danielson, Johan Walter, Per-Erik Malmnäs, Harko Verhagen, Paul Davidsson, and the participants of the *ICMAS'96 Workshop of Norms, Obligations, and Conventions* for discussions in relation to this paper.

References

Axelrod, R. 1986. An evolutionary approach to norms, *American Political Science Review* **80**(4), 1095–1111.

Baron, J. 1993. *Morality and Rational Choice*, Kluwer, Dordrecht.

Baron, J. 1994. Nonconsequentialist decisions. *Behavioral and Brain Sciences* **17**(1).

Boman, M. and Ekenberg, L. 1995. Decision making agents with relatively unbounded rationality, In *Proc DIMAS'95*, AGH, Krakow, pp. I/28–I/35.

Boman, M. 1996. Implementing Norms through Normative Advice, In *Proc ICMAS'96/IMSA'96 WS on Norms, Obligations, and Conventions*.

Boman, M. 1997. Norms as constraints on real-time autonomous agent action, In M. Boman & W. Van de Velde (eds.), *Multi-Agent Rationality (Proc MAAMAW'97)*, LNAI 1237, Springer-Verlag, pp. 36–44.

Conte, R. and Castelfranchi, C. 1995. *Cognitive and Social Action*, UCL Press, London.

Danielson, M. 1997. *Computational Decision Analysis*, Ph.D. diss., DSV, KTH, Stockholm.

Danielson, M. and Ekenberg, L. 1997. A framework for analysing decisions under risk, *European Journal of Operations Research*. To appear.

Ekenberg, L., Danielson, M., and Boman, M. 1996. From local assessments to global rationality, *Intelligent Cooperative Information Systems* **5**(2&3), 315–331.

Ekenberg, L. 1996. Modelling decentralised decision making, In *Proc ICMAS'96*, AAAI Press, pp. 64–71.

Ekenberg, L., Danielson, M., and Boman, M. 1997. Imposing security constraints on agent-based decision support, *Decision Support Systems* **20**, 3–15.

Ekenberg, L., Boman, M., and Linnerooth-Bayer, J. 1997. Catastrophic risk evaluation, IIASA Report No. IR-97-045. Intl Institute for Applied Systems Analysis, Laxenburg, Austria.

Jennings, N.R. 1992. Towards a cooperation knowledge level for collaborative problem solving, In *Proc 10th ECAI*, 224–228.

Jennings, N.R. and Campos, J.R. 1997. Towards a social level characterisation of socially responsible agents, In *IEEE Proc on Software Engineering*, Vol. 144, No. 1, pp. 11–25.

Kalenka, S. and Jennings, N.R. 1997. Socially responsible decision making by autonomous agents, In *Proc ICCS'97*.

Laskey, K. and Lehner, P. 1994. Metareasoning and the problem of small worlds, *IEEE Trans on Systems, Man, and Cybernetics* **24**, 1643–1652.

Lavoie, M. 1992. *Foundations of Post-Keynesian Analysis*, Edward Elgar.

Malmnäs, P.-E. 1995. *Methods of Evaluations in Supersoft Decision Theory*. Unpublished manuscript, available on the WWW using URL: http://www.dsv.su.se/ mab/DECIDE.

Rasmusen, E. 1994. *Games and Information*, 2nd edn, Blackwell.

Savage, L. 1954. *The Foundations of Statistics*, John Wiley & Sons.
Simon, H.A. 1982. *Models of Bounded Rationality*, MIT Press.
Walter, J. 1997. *A Decision Tool for Uncertain Decision Making*, M. Sc. thesis, NADA, KTH, Stockholm.
Wellman, M.P. 1996. Market-oriented programming: Some early lessons, In S. Clearwater (ed.) *Market-Based Control*, pp. 74–95.