J. Expt. Theor. Artif. Intell. ??(200?)1–33

Taylor & Francis
Taylor & Francis Group

The theory and practice of intention reconsideration

MARTIJN SCHUT

Department of AL, Vrije Universiteit, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
e-mail: schut@cs.vu.nl

MICHAEL WOOLDRIDGE

Department of Computer Science, University of Liverpool, Liverpool L69 7ZF, UK
e-mail: mjw@csc.liv.ac.uk

SIMON PARSONS

Department of Computer and Information Science, Brooklyn College, City University of New York, 2900 Bedford Avenue, Brooklyn, NY 11210, USA
e-mail: parsons@sci.brooklyn.cuny.edu

Abstract. One of the key problems in the design of belief-desire-intention (BDI) agents is that of finding an appropriate policy for *intention reconsideration*. Crudely, the idea is that at any given time, an agent will have a number of *intentions*, relating to states of affairs that the agent has committed to bring about. An agent chooses plans that are appropriate for bringing about these intentions; if a particular plan for a given intention fails, then the agent will typically replan, to find an alternative course of action for this intention. However, a rational agent's intentions will not be static. From time-to-time, it makes sense for such an agent to *reconsider* its intentions, for example when the intention is doomed never to be realized, or else when the agent would simply profit from adopting another, more fruitful goal. This paper presents a detailed investigation of the properties of intention reconsideration. The work builds on the foundational work of Kinny and Georgeff, who investigated the properties of various intention reconsideration strategies in environments that were to varying degrees *dynamic*, i.e. subject to unanticipated change. The present paper broadly falls into two distinct parts. In the first part, the authors extend work of Kinny and Georgeff, by investigating the properties of intention reconsideration strategies in environments that are also to varying degrees *(in) accessible* and *(non-)deterministic*. They then investigate two different models of intention reconsideration. In the first model, intention reconsideration is modelled as a process of *discrete deliberation scheduling*: intention reconsideration is modelled as an action that may be performed by an agent, and so lends itself to analysis in terms of conventional decision theoretic models of optimal action. In the second, intention reconsideration is modelled as a *partially observable Markov decision*

Journal of Experimental & Theoretical Artificial Intelligence ISSN 0952–813X print/ISSN 1362–3079 online

DOI: 10.1080/0952813041233109277

process (POMDP): solving the POMDP means finding an optimal intention reconsideration policy.

1

Keywords: ▮▮▮

Received December 2003; final version accepted August 2004

1. Introduction

Computation is a valuable resource for autonomous agents that are required to act in complex environments (Russell and Norvig 1995). Such agents cannot reason indefinitely, either about which goals to achieve, or what actions to perform in furtherance of these goals (Bratman *et al*. 1988). Any implemented agent will operate under very real resource bounds—in terms of computation power, memory and the time available to make decisions. It follows that the effective control of reasoning is a key factor in the success (or otherwise) of an agent system. Research on resource-bounded decision making and the control of reasoning originated in economics and the decision sciences (Good 1971, Simon 1982); in AI, such research falls under the banner of meta-level reasoning (Russell and Wefald 1991b); and in the agent literature, it falls under work on bounded optimality (Russell and Subramanian 1995).

1.1. *Overview*

Our chosen agent architecture for this study is the belief-desire-intention (BDI) model (Georgeff and Lansky 1987, Bratman *et al*. 1988). In BDI agents, decision-making is composed of two main activities: *deliberation* (deciding *what intentions* to achieve) and *means-ends reasoning* (deciding *how* to achieve these intentions) (Bratman *et al*. 1988). Deliberation is a computationally costly process, and in order for a BDI agent to operate effectively, it should choose to deliberate only when necessary; this requires an appropriate *intention reconsideration* policy (Bratman *et al*. 1988, Kinny and Georgeff 1991, Wooldridge and Parsons 1999).

In this paper, we present a detailed investigation of two distinct issues related to BDI agents. The first issue is that of the extent to which *environmental* factors determine the need for intention reconsideration. Our work here builds on, and extends that of Kinny and Georgeff, who studied the performance of different intention reconsideration policies in environments with varying degrees of *dynamism* (Kinny and Georgeff 1991). In our work, we investigate the performance of intention reconsideration policies in environments where we vary the following parameters (cf. Russell and Norvig 1995: p. 46): *dynamism* (the rate of change of the environment, independent of the activities of the agent), *accessibility* (the extent to which an agent has access to the state of the environment) and *determinism* (the degree of predictability of the system behaviour for identical system inputs). The environmental setting we use is the TILE-WORLD (Pollack and Ringuette 1990), which was also used as the basis of Kinny and Georgeff's experiments.

The second issue we address is that of *implementing* intention reconsideration strategies. We evaluate two models of intention reconsideration, both of which appear *a priori* to have something to offer as a mechanism for understanding how to

```
Algorithm:  BDI Agent Control Loop
1.
2.   B ← B0;
3.   I ← I0;
4.   π ← null;
5.   while (true) do
6.        get next percept ρ;
7.        update B on the basis of ρ;
8.        if (reconsider(B,I)) then
9.             D ← options(B,I);
10.            I ←filter (B,D,I);
11.            if (not sound(π,I,B)) then
12.                 π ← plan(B,I);
13.            end-if
14.       end-if
15.       if (not empty(π)) then
16.            α ← hd(π);
17.            execute(α);
18.            π ← tail(π);
19.       end-if
20.  end-while
```

Figure 1. The abstract BDI agent control loop. The loop consists of continuous observation, deliberation, planning and execution. To perform optimally, the *reconsider*(...) function decides whether deliberation and planning is necessary.

ultimately, it must fix upon some subset of its desires and commit to achieving them. These chosen desires are *intentions*.

A more formal description of the control loop of a BDI agent is shown in figure 1, which is based on the BDI agent control loop presented in Rao and Georgeff (1992) and Wooldridge (2000: 38). The idea is that an agent has *beliefs B* about the world, *intentions I* to achieve and a *plan* π to achieve intentions. In lines 2–4, the beliefs, intentions and plan are initialized. The main control loop is then in lines 5–20. In lines 6–7, the agent perceives and updates its beliefs; in line 8, it decides whether to reconsider or not; in lines 9–13 the agent deliberates, by generating new options and deliberating over these; in line 12, the agent generates a plan for achieving its intentions; and in lines 15–18 an action of the current plan is executed. Because the purpose of the functions used in this loop can be easily derived from their names, we omit the actual formalizations here for reasons of space, but direct the reader to Wooldridge (2000: ch. 2).

It is necessary for a BDI agent to *reconsider* its intentions from time to time (Bratman *et al.* 1988, Kinny and Georgeff 1991, Wooldridge and Parsons 1999). One of the key properties of intentions is that they enable the agent to be goal-driven rather than event-driven, i.e. by committing to intentions the agent can pursue long-term goals. But when circumstances have changed and, for example, an intention cannot be achieved anymore, the agent would do well to drop that intention. Similarly, when opportunities arise that enable intentions that the agent currently

has not adopted, the agent should reconsider. However, because reconsideration is itself a potentially costly computational process, one would not want the agent to reconsider its intentions at every possible moment, but merely when it is necessary to reconsider, that is when the set of intentions would change were it to reconsider. The purpose of the *reconsider*(...) function as shown in figure 1 is precisely this: to deliberate when it pays to deliberate (when deliberation will lead to a change in intentions), and otherwise not to deliberate, but to act.

Developing an appropriate intention reconsideration policy—which keeps an agent committed to its intentions just as long as it would be rational to do so—is thus a critical issue in the design of any BDI agent, and it is this issue that we address in this paper.

2. Environmental factors

The starting point for our study is an investigation of the extent to which *environmental factors*, (that is to say, issues independent of the properties of the agent), can play a part in determining when an agent should reconsider its intentions. We build on the work of Kinny and Georgeff (1991), who, in a series of experiments, investigated the relative performance of intention reconsideration strategies for BDI agents in different environmental settings. The experimental framework they used involved a PRS BDI system (Georgeff and Lansky 1987) that was situated in Pollack and Ringuette's TILEWORLD domain (Pollack and Ringuette 1990). Our contribution over the results that Kinny and Georgeff obtained are: first, the investigation of accessibility and determinism factors on the agent's effectiveness (as Kinny and Georgeff only considered dynamism); second, investigating the effect of *combined* environmental factors (dynamism, accessibility and determinism) on the agent's effectiveness.

2.1. *Background*

In essence, the TILEWORLD is a grid environment on which there are agents, tiles, obstacles and holes. An agent can move up, down, left or right, and can move tiles towards holes. An obstacle is a group of immovable grid cells. Holes have to be filled up with tiles by the agent. An agent scores points by filling holes with tiles, with the aim being to score as many points as possible. The TILEWORLD is inherently *dynamic*: starting in some randomly generated world state, based on parameters set by the experimenter, it changes over time in discrete steps, with the appearance and disappearance of holes. The experimenter can set a number of TILEWORLD parameters, including: the frequency of appearance and disappearance of tiles, obstacles, and holes; the shape of distributions of scores associated with holes; and the choice between hard bounds (instantaneous) or soft bounds (slow decrease in value) for the disappearance of holes. In the TILEWORLD, holes appear randomly and exist for as long as their *life-expectancy*, unless they disappear because of the agent's actions. The interval between the appearance of successive holes is called the *hole gestation time*.

The aims of Kinny and Georgeff's investigation were to '(1) assess the feasibility of experimentally measuring agent effectiveness in a simulated environment, (2) investigate how commitment to goals contributes to effective agent behaviour

and (3) compare the properties of different strategies for reacting to change' (Kinny and Georgeff 1991: 82). The full TILEWORLD domain was considered too complex for the experiment, and the testbed was therefore simplified in several ways. First, tiles were omitted: an agent scores points simply by moving to holes. In addition, the agent was assumed to have perfect, zero-cost knowledge of the state of the world. Finally, it was assumed that agents only form correct and complete plans, and only generate plans for visiting a single hole (rather than planning multiple-hole tours).

In Kinny and Georgeff's experiments, two different types of reconsideration strategy were used: *bold* agents, which never pause to reconsider their intentions before their current plan is fully executed, and *cautious* agents, which stop to reconsider after the execution of every action. These characteristics are defined by a *degree of boldness*, which specifies the maximum number of plan steps the agent executes before reconsidering its intentions. Dynamism in the environment is represented by the *rate of world change* and is manipulated by changing the ratio of the clock rates of the TILEWORLD and the agent. The effectiveness of the agent is represented by its score (the sum of values of holes filled) divided by the maximum score it could in principle have achieved (the sum of the scores of all holes appearing in the TILEWORLD during a trial). The results of the experiments show that a cautious agent outperforms a bold agent in highly dynamic environments; intuitively, because in dynamic environments, which change frequently, it pays to reconsider intentions frequently.

In Kinny and Georgeff's investigation, as mentioned previously, the agent has perfect zero-cost knowledge of the world. In later work by Kinny, Georgeff and Hendler (Kinny *et al.* 1992) a *sensing cost* was introduced, that represents the time cost of processing sensor information. The aim of this work was to show that an optimal sensing rate exists, depending on the degree of world dynamism and the sensing cost. A model was presented that captures the trade-off between time saved by early detection of change and time wasted by too frequent sensing. Applying a cost to sensing is different from varying the accessibility of the world. Varying accessibility essentially means varying the amount of information accessible to the agent, which implicates that it does not matter how much the agent attempts to obtain information. If a cost is applied to sensing, the information is available, but for a higher price.

The aim of the work described in this section is to experimentally investigate the performance of a range of intention reconsideration policies in environments with different properties. To do this, we make use of a simulation of a single agent inhabiting the TILEWORLD environment adapted in the way described by Kinny and Georgeff and with two further modifications: (i) we omitted obstacles from the TILEWORLD; and (ii) we allowed the agent to move diagonally over the grid (in addition to moving horizontally and vertically). Omitting obstacles simplifies the problem domain without trivializing it; allowing diagonal movement is an obvious extension.

Following Kinny and Georgeff (1991), we define the *effectiveness* ϵ of an agent as the ratio of the actual score achieved by the agent to the score that could in principle have been achieved. This measurement is thus independent of randomly distributed parameters in a trial. It also avoids problems such as machine-dependency and prevention of repetition of experiments on different machines, which would occur if

the effectiveness of an agent was based on such measures as CPU-time or elapsed time (Pollack and Ringuette 1990).

There are three main environmental attributes that we vary in our experiments:

- *Dynamism*: (an integer in the range 1 to 80 denoted by γ) represents the ratio between the world clock rate and the agent clock rate (Kinny and Georgeff 1991). If $\gamma = 1$, then the world executes one cycle for every cycle executed by the agent. Larger values of γ mean that the environment is executing more cycles for every agent cycle; if $\gamma > 1$ then the information the agent has about its environment may not necessarily be up to date.
- *Accessibility*: (a real value in the range 0 to 1 denoted by α) represents the proportion of the environment that is visible to the agent. If $\alpha = 1$, then the agent can see the entire TILEWORLD, and thus has complete, perfect information about its environment; if $\alpha = 0$, then the agent can see nothing of its environment but the grid point it currently occupies. Intermediate values of α give the proportion of the TILEWORLD that can be seen; the product of α and the maximum dimension of the TILEWORLD gives the 'distance' in grid locations that the agent can see.
- *Determinism*: (an integer in the range 0 to 100 denoted by δ) represents how certain it is that an action has the expected outcome. The idea is that an agent performs actions in order to bring about certain states of affairs. However, in most realistic environments, actions are non-deterministic, in that they can have a number of possible outcomes. Thus, δ represents the probability that an action will have its intended outcome, expressed as a percentage. If $\delta = 100$, then the agent can be certain that every action it performs will have the desired effect; as $\delta \to 0$ the probability that an action will have an undesirable outcome increases. In our scenario, actions are movements that can be made by an agent, either north, south, east, west or diagonally. We model non-determinism by allowing actions to move the agent in an unintended direction—for example, in attempting to move north, the agent may actually end up moving east. This represents the situation in mobile robotics, where a robot attempting to move in some direction can never be sure that it will succeed in moving in that direction.

The experiments we conducted on environmental factors are divided into two series: the *single parameter variation* series, in which we varied one parameter per experiment; and the *combined parameters variation* series, in which we systematically varied two parameters per experiment. In the single parameter variation, we respectively minimized or maximized the parameters other than the one varied: in the dynamism experiment, we maximized accessibility ($\alpha = 1$) and maximized determinism ($\delta = 100$). In the accessibility experiment, we minimized dynamism ($\gamma = 1$) and maximized determinism ($\delta = 100$). Finally, in the determinism experiment, we minimized dynamism ($\gamma = 1$) and maximized accessibility ($\alpha = 1$).

With respect to agent properties, we varied the *replanning rate* and the *planning cost*. The replanning rate represents the *boldness* of the agent. For each experimental condition, we set the rate to 1 (the agent replans every time before performing an action—a cautious agent) and ∞ (the agent never replans while executing a plan—a bold agent). The planning cost represents the time cost of planning: the number of

Table 1. Overview of the experiment parameters.

Parameter	Value/range
World dimension	20
Hole score	10
Hole life-expectancy	[240 960]
Hole gestation time	[60 240]
Dynamism (γ)	(1.80)
Accessibility (α)	(0.1)
Determinism (δ)	(0.100)
Number of time-steps	15 000
Number of trials	50
Replanning rate	0 or ∞
Planning cost (p)	0, 1, 2 or 4

time-steps required to form a plan. For each experimental condition, we set planning cost to 0, 1, 2 and 4. In what follows, we denote planning cost by p. In table 1, we give an overview of the values of relevant parameters that we used in the experiments ($[x, y]$ denotes a uniform distribution from x to y and (x, y) denotes the range from x to y). Note that each TILEWORLD was run for 15 000 time steps, and each run was repeated 50 times.

2.2. *Results and analysis*

In this section, we present the results of our experiments. The experiments with single parameter variation resulted in the graphs shown in figure 2. The experiments with combined parameter variation resulted in the graphs shown in figures 5, 6, and 7. The graphs for the combined parameter series generalize those of the single parameter series, and so in principle it would suffice to give the graphs of the combined parameter series only. However, in the interest of clarity, we included graphs for both series.[1] We refer to a plot of effectiveness ϵ as in figure 2 as an *effectiveness curve* and to a plot of ϵ as in figures 5, 6 and 7 as an *effectiveness surface*.

2.2.1. *Single parameter variation*

Dynamism: From the results of the dynamism experiment, as plotted in figures 2a and 2b, we observe that the shapes of the effectiveness curves are similar for bold and cautious agents, but the curves themselves differ. We can explain the shape of the effectiveness curves and the differences between the curves as follows. If the dynamism of the world is at a minimum ($\gamma = 1$), then holes appear and disappear sufficiently slowly that the agent can visit each hole before it disappears, which results in a perfect score ($\epsilon = 1$) of the agent. As γ increases, then at some point, holes start to disappear before the agent has visited them, and ϵ starts to drop below 1. The effectiveness curve first declines steeply, later more gradually and eventually asymptotically approaches zero.

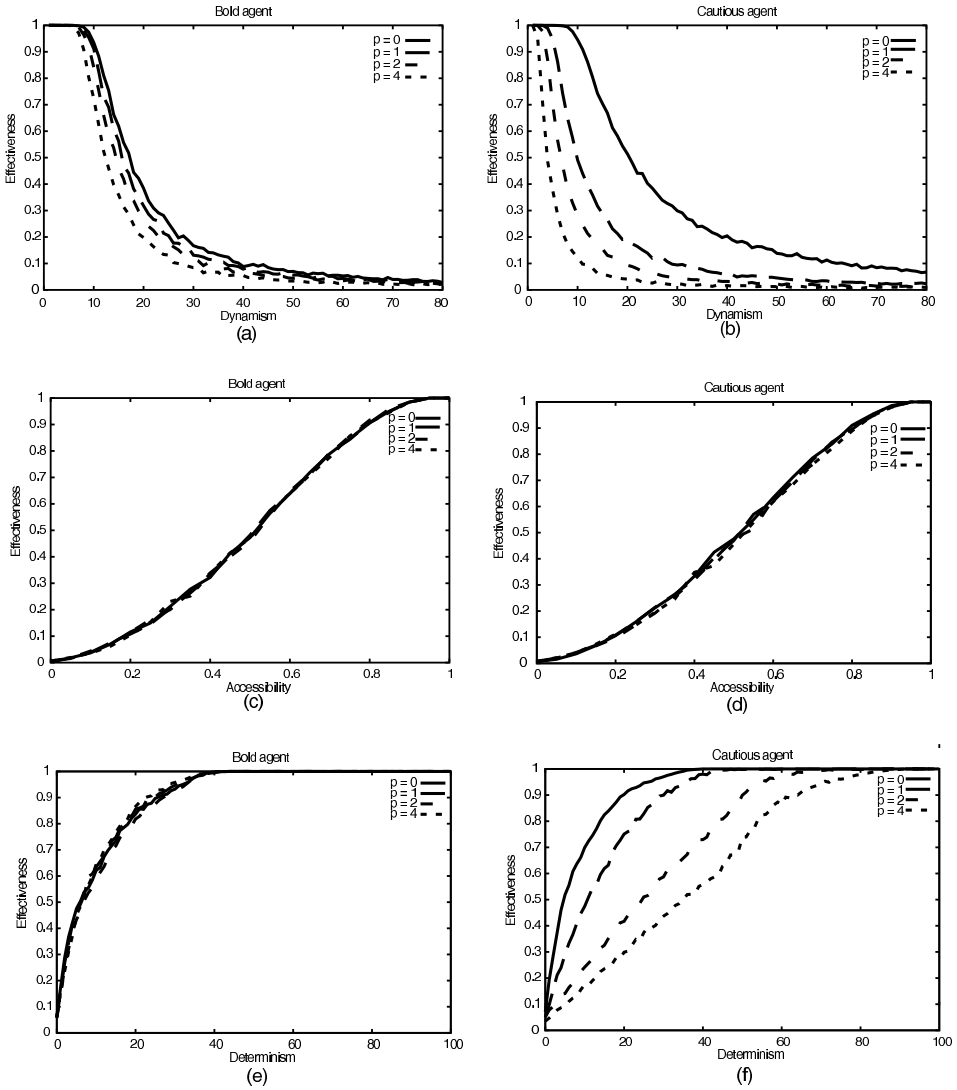


Figure 2. Experimental results (single parameter variation).

Some observations on the differences in the curves can be made directly. First, it is clear that varying the cost of planning has much more influence on the effectiveness of a cautious agent than on the effectiveness of a bold agent. Second, if planning is free ($p = 0$), then a cautious agent performs better than a bold agent if $\gamma > 7$. Third, if $p > 0$, then a cautious agent performs worse than a bold agent, independent of the dynamism of the world.

In an attempt to explain the shape of the graph in figure 2a, we used brute force computation to calculate the mean distance an agent has to travel to any hole in our TILEWORLD—as it turns out, the mean distance to any hole in our experiments is approximately 9. As previously stated, the effectiveness of an agent is the ratio of its actual score to the maximum score. This can be denoted by

$\epsilon = score_{agent}/score_{max}$. We can easily calculate the maximum score, namely $score_{max} = T/g$, where T denotes the number of time-steps and g denotes the hole gestation time. The agent's actual score can be calculated by $score_{agent} = T/f$, where f denotes the total time the agent takes to fill a hole. Similar to Kinny and Georgeff, we define f to be given by $f = d \times (p/k + m)$, where d is the hole distance, p is the planning cost, k is the reconsideration frequency and m is the time to move a single step (here always 1). If we set $k = d$, we have a bold agent, and when we set $k = 1$, we have a cautious agent. Now we can define the effectiveness of the agent as $\epsilon = g/(\gamma \times f)$. The curves in figures 2a and 2b can be approximated by this function, using the values from table 1 and a mean hole distance of 9.[2] This approximation is shown in figures 3 and 4 for a bold agent and cautious agent, respectively.

Accessibility: The shape of the effectiveness curves in figures 2c and 2d can be explained from the way we implemented the accessibility of the agent. If the accessibility is minimal ($\alpha = 0$), the agent can only see the point where it is currently located. With the exception of a hole appearing coincidentally on that location, the agent cannot score any points, and its effectiveness is minimal ($\epsilon = 0$). If the accessibility is maximal ($\alpha = 1$), the agent can see all points in the world, and has sufficient time to reach holes before they disappear, in which case its effectiveness is perfect ($\epsilon = 1$) thanks to the low value of dynamism. If $\alpha < 0.5$, then the curve is concave; if $\alpha > 0.5$, the curve is convex. This value can be explained from the fact that if $\alpha > 0.5$, it is possible for the agent to be at an optimal location, e.g. the middle of the grid, where it can see all the points in the world.

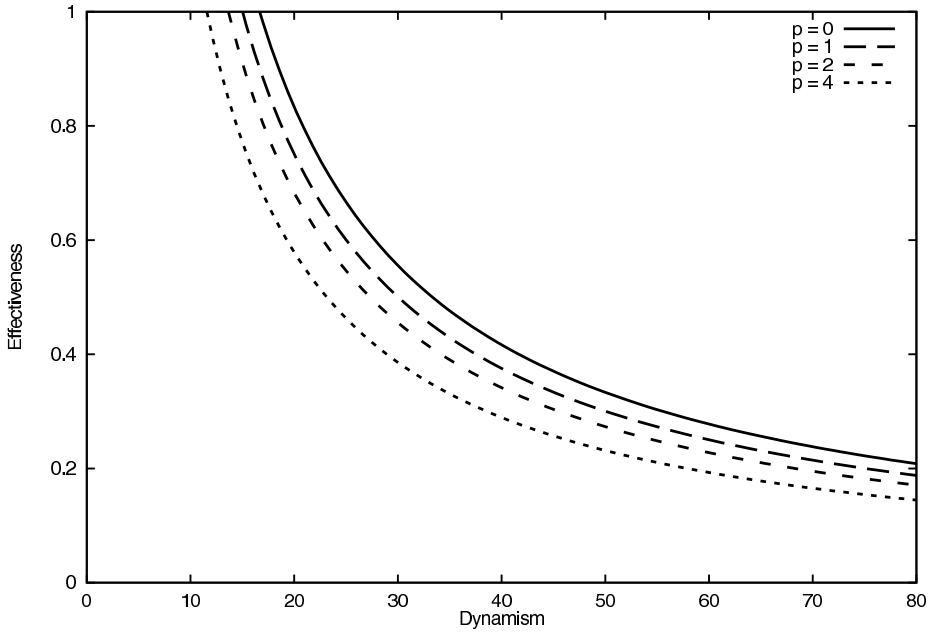


Figure 3. Theoretical effectiveness for a bold agent when dynamism is varied.

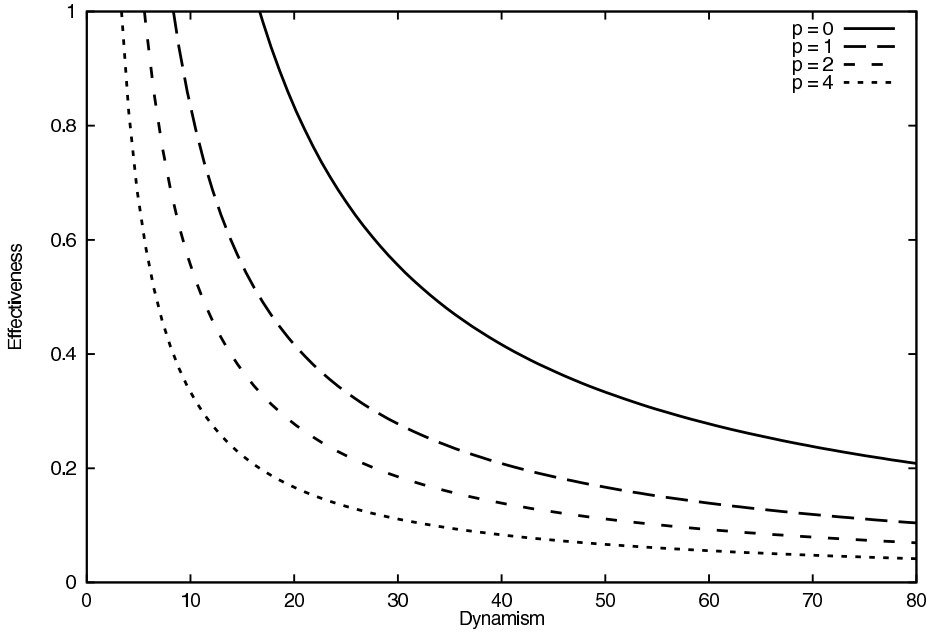


Figure 4. Theoretical effectiveness for a cautious agent when dynamism is varied.

From figures 2c and 2d it appears that there is no great difference between the results for the bold agent if planning cost is varied and between the curves for the cautious agent if planning cost is varied. Neither is there much difference between the curves for the bold agent and the curves for the cautious agent. A variance analysis on the experimental data confirms that the differences between the curves, within the bold and cautious agent effectiveness curves as well as between them, are not significant. An explanation for this might be that when accessibility is varied, the amount of deliberation an agent engages in does not influence the effectiveness of the agent. Intuitively, there is not enough information for the agent to deliberate over in order to increase its effectiveness.

Note that in addition to giving agents 'limited vision', we conducted a series of experiments in which we simulated agents with *noisy sensors*. The idea was that there would be a probability η that any given piece of information (percept) received by the agent was incorrect. If $\eta = 0$, then the agent's sensors would be perfect: all information available to the agent would be correct. If $\eta = 1$, then every piece of information available to the agent would be incorrect. We systematically varied the value of η from 0 to 1, and investigated the performance of bold and cautious agents for each, with different planning costs. These experiments yielded a linear relationship between effectiveness and η.

The shape of the graphs in figures 2c and 2d can easily be put on a theoretical footing. Because the world changes slowly enough for the agent to reach a hole when observed ($\gamma = 1$), the agent's effectiveness corresponds with its *visibility*—the number of grid points the agent can see around itself. Calculating this visibility by brute force computation resulted in a curve identical to the effectiveness curve as in 2c or 2d. Using a curve-fitting method, this visibility curve can be approximated by a biquadratic function. For example, for a 5×5 world the agent's visibility can be

described by $(-a^2+9a+5)^2/d^2$, where a denotes accessibility of the world before normalization ($a=\alpha \times \mathrm{d}$) and d denotes world dimension. The constant values in this function depend on the world dimension d.

Determinism: The effectiveness curves for the determinism experiment are plotted in figures 2e and 2f. If the determinism of the world is minimal ($\delta=0$), the outcomes of the agent's actions are never as intended by the agent. But because the agent can still encounter a hole by accident, it achieves a higher score than minimal ($\epsilon>0$). If determinism is maximal ($\delta=1$), the outcomes of the agent's actions are always the outcomes as intended by the agent, and the agent achieves a perfect score ($\epsilon=1$) again because we have set dynamism to its minimum value. The reason for this is that determinism is defined as the chance that the outcome of an agent's action is the outcome intended by the agent. If $\delta=0$, the agent never arrives at the location it intends. If $\delta=1$, the agent always arrives at the intended location. As δ increases, the agent slowly starts to arrive at the intended holes and thus increases its score. The curve inclines slowly at first and later steeper, until $\delta>40$, from where the effectiveness stays approximately perfect ($\epsilon\simeq 1$). We speculate that the agent can achieve a perfect score when $\delta>40$ for the following reason. If δ exceeds a certain threshold (here: $\delta>40$), the agent can compensate for failed actions by replanning. As long as the intended hole does not disappear, the agent can replan and in the end will reach the hole. This means an increase in deliberation, but a justified one, because it increases the effectiveness of the agent considerably.

When one considers the effectiveness curves for a bold agent, it is clear there is not much difference between them. As the planning cost p is increased, ϵ decreases. This decline is slight because the agent must replan completely after executing a plan, rather than because the agent does not need to reconsider its plans. This is also the reason why, with the exception of when planning is free ($p=0$), a bold agent performs better than a cautious agent. A cautious agent has to replan after every step, whereas a bold agent does not do this and therefore a bold agent can perform more effectively. However, when planning is free, the cautious agent outperforms the bold agent, because it does not need to execute its complete plan before replanning. In this case, a cautious agent's plans are more flexible and thus shorter. With reference to figure 2f, it is immediately obvious that planning cost has a significant impact on effectiveness for cautious agents in non-deterministic environments.

Before we leave this section, we note that the effectiveness of the agent depends on other characteristics of the environment, such as the life-expectancy of holes. If the life-expectancy of a hole is too short, then the agent cannot reach the hole by planning again. In this case, δ must be very high in order for the agent to score any points. On the other hand, if holes never disappear, the agent would achieve a perfect score, even when δ is very low.

2.2.2. *Combined parameter variation*

The experimental results from the combined parameter variation of dynamism and accessibility are shown in figure 5, of accessibility and determinism in figure 6 and of dynamism and determinism in figure 7. All effectiveness surfaces are consistent with the effectiveness curves individually. For example, in the variation of dynamism and accessibility, if dynamism is minimal ($\gamma=1$), the curve corresponds to the individual

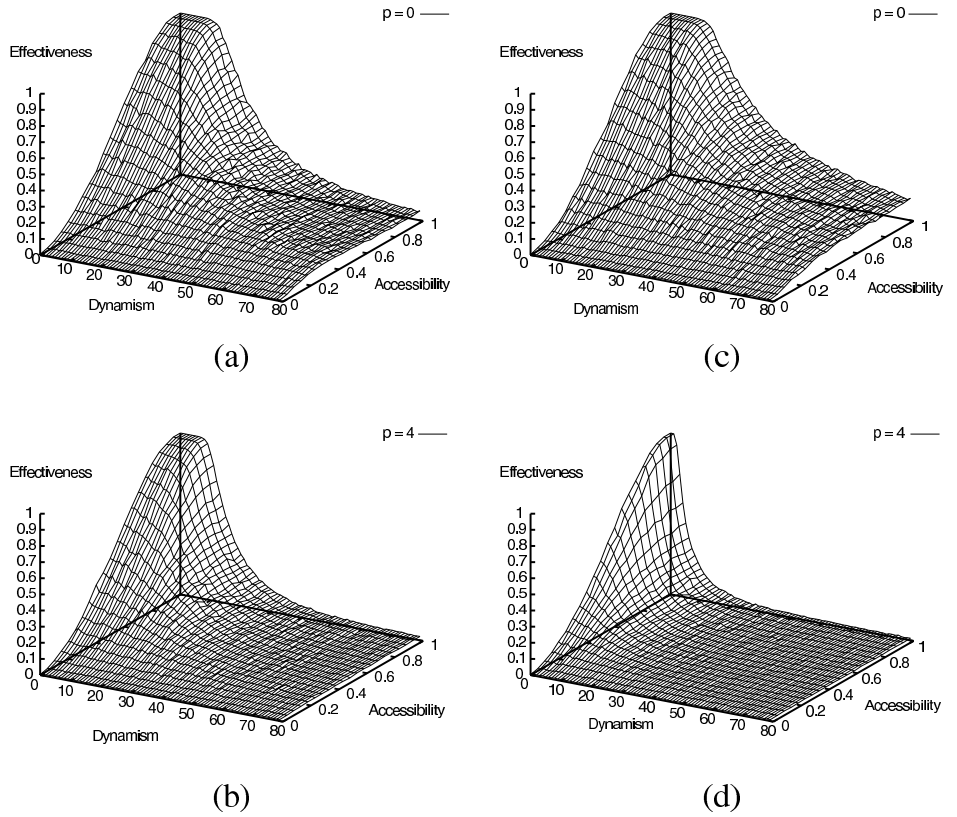


Figure 5. Dynamism and accessibility: the results for a bold agent are in (a) and (b), for a cautious agent in (c) and (d).

effectiveness curve for accessibility and if accessibility is maximal ($\alpha = 1$), the curve corresponds to the individual effectiveness curve for dynamism. For these values, the analysis is thus similar to the analysis for the single parameter variations.

With the combined parameter variation experiments we want to show which parameters dominate in complex environments. It is clear from the effectiveness surfaces in figure 5 that dynamism has more influence on the effectiveness of the agent than accessibility. This follows from the fact that the surfaces change more rapidly over the dynamism axis than over the accessibility axis. From maximal effectiveness ($\epsilon = 1$), where dynamism is minimal ($\gamma = 1$) and accessibility is maximal ($\alpha = 1$), the decline in effectiveness is much steeper when dynamism increases than when accessibility decreases.

It is clear from figure 6 that accessibility has more influence on the effectiveness of the agent than determinism. Even in a worst case scenario—a cautious agent where planning cost is 4—the decrease in effectiveness from maximal effectiveness ($\epsilon = 1$), where accessibility is maximal ($\alpha = 1$) and determinism is maximal ($\delta = 100$), is steeper when accessibility decreases than when determinism decreases. In other cases, effectiveness stays maximal until determinism is approximately 40 ($\delta = 40$). We explained the reason for this in section 2.2.1: the agent can compensate for non-determinism in the environment by replanning.

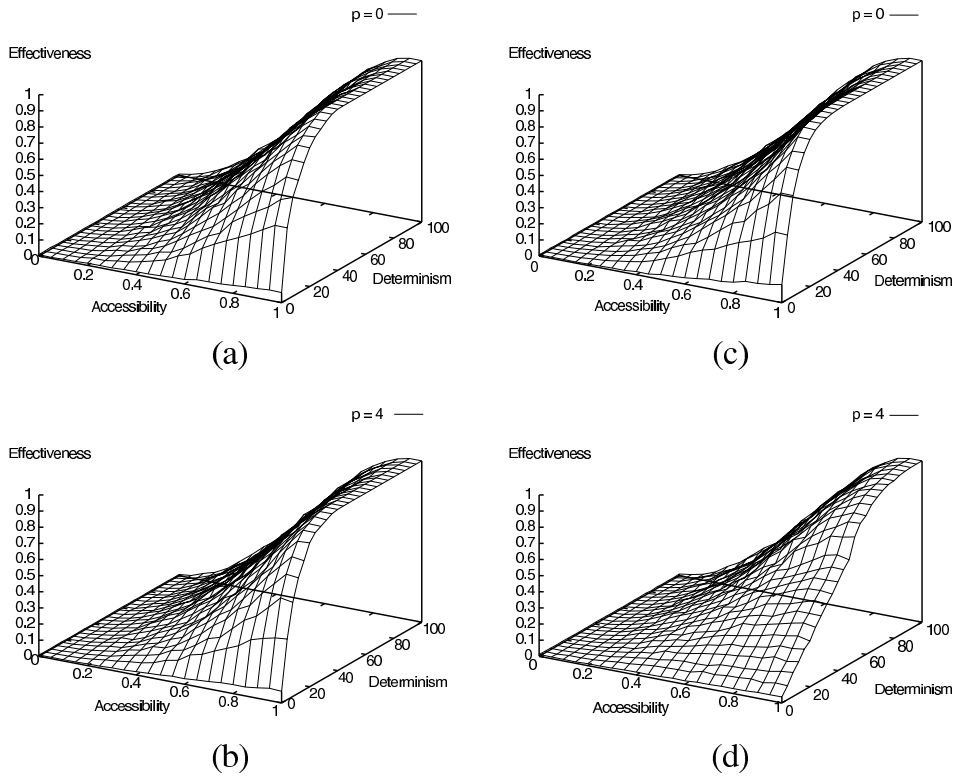


Figure 6. Accessibility and determinism: the results for a bold agent are in (a) and (b), for a cautious agent in (c) and (d).

Figure 7 shows that the agent's effectiveness changes faster over the dynamism axis than over the determinism axis, from which we conclude that dynamism has more influence on the effectiveness of the agent than determinism.

2.3. *Summary*

The experiments presented in this Section show that, for the modified TILEWORLD scenario, accessibility does not influence the effectiveness of an agent with respect to its reconsideration policy and planning cost; that determinism influences the replanning rate rather than the reconsideration rate; and that dynamism influences an agent's effectiveness the most.

Despite the fact that these results are only strictly valid for the TILEWORLD, we believe that they will extend to different domains. Indeed, we believe that the results will broadly carry over to any domain which is dynamic, non-deterministic, and have limited accessibility (all of which are features of many real world scenarios). If this estimate proves too optimistic, it certainly seems likely that similar results would be obtained for scenarios that are similar, including:

- RoboCup soccer, where agents have to move to the ball (which may be kicked away) or to block other players (who may move);
- Pursuer/Evader, where agents have to move towards, or away from, other agents (which are moving similarly); and

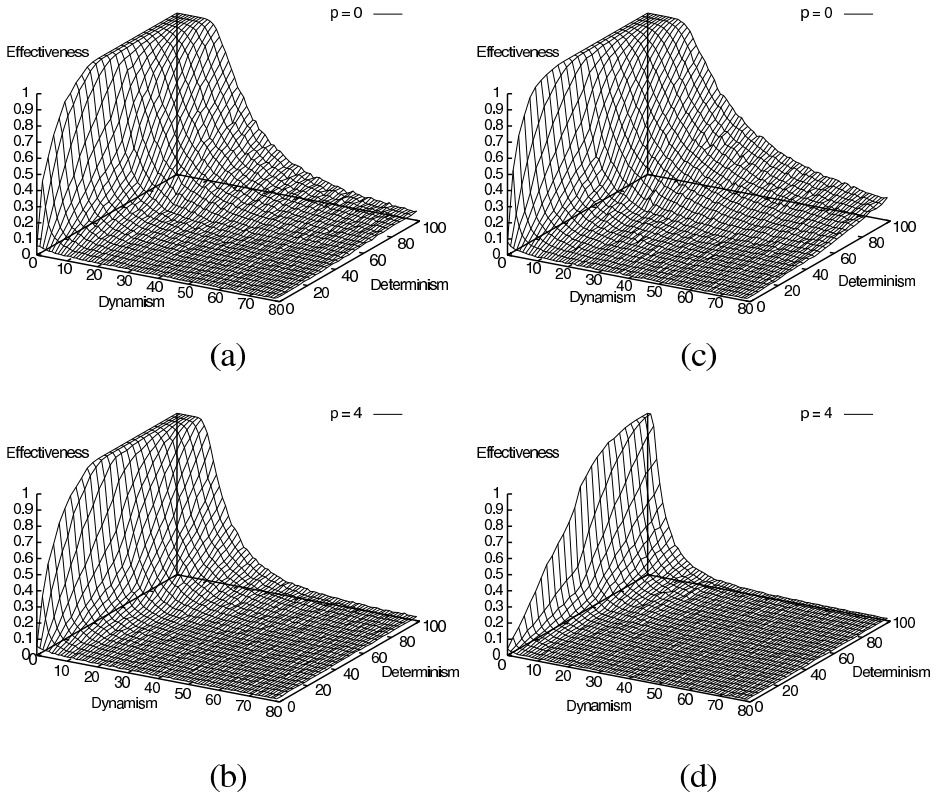


Figure 7. Dynamism and determinism: the results for a bold agent are in (a) and (b), for a cautious agent in (c) and (d).

- RoboCup rescue, where agents move through a simulated disaster zone rescuing victims (who move) and responding to changes in the environment.

In all of these scenarios, the environment is dynamic and non-deterministic, accessibility is limited and movement around the physical world is a major consideration.

2.4. *Towards more flexible intention reconsideration*

The bold and cautious strategies considered above are hard-wired into agents at compile-time. However, the issue of reconsideration is ideally one that an agent can manage autonomously. Towards this end, in the remainder of this article, we present two models that enable an agent to do precisely this. These models provide two flavours of adaptive agent that can autonomously manage their intentions.

- In the first, the notion that intention reconsideration is a form of meta-level reasoning—reasoning about how to reason—is taken seriously. We model intention reconsideration as *discrete deliberation scheduling*, using the ideas of Russell and Wefald (1991b). The key idea that informs this approach is that *deliberation is action*, and it is thus possible to determine an optimal deliberation action using the same mathematical techniques as decision theory uses to determine an optimal 'regular' action.

- In the second approach, we model intention reconsideration as a *partially observable Markov decision process* (POMDP) (Kaelbling *et al.* 1998). In this model, solving the POMDP, i.e. determining an optimal policy for the POMDP, means determining an optimal intention reconsideration policy.

For each of these models, we first describe and motivate the approach, and then give an evaluation in terms of the TILEWORLD scenario given above.

3. Discrete deliberation scheduling

Russell and Wefald (1991a) describe how an agent should schedule deliberation and action to achieve efficient behaviour. Their framework is known as *discrete deliberation scheduling*.[3] The key idea is that *deliberations are treated as if they were actions*. Decision theory gives us various models of how to determine the best possible action, of which the maximum expected utility model is perhaps the best known. Viewing deliberations as actions allows us to compute the utility of a deliberation action, and so makes it possible to apply the expected utility model as the meta-level reasoning component over all possible actions and deliberations. However, it is not difficult to see that this can be computationally hard. Russell and Wefald propose the following strategy in order to overcome this problem. Assume that at any moment in time the agent has some default action it can perform. The agent can either execute this action or deliberate, where deliberation can lead to a better action than the current default action. Their control algorithm then states that as long as there exist deliberations with a positive value, perform the deliberation with the highest value; otherwise, execute the default action.

We discuss the integration of the decision theoretic model for deliberation scheduling from Russell and Wefald and the BDI agent architecture.

3.1. *The DSS model*

Here we give a brief overview of the model of discrete deliberation scheduling. We explain the basic elements, present the control algorithm that uses these elements and discuss some additional assumptions required to make the algorithm computationally attractive. In the next Section we show how the model can be applied to the TILEWORLD.

The most important issue we are concerned with relates to the set of available actions A of the agent: we distinguish between *external* actions $A_{ext} = \{a, a', a'', \ldots\}$, affecting the agent's environment, and *internal* actions $A_{int} = \{d, d', \ldots\}$, affecting the internal state of the agent. We let $A = A_{ext} \cup A_{int}$ and assume $A_{int} \cap A_{ext} = \emptyset$. We assume the agent's environment, (i.e. everything external to the agent), may be in any of a set $E = \{e, e', e'', \ldots\}$ of environment states. We let *utility* be defined over environment states: $U_{\mathbf{e}}: E \rightarrow \mathbb{R}$. If the agent uses maximum expected utility theory (MEU) as a decision strategy, it chooses an action $a_{meu} \in A_{ext}$ for which the utility of the outcome state is maximal:

$$a_{meu} = \arg\max_{a \in A_{ext}} \sum_{e \in E} P(e \mid a) U_{\mathbf{e}}(e) \tag{1}$$

where $P(e \mid a)$ denotes the probability of state e occurring, given that the agent chooses to perform external action a. However intuitive this notion of decision

making is, many problems arise when MEU is used in the real world. It assumes $U_e(E)$ is known before deciding, that enough time is available to obtain a_{meu}, and it does not easily extend to *sequential* decision making.

Russell and Wefald (1991a) offer an alternative. The idea underlying their model is that the agent chooses between: (1) a default external action a_{def}, and (2) an internal action from the set of internal actions—at any moment, the agent selects an action from $\{a_{def}, d, d', \ldots\}$. The only purpose of an internal action is to revise the default external action, presumably to a better one. This algorithm does not guarantee an optimal choice, but computationally it can be a lot more attractive than MEU. In the remainder of this section, we outline the theory of discrete deliberation scheduling. For the formal details, we direct the interested reader to Russell and Wefald (1991a) and Schut and Wooldridge (2001).

The theory enables one to express utilities of deliberations and actions. Like a_{meu} was presented above for external actions, in discrete deliberation scheduling it is possible to make similar computations for internal actions (deliberations). This idea is exploited in the *decision control algorithm* (DCA), which ensures that an agent deliberates when there is a deliberation action with positive utility, and performs an external action otherwise.

Computing the utilities of deliberations is not trivial, mainly because of their reflective character. Performing a deliberation implies reasoning over one's own available actions to execute and this requires complicated reflective capabilities on the agent's side. Russell and Wefald tackle this problem pragmatically by subsequently demonstrating: first, how to compute utilities of external and internal actions; second, how to estimate the utilities of internal actions; and, finally, how to represent temporal constraints on the utilities of internal actions.

First, utilities for external actions are computed simply based on the expected utilities as shown in equation (1). The utility of an internal action is the utility of the external action that the internal action eventually leads to. Russell and Wefald refer to an internal action that immediately results in an external action as a *complete computation*, and to one that does not necessarily do so as a *partial computation*.

Second, an agent will need to estimate utilities if it has no immediate access to its utility function. This can be done by making utilities depend on sequences of actions undertaken to get an appropriate utility estimate. Such sequences can in practice be statistical knowledge collected from past situations.

Third, in practice, deliberation takes time. This can be captured by defining a cost function over internal actions. Such a function then denotes the difference between the intrinsic (time-independent) and total (time-dependent) utility of an internal action. This cost function can represent a negative effect (best actions should not be postponed) or a positive effect (longer deliberation leads to better external actions).

Finally, in the most advanced model (including estimation, cost functions, etc.) it is still not feasible in practice to assess the expected value of all continuations of a computation, because computations can be arbitrarily long. Russell and Wefald make two simplifying myopic assumptions: first, algorithms are *meta-greedy*, in that they consider single primitive steps, estimate their ultimate effect and choose the step appearing to have the highest immediate benefit; and, second the computation value of a complete computation is a useful approximation to its true value as a possibly partial computation (called the *single-step assumption*).

3.2. *Integrating BDI and DSS*

Having now defined both the BDI model and discrete deliberation scheduling, we discuss how the models can be integrated. The agent's control loop of our framework is the BDI agent control loop as shown in figure 1. As mentioned above, integrating the frameworks comes down to implementing the *reconsider*(...) function in this control loop. This implementation is shown in figure 8; it is based on Russell and Wefald's meta-reasoning model. The function *computeUtility*(...) computes the estimated utility of deliberation. The argument of this function is the agent's set of beliefs. These beliefs typically include the values of the necessary distributions for computing the estimates, e.g. the dynamism of the environment.

Because we use the BDI model, we treat deliberation on a very abstract level: we merely recognise deliberation as a way to alter the set of intentions. Therefore, we are only concerned with a single internal action: deliberation itself. The *reconsider*(...) function then decides whether to deliberate (indicated by *reconsider*(...) evaluating to 'true'), or act (*reconsider*(...) evaluates to 'false'). We can regard choosing to act as the default action a_{def} and choosing to deliberate as the single internal action. It is clear that this relates Russell and Wefald's model to the BDI model. We are left with two questions: what should the default action a_{def} be and how do we compute the utilities of choosing to deliberate versus choosing to act? We deal with these questions subsequently.

Let Π be the set of all plans. A plan is a recipe for achieving an intention; $\pi \in \Pi$ represents a plan, consisting of actions $\pi[0]$ through $\pi[n]$, where $\pi[i] \in A_{ext}$ and n denotes the length of the plan. The agent's means-ends reasoning is represented by the function *plan*: $\wp(B) \times \wp(I) \rightarrow \Pi$, used on line 12 in figure 1. At any moment in time, we let the default action a_{def} be $\pi[0]$, where the computation of the utility of a_{def} is derived via equation (1). This answers the first question.

The computation of the utility of deliberation is done using Russell and Wefald's model: we estimate the utility of deliberation, based on distributions which determine how the environment changes. These distributions are necessary knowledge because the optimality of intention reconsideration depends *only* on events that happen in the environment. For now, we assume that the agent knows these distributions and that they are *static* (they do not change throughout the existence

```
Function:  boolean reconsider(B, I)
1.
2.   get current plan π from I;
3.   a_def ← π[0];
4.
5.   U_a(a_def) ← Σ_{e∈E} P(e | a_def) U_e(e);
6.   Û_a(d) ← computeUtility(B);
7.
8.   if ((Û_a(d) ← U_a(a_def)) > 0) then
9.       return true;
10.  end-if
11.  return false;
```

Figure 8. The *reconsider*(...) function in the BDI agent control loop. It computes and compares the utilities of acting and deliberating, and decides, based on the outcome of this comparison, whether to deliberate or not.

of the environment) and *quantitative*. We estimate the utility of deliberation as the difference between the utility of the outcome of the deliberation (i.e. a revised $\pi[0]$), and a_{def} (i.e. the current $\pi[0]$). Situated in a real-time environment, the agent will discount the estimated utility of deliberation, based on the length of deliberating. The decision control algorithm DCA then prescribes to deliberate and execute the revised $\pi[0]$ if this estimate is positive, and to act—execute the current $\pi[0]$—otherwise.

This results in a meta level control function *reconsider*(...) which enables the agent at any time to compute the utility of $\pi[0]$ and also to estimate the utility of deliberating over its intentions, and then, according to these utilities, acts (by executing $\pi[0]$) or deliberates (by reconsidering its intentions). Next, we illustrate the theory with the same scenario that we introduced above.

3.3. *The* TILEWORLD

For our work in this section we need a formal model of the TILEWORLD. Let H represent the set of possible holes; an environment state is an element from the set $E = \wp(H)$ with members e, e', e'', ... We let $A_{ext} = \{noop, ne, e, se, s, sw, w, nw, n\}$, where each action denotes the direction to move next and the *noop* is a null action; if the agent executes *noop*, it stays still. The agent's only internal action is to deliberate, thus $A_{int} = \{d\}$. At any given time, if holes exist in the world, an agent has a single intended hole `IH`—the hole it is heading for—over which it is deliberating. If no holes exist, the agent stays still. Let $\texttt{dist}_h$ denote the distance between the agent and hole $h \in H$. Then $\texttt{mindist} = \min\{dist_h \mid h \in H\}$ denotes the distance to the hole closest to the agent. The agent's deliberation function d selects `IH`, based on `mindist`; the means-ends reasoning function *plan* selects a plan π to get from the agent's current location to `IH`. For example, if the agent is currently at location $(2, 0)$ and `IH` is at $(1, 3)$, then $\pi = [s; s; sw]$. We assume that d and *plan* are optimal, in that d selects the closest hole and *plan* selects the fastest route.

According to our model, the agent must at any time choose between executing action $\pi[0]$ and deliberating. Based on the utilities of these actions, the *reconsider*(...) function decides whether to act or to deliberate. Let the utility of an environment state be the inverse of the distance from the agent to its intended hole, $n(\texttt{dist}_{\texttt{IH}})$, where n is an order-reversing mapping.[4] Equation (1) then defines the utility of an external action. While in this domain, the utility of an external action is immediately known, the utility of internal actions is not immediately known, and must be estimated. In accordance with our model, we use pre-defined distributions here: the utility of an internal action is estimated using knowledge of the distribution of the appearance and disappearance of holes.[5] The reason for this is that the appearance and the disappearance of holes are events that cause the agent to change its intentions. For example, when the set of holes H does not change while executing a plan, there is no need to deliberate; but when H does change, this might mean that `IH` has disappeared or that a closer hole has appeared: reconsideration is necessary. Let `avedist` be the average distance from the agent to every location on the grid; this is a trivial computation. Let `newholes` be the estimated number of holes that have appeared since the last deliberation; this is calculated using the dynamism of the world and the gestation period of holes—the gestation period is the elapsed time in between two successively appearing holes. We deem `avedist/newholes` an appropriate estimate for the utility of deliberation.

```
Function:  boolean reconsider(B, I)
1.
2.  get dist_IH from B;
3.  get avedist from B;
4.  get newholes from B;
5.  get current plan π from I;
6.  a_def ← π[0];
7.
8.  U_a(a_def) ← n(dist_IH);
9.  Û_a(d) ← n(avedist/newholes);
10.
11. if ((Û_a(d) − U_a(a_def)) > 0) then
12.     return true;
13. end-if
14. return false;
```

Figure 9. The *reconsider*(...) function for the TILEWORLD.

The *reconsider*(...) function for TILEWORLD agents is shown in figure 9. We let the belief set of the agent at least consist of

$$B = \{\texttt{dist}_{\texttt{IH}}, \texttt{avedist}, \texttt{newholes}\}$$

and let the intention set be

$$I = \{\texttt{IH}\}$$

The *reconsider*(...) function computes the utility of executing π[0] and estimates the utility of deliberating: if

$$\texttt{dist}_{\texttt{IH}} < \frac{\texttt{avedist}}{\texttt{newholes}}$$

the agent acts, and if not, it deliberates.

As mentioned above, this does not guarantee optimal behaviour, but it enables the agent to determine its commitment to a plan autonomously. We empirically evaluate our framework in the next section, and demonstrate an agent using such an intention reconsideration scheme performs better than when a level of commitment is hardwired into the agent.

3.4. *Experimental results*

In this section, we present a series of simulations in which we utilize the TILEWORLD environment—as described above—inhabited by a single agent. The experiments are based on the same methodology as described above in section 2. In section 2, the performance of a range of intention reconsideration policies were investigated in environments of different structure. Here we carry out broadly the same set of experiments but, in addition to the bold and cautious agents studied before, we introduce an *adaptive* agent, which figures out for itself how committed to its plans it should be. The decision mechanism of this agent is based on the theory as described in section 3.1.

Table 2. Overview of the experiment parameters

Parameter	Value/range
World dimension	20
Hole score	10
Hole life-expectancy	[240 960]
Hole gestation time	[60 240]
Dynamism (γ)	(1.80)
Accessibility	20
Determinism	100
Number of time-steps	15 000
Number of trials	25
Planning cost (p)	0, 1, 2 or 4

2

We measured three dependent variables: the *effectiveness* ϵ, the *commitment* β and the *cost of acting* c of of an agent using the discrete deliberation scheduling approach to intention reconsideration. Effectiveness, as before, is the ratio of the actual score achieved by the agent to the score that could in principle have been achieved. Commitment is calculated as how many actions of a plan are executed before the agent replans as a fraction of the full plan. Thus commitment for a plan π with length n is $(k-1)/(n-1)$, where k is the number of executed actions. Observe that commitment defines a spectrum from a cautious agent ($\beta=0$, because $k=1$) to a bold one ($\beta=1$, because $k=n$). The cost of acting is the total number of actions the agent executes. While cost of acting can easily be factored into the agent's effectiveness, we decided to measure it separately in order to maintain clear comparability with previous results.

In table 2, we summarize the values of the experimental parameters ($[x, y]$ denotes a uniform distribution from x to y and (x, y) denotes the range from x to y).

3.4.1. *Results*

The experiments for dynamism resulted in the graphs shown in figure 10. In figure 11a we plotted commitment β of an adaptive agent, varying dynamism, with a planning cost p of 0, 1, 2 and 4, respectively. The collected data was smoothed using a Bezier curve in order to get these commitment graphs, because the commitment data showed heavy variation resulting from the way dynamism is implemented. Dynamism represents the acting ratio between the world and the agent; this ratio oscillates with the random distribution for hole appearances, on which the adaptive agent bases its commitment. The commitment of a cautious and bold agent are of course constantly 0 and 1 respectively. In figure 11b, the cost of acting c is plotted for the three agents for $p=4$. The cost of acting represents the number of time steps that the agent performed an action.

3.4.2. *Analysis*

For bold and cautious agents, we obtained the same results as from the series of experiments described above. When planning is free ($p=0$) as in figure 10a, it was shown in the experiments in section 2 that a bold agent outperforms a cautious agent. This out-performance, however, was negligible in a very dynamic environment. In these experiments, it is very clear that in a static world (where dynamism

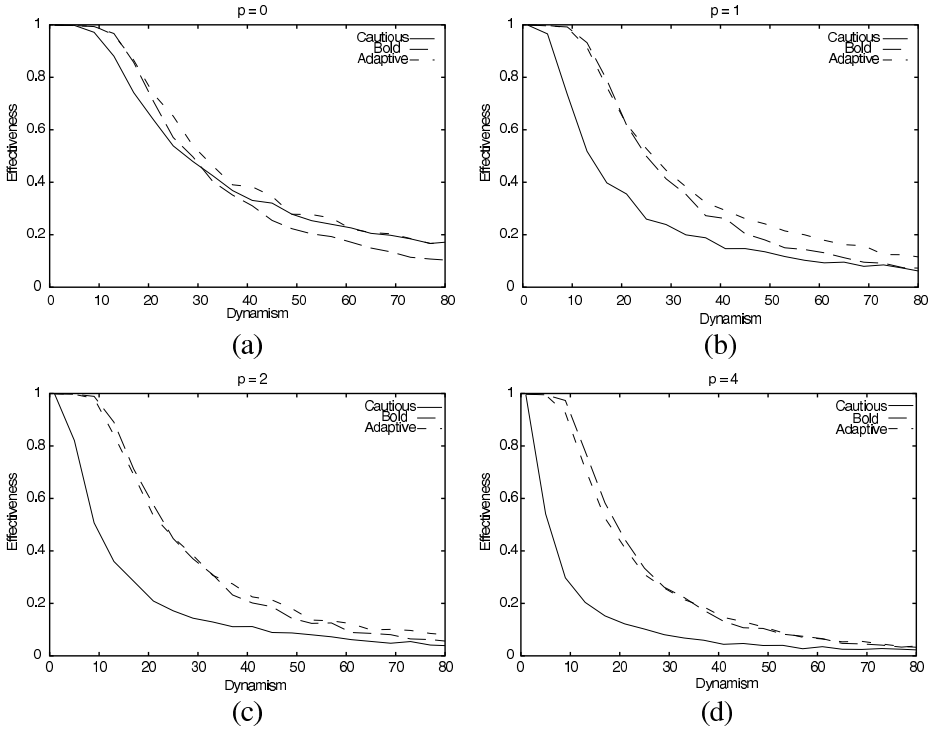


Figure 10. Performance of a cautious, bold and adaptive agent. Effectiveness is measured as a result of a varying degree of dynamism of the world. The four panels represent the effectiveness at different planning costs (denoted by p), ranging from 0 to 4.

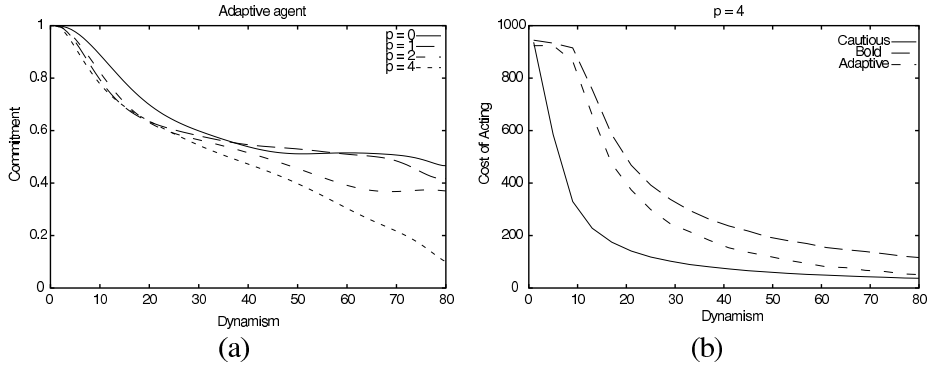


Figure 11. Commitment for an adaptive agent and cost of acting for a cautious, bold and adaptive agent. In (a), the commitment level is plotted as a function of the dynamism of the world for an adaptive agent with planning cost (denoted by p) of 0, 1, 2 and 4. In (b), the cost of acting—the number of time steps that the agent moves—is plotted as a function of the dynamism of the world for a cautious, bold and adaptive agent with a planning cost of 4.

is low), a bold agent indeed outperforms a cautious agent. But from some point onwards (dynamism is approximately 28), a cautious agent outperforms a bold one. This observation agrees with the natural intuition that it is better to stick with a plan as long as possible if the environment is not very likely to change much, and to drop it quickly if the environment changes frequently. More importantly, when planning is free, the adaptive agent outperforms the other two agents, independent of the dynamism of the world. This means that adaptive agents indeed outsmart bold and cautious agents when planning is free.

As planning cost increases, the adaptive agent's effectiveness gets very close to the bold agent's effectiveness. However, there is more to this: when we take the cost of acting into account, we observe that the adaptive agent's acting cost is much lower. Considering these costs, we can safely state that the adaptive agent keeps outperforming the bold agent. When planning is expensive ($p=4$) as in figure 10d, the cautious agent suffers the most from this increase in planning cost. This is because it only executes one step of its current plan and after that, it immediately plans again. It thus constructs plans the most often of our types of agents. We also observe that the bold agent and adaptive agent achieve a similar effectiveness. But again, as shown in figure 10d, the adaptive agent's acting costs are much lower.

We included the level of commitment for an adaptive agent, as shown in figure 10d, to demonstrate how commitment is related to the dynamism of the world. Some interesting observations can be made here. Firstly, we see that planning cost has a negative influence on commitment—as planning cost increases, the level of commitment decreases. The reason for this is that the cost of planning is the time it takes to plan; as this value increases, more events can take place in the world during the planning period, and it becomes more attractive to replan earlier rather than later. Secondly, we see that if dynamism increases, the level of commitment decreases. This can be easily explained from the intuition, as described above, that in a very fast changing world, it is better to reconsider more often in order to be effective.

4. Markov Decision Processes

In this section, we show how intention reconsideration may be modelled using the theory of Markov decision processes for planning in partially observable stochastic domains. We view an intention reconsideration strategy as a policy in a partially observable Markov decision process (POMDP): solving the POMDP thus means finding an optimal intention reconsideration strategy. We have shown above that an agent's optimal rate of reconsideration depends on the environment's dynamism, determinism, and observability. The motivation for using a POMDP approach here is that in the POMDP framework the optimality of a policy is largely based on exactly these three environmental characteristics.

Let P be a set of *propositions* denoting environment variables. In accordance with similar proposition-based vector descriptions of states, we let environment states be built up of such propositions. Then E is a set of *environment states* with members $\{e, e', \ldots\}$, and $e=\{p_1, \ldots, p_n\}$, where $p_i \in P$.

The internal state of an agent consists of beliefs and intentions. Let Bel: $E \rightarrow [0,1]$, where $\sum_{e \in E} Bel(e)=1$, denote the agent's *beliefs*: we represent what the agent believes to be true of its environment by defining a probability distribution over the possible environment states. The agent's set of *intentions*, Int, is a subset of the set of environment variables: $Int \subseteq P$. An internal state s is a pair $s=\langle Bel, Int\rangle$,

where Bel: $E \rightarrow [0,1]$ is a probability function and $Int \subseteq P$ is a set of intentions. Let S be the set of all internal states. For a state $s \in S$, we refer to the beliefs in that state as Bel_s and to the intentions as Int_s. We assume that it is possible to denote values and costs of the outcomes of intentions: an *intention value* V: $Int \rightarrow \mathbb{R}$ represents the value of the outcome of an intention; and *intention cost* C: $Int \rightarrow \mathbb{R}$ represents the cost of achieving the outcome of an intention.[6] The *net value* V_{net}: $Int \rightarrow \mathbb{R}$ represents the net value of the outcome of an intention; $V_{net}(i)$, where $i \in Int$, is typically $V(i) - C(i)$. We can express how 'good' it is to be in some state by assigning a numerical value to states, called the *worth* of a state. We denote the worth of a state by a function W: $S \rightarrow \mathbb{R}$, and we assume this to be based on the net values of the outcomes of the intentions in a state. For example, for a state s containing a single intention i, then $W(s) = V_{net}(i)$. Moreover, we assume that one state has a higher worth than another state if the net values of all its intentions are higher. This means that if $\forall s,\ s' \in S,\ \forall i \in Int_s,\ \forall i' \in Int_{s'},\ V_{net}(i) \geq V_{net}(i')$, then $W(s) \geq W(s')$. In the empirical investigation discussed below, we illustrate that a conversion from intention values to state worths is feasible, though we do not explore the issue here.[7] Finally, Ac denotes the set of physical actions the agent is able to perform; with every $\alpha \in Ac$ we identify a set of propositions $P_\alpha \subseteq P$, which includes the propositions that change value when α is executed.

In the remainder of this paragraph, we explain what a POMDP is; in the next Section, we explain how to use it for implementing intention reconsideration. A POMDP can be understood as a system that at any point in time can be in any one of a number of distinct states, in which the system's state changes over time resulting from actions, and where the current state of the system cannot be determined with complete certainty (Boutilier *et al.* 1999). In our case, the partial observability arises when environments are not completely accessible to the agent, in which case it cannot distinguish between states which vary only in details that it cannot observe. Partially observable MDPs satisfy the Markov assumption so that knowledge of the current state renders information about the past irrelevant to making predictions about the future. In a POMDP, we represent the fact that the knowledge of the agent is not complete by defining a probability distribution over all possible states. An agent then updates this distribution when it observes its environment.

Let a set of states be denoted by S and let this set correspond to the set of the agent's internal states as defined above. This means that a state in the MDP represents an internal state of the agent. We let the set of actions be denoted by A. (We later show that $A \neq Ac$ in our model.) An agent might not have complete knowledge of its environment, and must thus *observe* its surroundings in order to acquire knowledge: let Ω be a finite set of observations that the agent can make of the environment. We introduce an *observation function* O: $S \times A \rightarrow \Pi(\Omega)$ that defines a probability distribution over the set of observations; this function represents what observations an agent can make resulting from performing an action $a \in A$ in a state $s \in S$. The agent receives rewards for performing actions in certain states: this is represented by a *reward function* R: $S \times A \rightarrow \mathbb{R}$. Finally, a *state transition* function τ: $S \times A \rightarrow \Pi(S)$ defines a probability distribution over states resulting from performing an action in a state—this enables us to model non-deterministic actions.

Having defined these sets, we *solve* a POMDP by computing an *optimal policy*: an assignment of an action to each possible belief state such that the expected

sum of rewards gained along the possible trajectories in the POMDP is a maximum. Optimal policies can be computed by applying dynamic programming methods to the POMDP, based on backwards induction; value iteration and policy iteration are the most well known algorithms to solve POMDPs (Boutilier *et al.* 1999). A major drawback of applying POMDPS is that these kinds of algorithms tend to be highly intractable; we later return to the issue of computational complexity as it relates to our model.

4.1. *Intention reconsideration as a POMDP*

We regard the BDI as a *domain dependent object level* reasoner, concerned directly with performing the best action for each possible situation; the POMDP framework is then used as a *domain independent meta level reasoning* component, which lets the agent reconsider its intentions effectively. We define a meta level BDI-POMDP as a tuple $\langle S, A, \Omega, O, R, \tau \rangle$. We have explained above that a state $s \in S$ in this model denotes an internal state of the agent, containing a belief part and intention part. As intention reconsideration is mainly concerned with states, actions and rewards, we leave the implementation of observations Ω, the observation function O and the state transition function τ to the designer for now.

Since the POMDP is used to model intention reconsideration, we are merely concerned with two possible meta level actions: the agent either performs an object level action (*act*) or the agent deliberates (*del*). The possible actions $A = \{act, del\}$ correspond to the agent either acting (*act*) or deliberating (*del*). Because the optimality criterion of policies depends on the reward structure of the POMDP, we define the rewards for action *act* and deliberation *del* in state $s \in S$ as follows:

$$R(s, a) = \begin{cases} W(s_{int}) & \text{if } a = act \\ W(s) & \text{if } a = del \end{cases}$$

where $s_{int} \in S$ refers to the state the agent intends to be in while currently being in state s. Imagine a robot that has just picked up an item which has to be delivered at some location. The agent has adopted the intention to deliver the item, i.e. to travel to that location and to drop off the item. The reward for deliberation is the worth of the agent's current state (e.g. 0) whereas the reward for action is the worth of the intended state (e.g. 10) for having delivered the item. The robot consequently acts, which brings it closer to its 'correct' intentions. Intentions are correct in case the agent does not waste effort while acting upon them. An agent wastes effort if it is deliberating over its intentions unnecessarily. If an agent does not deliberate when that would have been necessary, the agent has wrong intentions.[8]

This structure of reward agrees with the intuition that the agent eventually receives a reward if it has correct intentions, it receives no reward if it has wrong intentions, and it receives no *direct* reward for deliberation. With respect to this last intuition, however, we must mention that the 'real' reward for deliberation is indirectly defined, by the very nature of POMDPs, as the expected worth of future states in which the agent has correct intentions. As intentions resist reconsideration (Bratman *et al.* 1988), the agent prefers action over deliberation and the implementation of the reward structure should thus favour action if the rewards are equivalent.

For illustrative purposes, consider the simple deterministic MDP in figure 12. This figure shows a 5×1 gridworld, in which an agent can move either right or left

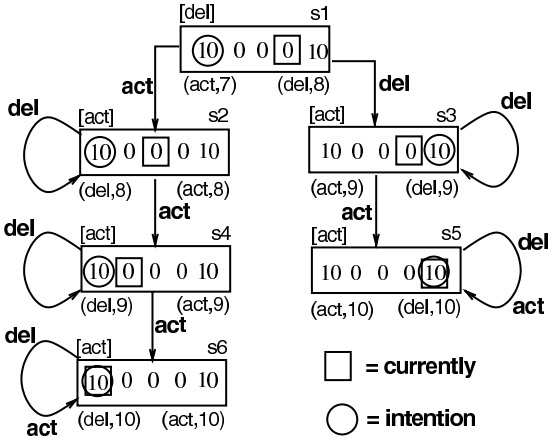


Figure 12. A 5×1 gridworld example which illustrates the definition of rewards in a BDI-POMDP. Rewards, being either 0 or 10, are indicated per location. With each state we have indicated the expected reward for executing a physical action and for deliberation; the best meta action to execute is indicated in square brackets.

or stay at its current location. The agent's current location is indicated with a square and the location it intends to travel to is denoted by a circle. Assume the agent is currently in state s_1: its location is cell 4 and it intends to visit cell 1. Action will get the agent closer to cell 1: it executes a move left action which results in state s_2. Deliberation results in dropping the intention to travel to cell 1, and adopting the intention to travel to cell 5 instead; this results in state s_3. Obviously, deliberation is the best meta action here and the expected rewards for the meta actions in s_1 reflect this: the expected reward for deliberation is higher than the one for action. In all other states, these expected rewards are equivalent, which means that the agent acts in all other states.

Solving a BDI-POMDP means obtaining an optimal intention reconsideration policy: at any possible state the agent might find itself in, this policy tells the agent either to act or to deliberate.

It is important that deciding whether to reconsider intentions or not is computationally cheap compared to the deliberation process itself (Wooldridge and Parsons 1999); otherwise it is just as efficient to deliberate at any possible moment. Using a POMDP to determine the reconsideration policy satisfies this criterion, since it clearly distinguishes between design time computation, i.e. computing the policy, and run time computation, i.e. executing the policy. We recognize that the design time problem of computing a policy is very hard; this problem corresponds with the general problem of solving POMDPs and we do not attempt to solve this problem in this paper. However, the computation that concerns us most is the run time computation, and in our model this merely boils down to looking up the current state and executing the action assigned to that state, i.e. either to act or to deliberate. This is a computationally cheap operation and is therefore suitable for run time execution.

4.2. *Experimental results*

In this section, we describe the use of the POMDP model in the TILEWORLD testbed and discuss the results obtained. The TILEWORLD implementation that we used is exactly that described above.

The TILEWORLD testbed is easily represented as a MDP.[9] Let L denote the set of locations, i.e. $L = \{i\colon 1 \leq i \leq n\}$ represents the mutually disjoint locations, where n denotes the size of the grid. A proposition p_i then denotes the presence ($p_i = 1$) or absence ($p_i = 0$) of a hole at location i. An intention value corresponds to the reward received by the agent for reaching a hole, and an intention cost is the distance between the current location of the agent and the location that the agent intends to reach. An environment state is a pair $\langle \{p_i, \ldots, p_n\}, m\rangle$, where $\{p_i, \ldots, p_n\}$ are the propositions representing the holes in the grid, and $m \in L$ is the current location of the agent.

Combining the $2^n \times n$ possible environment states with n possible intentions means that, adopting explicit state descriptions, the number of states is $2^n \times n^2$, where n denotes the number of locations. Computations on a state space of such size is impractical, even for small n. In order to render the necessary computations feasible, we *abstracted* the TILEWORLD state space. In the TILEWORLD domain, we abstract the state space by letting an environment state e be a pair $\langle p_1, p_2\rangle$, where p_1 refers to the location of the hole which is currently closest to the agent, and p_2 refers to the current location of the agent. Then an agent's internal state is $\langle p_1, p_2\rangle, \{i_1\}$ where i_1 refers to the hole which the agent intends to visit. This abstraction means that the size of the state space is now reduced to n^3. However, the agent now has to figure out at run time what is the closest hole in order to match its current situation to a state in the TILEWORLD state space. This computation can be done in time $O(n)$, by simply checking whether every cell is occupied by a hole or not. Because the main purpose of this example is merely to illustrate that our model is viable, we are currently not concerned with this increase in run time computation.

4.2.1. *Solving the TILEWORLD model off-line*

To summarize, the model that we have to solve off-line consists of the following parts. As described above, the state space S contains all possible internal states of the agent. Each state $s \in S$ is a tuple $\langle\langle p_1, p_2\rangle, \{i_1\}\rangle$, where p_1 refers to hole that is currently closest to the agent, p_2 refers to the current location of the agent, and i_1 denotes the hole which the agent intends to visit. The set of actions is $A = \{act, del\}$. (Note that the set of physical actions is $Ac = \{stay, n, ne, e, se, se, sw, w, nw\}$, but that is not of concern to us since we are concerned with the meta-level control problem rather than object level action selection) Since we assume full observability, the set of observations is $\Omega = S$. Finally, state transitions are defined as the deterministic outcomes of executing an action $a \in A$. As the agent deliberates in state s resulting in state s' (i.e. $\tau\ (s, del) = s'$), then $Bel_s = Bel_{s'}$, but possibly $Int_s \neq Int_{s'}$; as the agent acts (i.e. $\tau\ (s, act) = s''$), then $Int_s = Int_{s''}$, but possibly $Bel_s \neq Bel_{s''}$. Thus, deliberation means that the intention part of the agent's internal state possibly changes, and action means that the belief part of the agent's internal state possibly changes (both *ceteris paribus* with respect to the other part of the internal state). Although solving MDPs and POMDPs in general is computationally hard, we have shown above that by appropriate abstraction of the TILEWORLD state space, the computations become feasible.

4.2.2. *Results*

The experiments resulted in the graphs shown in figures 13, 14a and 14b. In every graph, the environment's dynamism and the agent's planning cost p (for values 0, 1, 2 and 4) are varied. In figure 13, the overall effectiveness of the agent is plotted. In Figure 14a, we plotted the agent's commitment level. (Once again, the collected data was smoothed using a Bezier curve.) Dynamism represents the acting ratio between the world and the agent; this ratio oscillates with the random distribution for hole appearances, on which the commitment level depends and in figure 14b the cost of acting.

4.2.3. *Analysis*

The most important observation we make from these experiments is that the results as presented in figure 13 are overall better than results as obtained in previous investigations into the effectiveness of reconsideration (as elaborated below). Our explanation for this observation is that solving the BDI-POMDP for our TILEWORLD domain delivers an optimal domain dependent reconsideration strategy: the optimal BDI-POMDP policy lets the agent deliberate when a hole appears that is closer than the intended hole (but not on the path to the intended hole), and when the intended hole disappears. This is exactly the reconsideration policy suggested by Kinny and Georgeff (1991). Besides this observation, we see in figure 14a that our

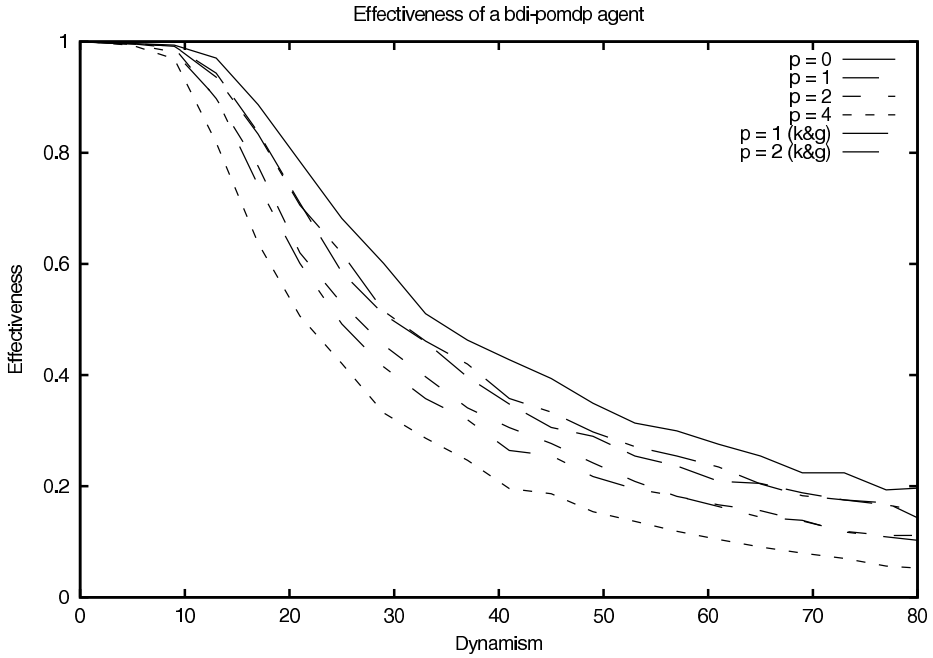
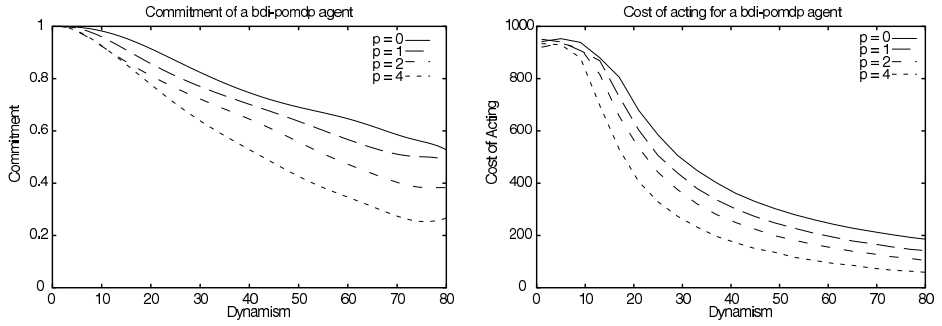


Figure 13. Overall effectiveness of a BDI-POMDP agent. Effectiveness is measured as the result of a varying degree of dynamism of the world. The four curves show the effectiveness at a planning cost (denoted by p) from 0 to 4. The two other curves show the effectiveness at $p=1$ and $p=2$ of Kinny and Georgeff's best reconsideration strategy (from Kinny and Georgeff 1991).



(a) (b)

Figure 14. (a) Average commitment level for a BDI-POMDP agent. The commitment level is plotted as a function of the dynamism of the world with planning cost (denoted by p) of 0, 1, 2 and 4. (b) Average cost of acting for a BDI-POMDP agent. The cost of acting—the number of time steps that the agent moves—is plotted as a function of the dynamism of the world with planning cost (denoted by p) of 0, 1, 2 and 4.

BDI-POMDP agent is able to determine its plan commitment at run time, depending on the state of the environment. This ability contributes to increasing the agent's level of autonomy, since it pushes the choice of commitment level from design time to run time.

In the context of *flexible* strategies, we can compare our results to those of the previous section. The main conclusion we draw from comparing the results from the two strategies is that the empirical outcomes suggest neither approach dominates the other. Comparing the graphs from figures 10 and 13 we observe that the agent's effectiveness is generally higher for our BDI-POMDP model; when we compare the graphs from figure 14 to figure 11, we see that the cost of acting is lower overall in the discrete deliberation model. However, in our BDI-POMDP model, the level of commitment is more constant, since the BDI-POMDP agent's decision mechanism depends less on predictions of appearances and disappearances of holes.

4.3. *Discussion*

While we have evaluated our two approaches to flexible intention reconsideration in the context of the TILEWORLD, we see no difficulty in adpating the techniques we are proposing to different domains. Indeed, the only domain specific information that is used in the formulation of the discrete deliberation scheduling is in the definition of the utility functions. Thus, for any domain, in theory, in which agents engage in a mixture of deliberation and action, it should be possible to apply the deliberation scheduling approach to manage an agent's intention reconsideration. Very similar considerations apply to the POMDP model. The model is only concerned with the meta-level decision about whether to deliberate or act, and this decision is based only upon the utilities of the intention currently being acted on, and the utilities of other potential intentions. Thus, the approach could, in theory, be applied to any domain in which these utilities are available.

5. Related work

The notion of commitment has been widely studied in the agent literature. Two different fields of research can be easily distinguished: a *single agent* and *multi-agent* case. In both fields, only recently investigation has been initiated on run-time decision making. Until now, the majority of previous work presupposed the problem as a design-time one. Whereas in the single agent field, commitment is mostly referred to as *a deliberation and action trade-off*, in the multi-agent field it is a *'pledge' to undertake a specified course of action* (from Jennings 1993) and, obviously, more related to the social property of agents.

Our work originates in the research on the role of intentions in the deliberation process of practical reasoning agents, which was initiated by Bratman *et al.* (1988). Since then, Pollack has investigated the issue of commitment in single practical reasoning agent systems by means of *overloading intentions* (Pollack 1991). The idea behind overloading is closely related to the filter override mechanism in the initial BDI agent model as described in Bratman *et al.* (1988): the agent makes use of opportunities that arise in the world, based on the intentions it has already adopted. This research is more focused on the optimal usage of the current set of intentions, rather than the actual process of deliberating over intentions.

More recently, Veloso *et al.* (1998) used a *rationale based monitoring* (RBM) method to control of reasoning in intentional systems. The idea behind RBM is that plan dependent features of the world are monitored during plan execution; if a feature changes value, this is reason to replan. It must be noted here that the determination of such monitors is a very domain-dependent task and this might hinder the way to a more general domain-independent theory of control of reasoning.

6. Conclusions

We have described research into the efficient and effective reconsideration of intentions in autonomous belief-desire-intention (BDI) agents. We have investigated how reconsideration depends on the environment in which an agent is situated. We used the results of this investigation as the basis of two novel methodologies with which agents can choose appropriate reconsideration strategies. Both these methodologies are decision theoretic: the first based on deliberation scheduling and the second on partially observable Markov decision processes. We regard this kind of management of intentions as a meta-level method for the control of reasoning for an agent to deal with limited computational resources.

We investigated the relationship between the agent's reconsideration rate and its environment. For this, we let the agent have a fixed reconsideration rate and observed the effectiveness of such a rate in different environments. We characterized environments in terms of dynamism, determinism and accessibility. These characteristics represent the rate of change, the predictability of actions and the access to information in the environment, respectively. We found that all characteristics influence the effectiveness of the agent; that the reconsideration rate influences the effectiveness, e.g. a bold agent performs best in a static environment and a cautious agent performs best in a dynamic environment; accessibility has no influence on effectiveness with respect to the agent's reconsideration rate; and reconsidering in highly non-deterministic environments only pays off if planning is free. From the

experimental series in which we vary more than one environmental characteristic, we found that dynamism influences the agent's effectiveness the most.

The results of this investigation suggest that in real world situations, which tend to be dynamic, non-deterministic and inaccessible, agents may be better off by selecting an appropriate reconsideration rate themselves. The two methodologies we developed enable an agent to autonomously decide how strongly committed it is to its intentions. These methodologies are the main contribution of our work. Our approaches give well-founded means of establishing domain-dependent reconsideration strategies (optimal in the case of the POMDP approach). This makes it possible to program agents with essentially domain independent strategies, which they then use to compute domain dependent strategies (off-line in the case of the POMDP model, online in the case of the discrete deliberation scheduling model). Until now, empirical research on meta level reasoning aimed at efficient intention reconsideration has, to the best of our knowledge, involved hardwiring agents with domain dependent strategies.

In the first methodology we use the technique of discrete deliberation scheduling for implementing the agent's meta-level control function that selects an appropriate reconsideration rate. The main idea behind this approach is that while executing a plan to achieve some intention, a trade-off is calculated to either execute a next action or to deliberate and possibly adopt another intention. The trade-off is decision-theoretically determined on the basis of the expected utility of the particular intention. We empirically evaluated the method in the TILEWORLD, where we varied dynamism and planning time. This evaluation demonstrated that: first, the agent outperforms agents with a fixed reconsideration rate (bold/cautious); secondly, the reconsideration rate increases as dynamism increases; and, finally, as planning time increases, the rate of reconsideration increases.

The second methodology is based on the theory of partially observable Markov decision processes (POMDPs). We let the meta-level process of reconsideration be implemented as POMDP. The actions in such a POMDP consist of two meta-level choices to the agent, i.e. either to act or deliberate. In every possible situation, the agent may find itself in, the POMDP indicates whether the agent should act or deliberate. This decision on reconsidering or not is enforced by the POMDP through computing the expected utilities of these meta-level choices. Because of the very nature of constructing and solving MDPs in general, this computation happens completely off-line. Therefore, the online reconsideration process is merely a matter of looking up the current situation and making the related best choice to either deliberate or act. The main problem in constructing the POMDP is the choice of appropriate rewards for action and deliberation, since this is the basis of their expected utility. As intuition prescribes, we let the reward for action be the value of the state one intends to be in and the reward for deliberation the value of the current state. An important advantage of using POMDPs is that we can formalise many reconsideration issues independent of the domain in which an agent is placed. Again, we demonstrated empirically that this approach toward agent design gives better performance than approaches in which the reconsideration rate is fixed. The approach is also better in some aspects than reconsideration based on discrete deliberation scheduling. This is because the the optimal reconsideration policy is computed (off-line), leading to an optimal reconsideration rate during online operation. In addition, the level of commitment is more constant when using the POMDP, since reconsideration depends less on the predictions about the dynamism

of the environment. However, the total number of executed actions, in other words the cost of acting, is lower in the discrete deliberation model. Furture work will further investigate the relative merits of the two approaches.

Acknowledgements
We are grateful to the anonymous reviewers for helpful comments.

Notes

1. To save space, we omitted the graphs from the combined parameter variation series for planning cost equal to 1 and 2, as although we conducted these experiments, the results were consistent with those of the single parameter variation experiments.
2. The function, however, assumes that the life-expectancy of holes is infinite.
3. This contrasts with the *continuous* deliberation scheduling framework, which is the term mainly used to cover work such as anytime algorithms (see e.g. Boddy and Dean 1989).
4. The TILEWORLD is a domain in which it is easier to express costs (in terms of distances) rather than utilities. With an order-reversing mapping from costs to utilities, we can continue to use utilities, which fits our model better.
5. Note that we do *not* let the agent know when or where holes appear, we merely give it some measure of how fast the environment changes.
6. We clearly distinguish intentions from their outcome states and we do not give values to intentions themselves, but rather to their outcomes. For example, when an agent *intends* to deliver coffee, an *outcome* of that intention is the state in which coffee has been delivered.
7. Notice that this problem is the inverse of the utilitarian *lifting problem*: the problem of how to lift utilities over states to desires over sets of states. Discussing the lifting problem, and its inverse, is beyond the scope of this paper, and therefore we direct the interested reader to the work of Lang *et al.* (2001).
8. This approach does not take the time for deliberation into account, but we could do this by reducing the utility of the agent by the utility of the action it would otherwise carry out.
9. Although we have set up the theory as such to allow for partial observability (POMDPs), the example presented here concerns full observability and thus a MDP.

References

Allen, J. F., Hendler, J., and Tate, A. (eds), 1990, *Readings in Planning* (San Mateo, CA: Morgan Kaufmann Publishers).

Boddy, M., and Dean, T., 1989, Solving time-dependent planning problems. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence (IJCAI-89)*, Detroit, MI, pp. 979–984.

Boutilier, C., Dean, T., and Hanks, S., 1999, Decision-theoretic planning: structural assumptions and computational leverage. *Journal of AI Research*.

Bratman, M. E., Israel, D. J., and Pollack, M. E., 1988, Plans and resource-bounded practical reasoning. *Computational Intelligence*, **4**: 349–355.

Brooks, R. A., 1999, *Cambrian Intelligence* (Cambridge, MA: The MIT Press).

Bylander, T., 1994, The computational complexity of propositional STRIPS planning. *Artificial Intelligence*, **69**(1–2): 165–204.

Chapman, D., 1987, Planning for conjunctive goals. *Artificial Intelligence*, **32**: 333–378.

Georgeff, M. P., and Lansky, A. L., 1987, Reactive reasoning and planning. In *Proceedings of the Sixth National Conference on Artificial Intelligence (AAAI-87)*, Seattle, WA, pp. 677–682.

Good, I. J., 1971, Twenty-seven principles of rationality. In V. P. Godambe, and D. A. Sprott (eds) *Foundations of Statistical Inference* (Toronto: Holt Rinehart Wilson), pp. 108–141.

Jennings, N. R., 1993, Commitments and conventions: The foundation of coordination in multi-agent systems. *The Knowledge Engineering Review*, **8**(3): 223–250.

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R., 1998, Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, **101**: 99–134.

Kinny, D., and Georgeff, M., 1991, Commitment and effectiveness of situated agents. In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence (IJCAI-91)*, Sydney, pp. 82–88.

Kinny, D., Georgeff, M., and J. H., 1992, Experiments in optimal sensing for situated agents. In *Proceedings of the Second Pacific Rim International Conference on AI (PRICAI-92)*.

4 Lang, J., Torre, L. V. D., and Weydert, E., 2001, Utilitarian desires. *Autonomous Agents and Multi-Agent Systems*.

Mueller, J. P., 1997, *The Design of Intelligent Agents (LNAI Volume 1177)*. (Berlin: Springer-Verlag).

Pollack, M. E., 1991, Overloading intentions for efficient practical reasoning. *Noûs*, **25**(4): 513–536.

Pollack, M. E., and Ringuette, M., 1990, Introducing the Tileworld: Experimentally evaluating agent architectures. In *Proceedings of the Eighth National Conference on Artificial Intelligence (AAAI-90)*, Boston, MA, pp. 183–189.

Rao, A. S., and Georgeff, M. P., 1992, An abstract architecture for rational agents. In C. Rich, W. Swartout, and B. Nebel (eds), *Proceedings of Knowledge Representation and Reasoning (KR & R-92)*, pp. 439–449.

5 Russell, S., and Norvig, P., 1995, *Artificial Intelligence: A Modern Approach* (Prentice-Hall).

Russell, S., and Subramanian, D., 1995, Provably bounded-optimal agents. *Journal of AI Research*, **2**: 575–609.

Russell, S., and Wefald, E., 1991a, Principles of metareasoning. *Artificial Intelligence*, **49**(1–3): 361–395.

Russell, S. J., and Wefald, E., 1991b, *Do the Right Thing – Studies in Limited Rationality* (Cambridge, MA: The MIT Press).

Schut, M., and Wooldridge, M., 2001, The control of reasoning in resource-bounded agents. *The Knowledge Engineering Review*, **16**(3): 215–240.

Simon, H. A., 1982, Rational choice and the structure of the environment. In H. A. Simon (ed.), *Models of bounded rationality, Volume 2* (Cambridge, MA: The MIT Press), pp. 259–268.

Veloso, M., Pollack, M., and Cox, M., 1998, Rationale-based monitoring for planning in dynamic environments. In *Proceedings of the Fourth International Conference on Artificial Intelligence Planning Systems (AIPS 1998)*.

Wooldridge, M., 2000, *Reasoning about Rational Agents* (Cambridge, MA: The MIT Press).

Wooldridge, M., and Jennings, N. R., 1995, Intelligent agents: theory and practice. *The Knowledge Engineering Review*, **10**(2): 115–152.

Wooldridge, M., and Parsons, S. D., 1999, Intention reconsideration reconsidered. In J. P. Müller, M. P. Singh, and A. S. Rao (eds), *Intelligent Agents V (LNAI Volume 1555)*, (Berlin: Springer-Verlag), pp. 63–80.