Cooperative Plan Selection Through Trust

Nathan Griffiths and Michael Luck

Department of Computer Science, University of Warwick, Coventry, CV4 7AL, UK
Email: {Nathan.Griffiths, Michael.Luck}@dcs.warwick.ac.uk

Abstract. Cooperation plays a fundamental role in multi-agent systems in which individual agents must interact for the overall system to function effectively. However, cooperation inherently involves an element of risk, due to the unpredictable nature of other's behaviour. In this paper, we consider the information needed by an agent to be able to assess the degree of risk involved in a particular course of action. In particular, we consider how this information can be used in the process of plan selection in BDI-like agents.

1 Introduction

BDI agent architectures represent an important class of system that has been used in an increasing number of applications. Based on the folk-psychology mental notions of belief, desire and intention, they are also distinctly popular among the plethora of existing agent architectures. Apart from the intuitive understanding of the BDI model, this popularity may be due to the successful practical application of BDI systems such as PRS [8] and dMARS [5] to diverse areas including malfunction handling on the space shuttle and air traffic control, for example.

Though BDI architectures occupy an an important place in the design of intelligent agents, the limitation of the approach is that it is typically focussed on what might be called *standard* task planning and execution for individual agents. In contrast, our work is concerned with the *extension* of a BDI-like architecture to include those higher-level control strategies and other modifications so that domain properties of multi-agent environments can be used to hone agent behaviour. In this paper, we consider how to augment the process of plan-selection in such agents to provide a richer and potentially more effective mechanism.

We begin by reviewing key components of the target architecture, including the basic control cycle and details of the actions and plans that form the basis of this work. Then we describe the problem of plan selection and examine some of the relevant factors that may be used in multi-agent domains, and proceed to develop a detailed model of plan selection. Before finally assessing the value and contribution of this work, a mechanism for the critical application of the model to partial plans is presented.

2 The Base Architectural Model

The basic operation of BDI agents is based around their beliefs, desires and intentions. An agent has beliefs (about itself, others and the environment), desires (in terms of

the states it wants to achieve in response) and intentions as adopted plans. In addition, agents also maintain a repository of available plans, known as the *plan library*. Agents respond to changes in their goals and beliefs, resulting from perception, by selecting appropriate plans from the plan repository and then instantiating one of these plans as an intention. Intentions comprise actions and subgoals to be achieved, with the latter giving rise to the addition of new subplans to that intention.

This control cycle, while proving generally effective and useful in many domains, does not, however, relate to the specific issues that arise in multi-agent scenarios where cooperation among multiple interacting agents is either necessary or desirable. For example, the questions of who to cooperate with, and how, are not addressed at all. In this paper, we are specifically concerned with the impact of multi-agent plans on the plan-selection process in this kind of architecture. However, before considering plan selection itself, we must describe the nature of such plans.

For an agent situated in a multi-agent environment to take advantage of others, its plans must include a means for it to interact with those others. Cooperation may take the form of an agent performing an action on behalf of another, a group of agents performing an action together or set of actions performed at the same time. Thus there are three distinct types of action that a plan may include, described below.

Individual actions are those performed by an individual agent, without the need for assistance from others. An individual action may be executed by the agent owning the plan in which it is contained, or by another agent on its behalf.

A *joint action* is a composite action, made up of individual actions that must be performed together by a group of agents. Each agent involved in executing a joint action makes a simultaneous *contribution* to the joint action, corresponding to the component action that it performs. For example, if agents α and β perform the joint action of lifting a table, then α must make the contribution of lifting one end of the table simultaneously with β lifting the other.

Concurrent actions are those that can be performed in parallel by different agents, without the need for synchronisation (except at the beginning and end of a set of concurrent actions). As with joint actions, the action an agent performs as part of a set of concurrent actions is its *contribution*. For example, if agents α and β each write a chapter for a book, and they perform their actions in parallel, then α and β perform concurrent actions where each agent's contribution is the action of writing the appropriate chapter.

Our definitions of joint and concurrent actions are related to the notions of strong and weak parallelism described by Kinny *et al.* [9]. The key difference is that while we consider the component actions, or contributions, that make up a joint action, Kinny represents joint actions as a "black-box" without explicit contributions. These are primitive actions from which others can be constructed. Thus, related and dependent actions that do not fit directly into these categories can be built up from them.

In the BDI model, the plans in a plan library are *partial plans* in that they are incomplete, and contain subgoals in addition to actions. We do not consider the arguments for and against such a choice of representation here, since it is has been addressed elsewhere, but note that this is one standard form of organisation [1].

We thus define a plan as sequence of steps, where a step is either an individual action, a joint action, a set of concurrent actions, or a subgoal. In addition, since plans apply only to particular situations, they must also have preconditions.

3 Cooperative Plan Selection

In the BDI model, an agent's actions are determined by its intentions. When an agent forms an intention to achieve a given goal, it does so by committing to a plan to achieve that goal. However, for any particular goal there may be several plans to achieve it that are *applicable* in the current situation, since their preconditions are satisfied. Some of these plans may contain actions beyond the agent's capabilities (or joint or concurrent actions) which, if chosen, require assistance from another agent.

Thus, an agent's choice of plan determines whether it must cooperate to achieve its goal. If all the applicable plans for a goal contain actions that cannot be performed by the agent alone, cooperation is *necessary*. If there is a choice between plans that are performable alone and those that are not, then cooperation is *optional*. If choosing to cooperate in this case, there must be some inherent advantage to the cooperation, for example by minimising effort, since it can also be achieved by the agent alone.

In existing work, several researchers have considered the situation where cooperation is necessary. For example, Castelfranchi *et al.* have developed a model of cooperation based upon the notion of *dependence*, where an agent is dependent on another for an action if it is unable to perform that action itself [3]. However, the issues involved in determining why an agent might choose to cooperate when this is optional, have been largely unaddressed. One exception is Wooldridge and Jennings' [15] formalisation of cooperative problem solving, in which they argue that it begins with an agent recognising the potential for cooperation, either because it is unable to achieve its goal alone (and cooperation is necessary), or because it prefers assistance (and cooperation is optional). Since their work is relatively high level, though, many details such as *why* an agent might prefer assistance are not considered. The approach to plan selection described in this paper is an attempt to answer this question.

3.1 Plan Selection Criteria

The problem of plan selection amounts to choosing the best plan — the plan that is most likely to be successful, with least cost in terms of time and resources, and the least *risk*. (While in some circumstances, such as gambling, the influence of these factors may be contradictory, requiring an agent to make a trade-off between the two, we assume that in general an agent's high-level desires are likely to be such as to attempt to minimise both the risk and the cost of its actions.) When the plans involved do not involve other agents, standard plan selection criteria (or planning heuristics) can be used to assess cost. However, when one or more of the agent's plans do involve others, an element of *risk* is introduced by the inherent uncertainty of interaction. In addition to a measure of the cost of a plan, therefore, we need to be able to assess the likelihood of finding an agent (or agents) for actions that are required for successful plan execution; the likelihood that once such agents are identified they will agree to cooperate; and the

likelihood that once a commitment has been given, the agents concerned will fulfill their commitments.

We identify four primary factors relevant in comparing plans in respect of risk: knowledge of other's capabilities, risk from others, knowledge of view of self, and knowledge of other's preferences. Certainly, risk may be introduced for any number of other reasons, but these are the key domain-independent general issues.

Agent Capabilities Knowledge of others' capabilities helps to determine which agents might perform the required actions. If many agents are known to have the target capabilities, then successful execution of the plan is more likely than if fewer or no agents do so. However, in line with the motivating concerns of dynamic environments and uncertain and incomplete knowledge, we cannot assume that an agent's knowledge of others faithfully represents them, and success at execution time may be possible even if it is not at evaluation time, just as failure is also possible. In general, though, we assume that there is sufficient stability for this to be useful in assessing plans prior to execution.

Risk from Others Once potential cooperating agents are identified, they may be evaluated in terms of the risk involved in interacting with them. Plans involving agents with whom interaction is more likely to be successful, should be rated higher than those involving interactions less likely to be successful.

Risk from view of Self Knowledge of the view of oneself in the eyes of others in terms of risk of interaction may also be useful in assessing plans. It can provide a measure of the likelihood that another agent will agree to cooperate, since an agent is more likely to cooperate with another if it has confidence in the success of that interaction. Thus, the agents identified in competing plans can be evaluated in respect of their view of the risk involved in cooperating with the planning agent. It is, however, difficult to maintain an assessment of how one is viewed by others.

Agent Preferences It might also be possible to assess plans in relation to the higher-level motivations of the agents involved in them, and whether cooperation would be likely. This would require a detailed model of the motivations and goals of other agents, however, which is unlikely to be accurate.

3.2 Trust

How then to assess risk in interaction? Fortunately, as recognised by several researchers [2, 4, 7, 10, 11], this has a relatively simple solution in the form of *trust*. The risk of whether to cooperate and with whom, may be determined by, among other things, the degree of confidence or *trust* in other agents. Despite the notion of *trust* being commonplace in our everyday interactions, there are few formal definitions. However, it is generally accepted that trust implies some form of risk, and that entering into a trusting relationship is choosing to take an uncertain path that can lead to either benefit or cost depending on the behaviour of others [12].

In this paper, we view trust as one of the means available to an agent for estimating the risk involved in cooperation, in terms of an estimation of the degree of expectation that others will do what they agree to do, i.e. an *expectation of risk*. This is a synthetic notion of trust since, unlike Deutsch [4] and Luhmann [10], for example, we are not concerned with how trust operates in humans, but with how the concept of trust can be used in relation to cooperation between artificial agents. We are also primarily concerned with *how* an agent can use the degree of trust it has in another in reasoning about cooperation, rather than how an agent determines this degree of trust in the first place.

4 A Model of Cooperative Plan Selection

4.1 Plan Ratings

The problem of plan-selection is essentially the same as that of finding effective heuristics for plan construction. In that sense, we can apply standard domain-independent heuristics for evaluating plans which perform a valuable, if limited, service. These heuristics include, for example, the length of a plan as the number of its actions, the cost based on the cost of the actions it contains, and the duration of plan execution based on the duration of individual actions. We will not consider this further in this paper, since these issues are well addressed by textbooks (for example [14]), but suffice it to state that any such heuristics may be used to arrive at an assessment of a plan in terms of its *standard rating*.

This evaluation of a plan does not, however, address our key concerns of assessing plans in relation to the dynamic multi-agent nature of the environment. If one or more of the plans available to an agent requires interaction with another, the *standard rating* is inadequate, since this interaction introduces an element of risk. A second rating is therefore necessary in these terms, which we call the *cooperative rating*.

4.2 Trust

The perceived risk of cooperating with a particular agent is determined by that agent's reliability, honesty, etc., embodied by the notion of *trust*. Thus an agent can use its trust in others as a means of assessing the risk involved in cooperating with them. Describing *trust* in terms of *risk* allows us to consider the limits of trust more precisely, and to quantify it. An agent with a high trust value is more trusted than an agent with a low trust value, and represents less risk in terms of cooperation, for example. This suggests an inverse relationship between trust, T, and risk, R, as follows.

$$R = \frac{1}{T} \tag{1}$$

An agent's trust of another is dependent on a variety of factors, including the other's believed reliability, honesty, veracity, etc. However, modelling all such potentially relevant factors is excessive, and can add to the complexity of the solution, when typically they will not be needed. Consequently, we base our model of trust upon Marsh's formalism [11] and the work of Gambetta [7], and define the trust in an agent α, to be a

value from the interval between 0 and 1: $T_\alpha \in [0, 1]$. The numbers merely represent comparative values, and are not meaningful in themselves. Values approaching 0 represent complete distrust, and those approaching 1 represent complete, blind trust. In this paper we are not concerned with how an agent should update its trust of others, but Marsh [11] describes a possible approach that will suffice. This representation of trust corresponds to Marsh's notion of *general trust*. However, Marsh also introduces *situational trust*, where an agent's trust in another is dependent on the importance of the situation being considered. For example, while an agent may trust another to extract product information from a database, it might not trust it to determine which product represents the best value for money. Although conceptually this situational trust is a more powerful mechanism that general trust, the computational overhead involved in identifying trust in *tasks* can be prohibitive, and we do not consider it further.

4.3 Agent Models

In order to choose between plans that may require cooperation for their execution, an agent needs some knowledge about the other agents that it may cooperate with. Durfee [6] notes that in order to cooperate effectively an agent may need to know certain information about others, about themselves, about how they view others and are viewed themselves and so on. However, since an agent's reasoning is resource bounded, if taken to an extreme, the amount of knowledge an agent possesses to facilitate its cooperation might overwhelm its limited reasoning capabilities. Thus agents need just enough knowledge to coordinate well, and no more, since any additional knowledge may simply hinder the reasoning process of the agent.

An agent has a *model* of each other agent with which it may interact, that contains its knowledge of the other's capabilities and the degree to which it is trusted. These agent models form part of the agent's wider knowledge base, or beliefs. The conceptual form such models may take in an agent's knowledge base is shown in Figure 1, which represents an agent's models of two others, α and β. For each agent, the model contains a set of capabilities, and the degree of trust in that agent.

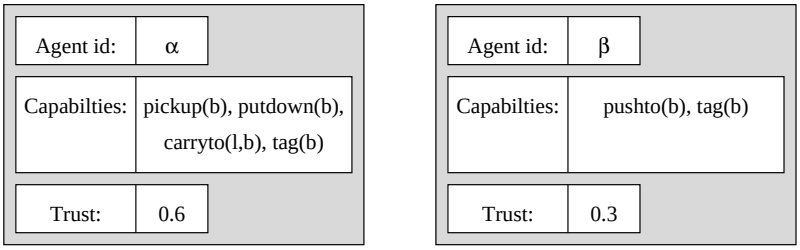

Fig. 1. Example agent models

4.4 Assessing Actions

In assessing the merit of a plan, an agent must make a judgement about the risk attached to each action in the plan requiring cooperation, by examining the trust value in its model of each of the possible cooperating agents. Suppose that an agent knows of n others, $\alpha_1, \alpha_2, \ldots, \alpha_n$, with the required capabilities for performing a given action, and ordered such that $T_{\alpha_{x-1}} \geq T_{\alpha_x}$, where T_{α_x} denotes the trust in α_x. Several possibilities for assessing the risk involved in cooperating with others are discussed below.

We might only consider trust in the *most trusted* agent involved so that the risk of a particular action would be as follows.

$$R_{\text{action}} = \frac{1}{T_{\alpha_1}} \tag{2}$$

Though simple, the problem with this approach is that this most trusted agent might not be the actual agent involved in the cooperative action, for any number of reasons. In particular, the autonomous nature of agents underlying this model suggests that it is impossible to determine the behaviour of another agent in advance. As a consequence, cooperation with less trusted others may be needed, and this must be factored into the measure of risk. Alternatively, then, we might consider the additive total of trust in all agents in the set of potential agents for the action.

$$R_{\text{action}} = \frac{1}{\sum_{i=1}^{n} T_{\alpha_i}} \tag{3}$$

This avoids the problem of only considering the most trusted agent, and considers all agents to an equal extent, but does not address the decreased likelihood of cooperation with less trusted agents. An agent would first try to cooperate with α_1 and, if unsuccessful, would then try α_2, and so on, but for each successive agent, the likelihood of success decreases. To address this, we can adjust the formula to increase the significance of more trusted agents, by dividing the trust of successive agents by a correspondingly increasing factor.

$$R_{\text{action}} = \frac{1}{\sum_{i=1}^{n} \frac{T_{\alpha_i}}{i}} \tag{4}$$

Thus, trust in all relevant agents is considered, but in relation to the likelihood of cooperation with them. Using this measure of risk, we can determine the *cooperative rating* of a plan by summing the risk associated with each action in it. Thus a plan with few high risk actions may be rated better (or less risky) than a plan with many low risk actions. For a plan with m actions, $a_1, a_2, \ldots, a_m$, the cooperative rating C for that plan is given by the following equation.

$$C = \sum_{i=1}^{m} R_{a_i} \tag{5}$$

4.5 Plan Quality

Once both the *standard* and *cooperative* ratings of a plan have been determined, they must be combined to form an overall measure of plan quality to select between alternative applicable plans. It would not be sensible simply to add the two values together, since one measures the *cost* of the plan, and the other the *risk* involved in it, and the relative importance of these may vary for each agent. We therefore include a weighting for these ratings for a particular agent in the overall quality measure, Q, as follows, where w_s and w_c represent the influence weighting applied to the *standard rating*, S and *cooperative rating*, C, respectively.

$$Q = (w_s * S) + (w_c * C) \tag{6}$$

Different agents may use different weightings, the values used reflecting, in part, an agent's predisposition, since agents that place greater importance on the *standard rating* are inclined to minimise the cost of achieving their goals, whether or not this requires cooperation. Conversely, agents that place most importance on the *cooperative rating* are predisposed to minimising the risk involved in cooperating with others, even if this increases the cost involved in achieving their goals. Thus agents that place more importance on the standard rating are more inclined to take risks associated with cooperation in order to minimise the cost of their plans, when compared to agents that place more importance on the cooperative rating. The values of the weighting that provide the best selection of plans depends on an agent's environment.

5 Cooperation in Partial Plans

5.1 Plan Evaluation

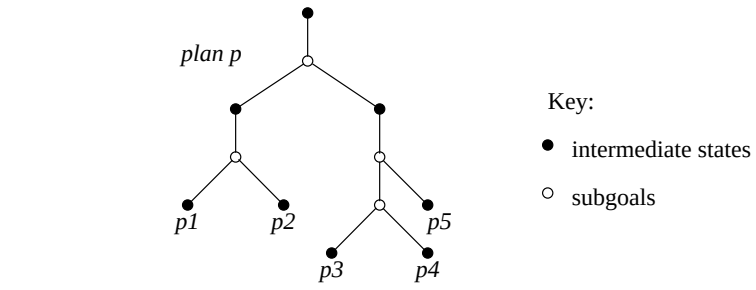

Fig. 2. Example plans

Figure 2 shows a graphical representation of a plan that includes all possible elaborations, where the edges represent actions, solid bullets correspond to intermediate

states between actions, and outline bullets correspond to subgoals. For each subgoal in the plan, there are a set of applicable plans, each of which forms a branch of possible elaboration from that subgoal. The set of plan elaborations is the set of paths from the root of the graph to the leaves. Thus, for plan p possible elaborations are paths from the root to the nodes labelled $p1$, $p2$, $p3$, $p4$, and $p5$. If this set has been determined, the alternatives can be evaluated in respect of the criteria developed for fully elaborated plans, and an appropriate plan selected.

A naive solution would thus be to require an agent to fully elaborate each of its applicable plans in order to choose between them. While this would indeed allow direct use of the criteria described above, it also requires a premature commitment to a particular plan. Such a requirement would negate the benefit of using partial plans in the possibility of interleaving execution and deliberation to cope with the environmental change that is typical of multi-agent scenarios. More importantly, it demands a search through the entire tree of plans so that the quality of each possible path solution can be measured. This is prohibitively expensive to be performed in real-time.

We assume, for reasons of simplicity, that plans are not recursive, meaning that a plan should not contain a subgoal that is the same as the top-level goal that plan achieves.

5.2 Pre-Execution Plan Assessment

If we are to avoid constructing the entire search tree at the time of plan selection, we must be able to make a choice based on a limited number of alternatives, such as the top-level applicable plans. An informed choice at this level is only possible, however, if we have some measure of the value of plans in terms of the standard and cooperative ratings, but clearly, this is not possible to do on the fly. Instead, we perform an off-line *pre-execution assessment* of the plan library in which all of the plans in it are evaluated in a coarse fashion with respect to the agents required for successful execution. This approach represents a compromise between the desire to minimise the computational overhead and that of maximising the quality of any measure of the value of a plan.

Starting with the plans that require no further elaboration, since these are the only ones which can be directly evaluated, the *standard* and *cooperative* ratings are determined. These ratings must then be fed back into the other plans as values for subgoals within them. For each plan containing actions that cannot be performed by the planning agent, the set of all agents known to have the relevant capability is generated through inspection of its agent models, so that these ratings can be calculated as described earlier. There are two possible approaches to incorporating these values for fully elaborated plans into the larger partial plans of which they might form subplans.

Firstly, these values can be used in subsequent levels of plans in the library for which the plans *best* satisfy subgoals, and so on until each plan has an overall quality measure. This quality measure is an assessment of the *best-case* solution.

An alternative solution is to take into account *all* possible elaborations and calculate a *mean* rating for competing plans, so that there is less reliance on one individual plan that may not be possible at execution time. This provides a less sensitive measure, but one which is more likely to be useful in a dynamic environment, since it might still be relevant. The balance between the *best-case* and *mean* ratings amounts to a trade-off

between an agent trying to find the best final plan and minimising the chance of the final plan being poor due to environmental change (in terms of these ratings). These best-case and mean ratings for agent plans will need periodic reassessment as the agent's knowledge of other's capabilities (and its trust in them) changes.

The *best-case advantage* (BCA) of one plan over other applicable plans is the advantage of that plan over others if its final elaboration has the best quality rating. Thus, for two applicable plans, p and q, with best-case ratings of $Q_b(p)$ and $Q_b(q)$ respectively the BCA is equal to the difference between the quality rating for p and that for q, $|Q_b(p) - Q_b(q)|$. If there are more than two applicable plans, as is typical, then the BCA is equal to the difference between the minimum and maximum best-case ratings. Thus with applicable plans $p, q, \ldots, z$ the BCA is determined by the following equation.

$$BCA = max\{Q_b(p), Q_b(q), \ldots, Q_b(z)\} - min\{Q_b(p), Q_b(q), \ldots, Q_b(z)\} \quad (7)$$

The *mean-case advantage* (MCA) of one plan over other applicable plans is the typical (or mean) extra advantage. This is a general case measure that incorporates more information, since it takes into account all possible elaborations of the applicable plans. With mean ratings for p and q of $Q_m(p)$ and $Q_m(q)$, the MCA is equal to $|Q_m(p) - Q_m(q)|$. As above, if there are more than two applicable plans, the MCA is equal to the difference between the minimum and maximum mean ratings. Thus with plans $p, q, \ldots, z$ the MCA is as follows.

$$\begin{aligned} MCA = {} & max\{Q_m(p), Q_m(q), \ldots, Q_m(z)\} \\ & - min\{Q_m(p), Q_m(q), \ldots, Q_m(z)\} \end{aligned} \quad (8)$$

Selecting between partial plans There is a trade-off between maximising the best-case advantage and the mean-case advantage. If the best-case advantage of a plan p over another, q, outweighs the mean-case advantage of q over p, then p should be selected, but if the mean-case advantage of q over p is greater than the best-case advantage of p over q, then q should be selected.

More generally, the advantage should be maximised, regardless of whether it is best case or mean-case. If $BCA > MCA$ then the best-case rating should be used to select plan x, such that $Q_b(x) < Q_b(p) \wedge Q_b(x) < Q_b(q) \wedge \ldots \wedge Q_b(x) < Q_b(z)$. Alternatively, if $MCA > BCA$ then the mean-case rating of the applicable plans should be used.

Certainly, more sophisticated mechanisms involving the likelihood of elaboration of individual plans are possible, but these require much more extensive knowledge of the relationship of plans and environments, and the nature of change in environments, as well as significantly more costly computation. Given that the environment is largely unpredictable, there is unlikely to be any significant advantage, however.

This approach is suited to situations in which the likelihood of the environment and the agent models remaining the same is high so that plan elaboration at execution time is likely to reflect the plan quality values determined in advance for the overall partial plan concerned. Reassessment of these quality measures will be required periodically to ensure they are consistent with the changes in trust of others. Although we do not address

action	effect and cost	performable by
$pickup(b)$	pick up box b, (cost 2)	α_1
$putdown(b)$	put down box b, (cost 2)	α_1
$carryto(l, b)$	carry box b to location l, (cost 2)	α_1
$pushto(l, b)$	push box b to location l, (cost 4)	α_2
$shelve(b)$	put box b on the nearest shelf, (cost 2)	α_3
$tag(b)$	put a tag on box b, (cost 1)	α_1, α_2

plan p_1 — $[tag(box), pushto(long\_term\_storage, box)]$
plan p_2 — $[tag(box), pickup(box), carryto(standard\_storage, box), goal(stored(box))]$
plan p_3 — $[putdown(box)]$
plan p_4 — $[putdown(box), shelve(box)]$

Fig. 3. Actions and plans in the warehouse domain

this issue in this paper, a simple strategy is for an agent to perform this reassessment when it is not otherwise occupied, or when the change in its trust of others exceeds some threshold. Although there will be some significant computation involved, it is limited in the number of capable agents, the number of plans and the the numbers of actions in those plans. Moreover, since assessment is carried out in a *pre-execution* strategy combined with periodic reassessment, the overhead placed on an agent for plan selection at run time is relatively low, especially if computation relating to plan reassessment is performed when the agent is idle.

6 Warehouse Example

To illustrate this scheme, consider the example of a warehouse domain, comprising three agents α_1, α_2, and α_3. The warehouse has three areas: a delivery area, a standard storage area, and a long term storage area, such that boxes arrive in the delivery area and must be moved to one of the storage areas. On the arrival of a box it is unknown how soon it will be needed, and whether it should be put in standard or long term storage. Boxes in the standard storage area can be kept on the floor or on shelves, with the only constraint being that a box cannot be placed on top of another, to allow easy access. Thus, if an agent wishes to store a box in the standard storage area, and the floor is full, it must be stored on the shelves. In the long term storage area boxes can be stored on the floor or on other boxes. The possible actions, the agents that are able to perform them, along with an example set of plans, are shown in Figure 3. Each action has an associated cost, shown in parentheses, corresponding to its standard rating.

The warehouse requires that once boxes arrive in the delivery area they are moved to one of the storage areas, and that the first agent to perceive a box in the delivery area should adopt the goal to move the box. Suppose that agent α_1 notices a box in the delivery area, and forms the goal of the box being placed in a storage area. Now, it has two applicable plans for this goal, p_1 and p_2. Plan p_1 is fully elaborated and can be

executed without further elaboration, while p_2 is partial and requires elaboration before it can be fully executed. There are two plans, p_3 and p_4, that can be used to elaborate p_2. Which of these plans will be used for elaboration depends on the circumstances at the time of elaboration. For example, if there is sufficient floor space in the standard storage area then p_3 can be used, but if there is no free space then p_4 must be used. We use the notation $p_{2(3)}$ to refer to p_2 when elaborated with p_3, and $p_{2(4)}$ when with p_4. Note that agent α_1 has sufficient capabilities to execute $p_{2(3)}$ by itself, but it must cooperate to execute both p_1 and $p_{2(4)}$.

The *standard rating* for the possible plans can be determined from the cost of the actions, and is equal to 5 for p_1, 7 for $p_{2(3)}$ and 9 for $p_{2(4)}$. However, since agent α_1 is unable to execute p_1 without assistance, and may be unable to execute p_2 without assistance (depending on how it is elaborated), it must consider the plan's *cooperative rating*. Suppose that α_1 has a trust value of 0.8 for both agent α_2 and α_3. The only agent capable of performing the action required for p_1, $pushto(l, b)$, is α_2, so the *cooperative rating* for plan p_1 is equal to $\frac{1}{0.8} = 1.25$. For p_2, if the agent elaborates the plan with p_3 then cooperation is not needed, so the rating is 0. However, if elaborated with p_4, the agent must cooperate with α_3, so the rating is $\frac{1}{0.8} = 1.25$.

If we suppose for simplicity that the weighting used in combining the *standard* and *cooperative* rating (i.e. w_s and w_c) are both equal to 1, then the overall rating for the plans can be determined. Since p_1 is fully elaborated the rating is simply arrived at from the formula $(w_s * S) + (w_c * C)$, i.e. $5 + 1.25 = 6.25$. The rating for p_2 depends on the ratings for each of its possible elaborations, $p_{2(3)}$ and $p_{2(4)}$. The rating for $p_{2(3)}$ is equal to $7 + 0 = 7$, and similarly $9 + 1.25 = 10.25$ for $p_{2(4)}$. Thus the best-case rating for p_2 is 7, while the mean rating is $\frac{7+10.25}{2} = 8.625$. In this example the best-case advantage is $7 - 6.25 = 0.75$, and the mean-case advantage is $8.625 - 6.25 = 2.375$. The mean-case advantage is greater so the agent should use the mean rating in plan selection. Since p_1 has the lowest mean-rating, it should be selected by the agent.

Alternatively, if α_1 has a trust value of 0.2 for agent α_2, the best-case, and mean, rating for p_1 becomes $5 + \frac{1}{0.2} = 10$. Here, the best-case advantage outweighs the mean-case advantage, so the best-case rating is used to select the best plan, in this case p_2.

7 Conclusions

In this paper we have presented a mechanism for plan selection in BDI-like agents that takes into account the inherent risk involved in cooperation. We describe how an agent can assess the risk for a given plan in the light of its knowledge of others' capabilities, and its trust in them. Plans are judged both in terms of the risk they involve and their cost according to standard criteria. However, computational constraints mean that a full analysis of plans is not possible at execution time and a pre-execution assessment is performed instead, allowing an agent to make an informed selection between plans.

The work described in this paper is part of a wider effort investigating the process of cooperation with respect to BDI-like agents. As part of this, several questions remain open with respect to the mechanisms described in this paper. Firstly, we might consider how to incorporate the notion of an agent's *rights* [13] to perform actions, both in terms of an agent not having right to perform an action and so needing to cooperate, and also

when assessing the risk involved in a plan in relation to the rights of other. Secondly, as Marsh [11] points out, an agent's trust in another is dependent on the action being considered. This would provide a richer basis for plan selection if incorporated into the assessment of plans, but at a cost of increasing the overhead of modelling others. In order to take further advantage of the dynamic multi-agent nature of the environment further exploration of these and other issues will be required. Nevertheless, in this paper we propose an effective mechanism for cooperative plan selection that moves us a step forward towards to better exploiting the potential benefits of the multi-agent domain.

Acknowledgements Thanks to Kevin Bryson and the anonymous referees for many helpful comments.

References

1. M. E. Bratman, D. Israel, and M. Pollack. Plans and resource-bounded practical reasoning. *Computational Intelligence*, 4:349–355, 1988.
2. C. Castelfranchi and R. Falcone. Principles of trust for MAS: Cognitive anatomy, social importance, and quantification. In *Proceedings of the Third International Conference on Multi-Agent Systems*, pages 72–79, 1998.
3. C. Castelfranchi, M. Miceli, and A. Cesta. Dependence relations among autonomous agents. In E. Werner and Y. Demazeau, editors, *Decentralized AI 3*, pages 215–227. Elsevier Science Publishers, 1992.
4. M. Deutsch. Cooperation and trust: Some theoretical notes. In M. R. Jones, editor, *Nebraska Symposium on Motivation*, pages 275–319. University of Nebraska Press, 1962.
5. M. d'Inverno, D. Kinny, M. Luck, and M. Wooldridge. A formal specification of dMARS. In Singh, Rao, and Wooldridge, editors, *Intelligent Agents IV*, pages 155–176. Springer, 1998.
6. E. H. Durfee. Blissful ignorance: Knowing just enough to coordinate well. In *Proceedings of the First International Conference on Multi-Agent Systems*, pages 406–413, 1995.
7. D. Gambetta. Can we trust trust? In D. Gambetta, editor, *Trust: Making and Breaking Cooperative Relations*, pages 213–237. Blackwell, 1988.
8. M. P. Georgeff and A. L. Lansky. Reactive reasoning and planning. In *Proceedings of the Sixth National Conference on Artificial Intelligence*, pages 677–682, 1987.
9. D. Kinny, M. Ljungberg, A. Rao, E. Sonenberg, G. Tidhar, and E. Werner. Planned team activity. In *Proceedings of the Forth European Workshop on Modelling Autonomous Agents in a Multi-Agent World*, pages 227–256, 1992.
10. N. Luhmann. Familiarity, confidence, trust: Problems and alternatives. In D. Gambetta, editor, *Trust: Making and Breaking Cooperative Relations*, pages 94–107. Blackwell, 1988.
11. S. Marsh. *Formalising Trust as a Computational Concept*. PhD thesis, University of Stirling, 1994.
12. S. Marsh. Trust in distributed artificial intelligence. In C. Castelfranchi and E. Werner, editors, *Artificial Social Systems*, pages 94–112. Springer, 1994.
13. T. J. Norman, C. Sierra, and N. R. Jennings. Rights and commitments in multi-agent agreements. In *Proceedings of the Third International Conference on Multi-Agent Systems*, pages 222–229, 1998.
14. S. Russell and P. Norvig. *Artificial intelligence: A modern approach*. Prentice Hall, 1995.
15. M. Wooldridge and N. R. Jennings. Formalizing the cooperative problem solving process. In *Proceedings of the Thirteenth International Workshop on Distributed Artificial Intelligence*, pages 403–417, 1994.