Intention Reconsideration as Metareasoning

Marc van Zee
Department of Computer Science
University of Luxembourg
marcvanzee@gmail.com

Thomas Icard
Department of Philosophy
Stanford University
icard@stanford.edu

1 Motivation: Intention Reconsideration

The commonplace observation that agents—human and artificial alike—are subject to resource bounds makes salient the possibility that an agent might have the capability to *control* its own reasoning and decision making abilities, to tune itself so that it has a better chance of spending time thinking about the right things at the right times. The general study of *metareasoning* aims to understand this "reasoning about reasoning" in the context of an agent that needs to budget its time and resources in the optimal way, to achieve the best possible expected outcome.

Much of the work on metareasoning in AI has focused on discovering smart methods for focusing an agent's computational effort in the most useful ways, e.g., in the context of a hard search problem [5, 4]. Meanwhile, much of the work in psychology has considered the very important issue of *strategy selection* in problem solving and related tasks (see, e.g., [3] and references therein). Most of this work views metareasoning through the lens of *value of computation*, an appropriation of the notion of *value of information*, where the information-producing actions are internal computations (this idea goes back to I.J. Good). The work we describe here also pursues this general line.

In this project we are interested in understanding a specific aspect of bounded optimality and metareasoning, namely the control of *plan* or *intention reconsideration*. This problem is more circumscribed than the general problem of metareasoning, but it also inherits many of the interesting and characteristic features. The basic problem is as follows: Suppose an agent has devised a (partial) plan of action for a particular environment, as it appeared to the agent at some time t. But then at some later time $t' > t$—perhaps in the course of executing the plan—the agent's view on the world changes. When should the agent *replan*, and when should the agent keep its current (perhaps improvable, possibly dramatically) plan? In other words, in the specific context of a planning agent who is learning new relevant facts about the world, when should this agent stop to *rethink*, and when should it go ahead and *act* according to its current plan?

This problem was considered early on in philosophy (sometimes called "Hamlet's Problem"), and was then considered in AI as well (see, e.g., [1]). We would like to understand optimal solutions to this problem, and in that direction, we have been investigating different metareasoning strategies—that is, strategies for making the think/act decision in this specific context—and how they fare in different *classes of environments*. The ultimate aim is to be able to determine, from the characteristics of the environment, combined with what we know about the agent, what kind of intention/plan reconsideration strategy will be (at least approximately) optimal. We are also ultimately interested in meta-meta-level strategies, concerning how an agent might interpolate among meta-level reconsideration strategies given observed statistics of some novel environment.

Our work builds on earlier, largely forgotten (regrettably, in our view) work in the *belief-desire-intention* (BDI) agent literature, by Kinny and Georgeff [2] (see also [6]). They compare some rudimentary reconsideration strategies, as a function of several environmental parameters, in simple *Tileworld* experiments. We reproduce their results, and also compare their reconsideration strategies to the *optimal* reconsideration strategies for these environmental parameter settings.

In this abstract we first present a theoretical framework for the intention reconsideration problem in MDPs, in the same spirit as much other work on metareasoning. This involves the construction

of a meta-level MDP in which the two actions are 'think' or 'act'. We then consider Kinny and Georgeff's framework as a special case, reproducing their results, and comparing their agents to an "angelic" agent who decides optimally when to think or act. Interestingly, even the very simple agents Kinny and Georgeff considered behave nearly optimally in certain environments. However, no agent performs optimally across environments. Our results suggest that meta-meta-reasoning may indeed be called for in this setting, so that an agent might tune its reconsideration strategy flexibly to different environments.

2 Theoretical Framework

We formalize intention reconsideration as a metareasoning problem. At each time step, the agent faces a choice between two meta-level actions: *acting* (i.e., executing the optimal action for the current decision problem, based on the current plan) or *deliberating* (i.e., recomputing a new plan). We assume that the agent's environment is inherently dynamic, potentially changing at each time step. As a result, some plan that may be optimal at a certain time may no longer be optimal, or worse, may not be executable at a later time moment. We formalize the sequential decision problem as an MDP (S, A, T, R), where S is a set of states, A is a set of actions, $T : S \times A \times S \to [0, 1]$ is a transition function, and $R : S \times A \times S \to \mathbb{R}$ is a reward function. An agent's view on the world is captured by a *scenario* $\sigma = (S, A, T, R, \lambda)$, where (S, A, T, R) is an MDP, and $\lambda \in S$ is the agent's location in the MDP. At any given time the agent also maintains a policy, or plan, $\pi : S' \to A'$ for some set of states S' and set of actions A', which may or may not equal S and A. Thus, the domain and range of the agent's policy may not even coincide with the current set of states and actions.

We also assume an agent might have a memory store μ, which in the most general case simply consists of all previous scenario/plan pairs: $\mu = \langle\langle\sigma_1, \pi_1\rangle, \ldots, \langle\sigma_{n-1}, \pi_{n-1}\rangle\rangle$. (We will typically be interested in agents with significantly less memory capacity.) Summarizing, an agent's overall state (σ, π, μ) consists of a scenario σ, a plan π, and a memory μ.

2.1 Meta-Level Actions: Think or Act

If the environment were static, then there would be no reason to revise a perfectly good plan.[1] However, environments are of course rarely static. States may become unreachable, new states may appear, and both utilities and probabilities may change. This raises the question of plan reconsideration. We assume that at each time moment, an agent has a choice between two meta-level actions, namely whether to act or to think (deliberate). When the agent decides to act, it will attempt the optimal action according to the current plan. When the agent decides to think, it will recompute a new plan based on the current MDP. The cost of deliberation can either be charged directly, or can be captured indirectly by opportunity cost (missing out on potentially rewarding actions).

2.2 The Dynamics of the Environment

An environment specifies how a state $s = \langle\sigma, \pi, \mu\rangle$, and a choice of meta-decision $\alpha \in \{\mathsf{think}, \mathsf{act}\}$, determine (in general stochastically, according to P_d and P_a) a new state $s' = \langle\sigma', \pi', \mu'\rangle$:

$$\langle\sigma, \pi, \mu\rangle \xrightarrow{\alpha} \langle\sigma', \pi', \mu'\rangle$$

- $\mu' = \langle\langle\sigma_1, \pi_1\rangle, \ldots, \langle\sigma_{n-1}, \pi_{n-1}\rangle, \langle\sigma, \pi\rangle\rangle$;
- if $\alpha = \mathsf{think}$:
  - σ' is some perturbation of σ: $\sigma' \sim P_d(\cdot \mid \sigma)$.
  - π' is a new policy for σ.
- if $\alpha = \mathsf{act}$:
  - σ' is a noisy result of taking action $a = \pi(\lambda)$: $\sigma' \sim P_a(\cdot \mid \sigma)$.
  - $\pi' = \pi$.

[1] Of course, there still might be a question of whether further thought might lead to a better plan in case the current plan was itself selected heuristically or sub-optimally.

Let $\mathcal{S}$ be the set of all possible environment states, which are the scenarios that we introduced in the first subsection, and let $\mathcal{A}$ be the set of all possible actions. Let us assume we have specified concrete perturbation functions P_d and P_a for $a \in \mathcal{A}$. We can lift these to a general transition function $\mathcal{T} : \mathcal{S} \times \{\mathsf{think}, \mathsf{act}\} \times \mathcal{S} \to [0, 1]$, so that

$$\mathcal{T}(s, \alpha, s') \quad = \quad \begin{cases} P_d(\sigma' \mid \sigma) & \text{if } \alpha = \mathsf{think} \text{ and } \pi' \text{ is the revised plan for } \sigma \\ P_{\pi(\lambda)}(\sigma' \mid \sigma) & \text{if } \alpha = \mathsf{act} \text{ and } \pi' = \pi \\ 0 & \text{otherwise} \end{cases}$$

We can also lift the reward functions R over S to reward functions $\mathcal{R}$ over $\mathcal{S}$:

$$\mathcal{R}(s, \alpha, s') = \begin{cases} R(\lambda, a, \lambda') & \text{if } \alpha = \mathsf{act} \\ 0 & \text{if } \alpha = \mathsf{think}, \end{cases}$$

where λ' is the agent's location in scenario σ'. This defines a new meta-level MDP as follows:

$$\langle \mathcal{S}, \{\mathsf{think}, \mathsf{act}\}, \mathcal{T}, \mathcal{R} \rangle$$

Thus, once the set $\mathcal{S}$ and the function $\mathcal{T}$ are specified, we have a well defined MDP, whose space of policies can be investigated just like any other MDP.

3 Experiments

Computing an optimal policy for the meta-level MDP is difficult in general. In this section, we present experimental simulation results on specific classes of environments and agents. We have implemented the general framework from the previous section in Java.[2] While we have also been investigating this general setting, in this abstract we focus on one set of experiments reproducing the aforementioned Tileworld experiments by Kinny and Georgeff, with comparison to an "angelic" metareasoner, who solves the think/act tradeoff approximately optimally.

3.1 Experimental Setup

Kinny and Georgeff present the Tileworld as a 2-dimensional grid on which the time between two subsequent hole appearances is characterized by a gestation period g, and holes have a life-expectancy l, both taken from a uniform distribution. Planning cost p is operationalized as a time delay. The ratio of clock rates between the agent's action capabilities and changes in the environment is set by a *rate of world change* parameter γ. This parameter determines the *dynamism* of the world. When an agent plans, it selects the plan that maximizes hole score divided by distance (an approximation to computing an optimal policy in this setting). The performance of an agent is characterized by its *effectiveness* ϵ, which is its score divided by the maximum possible score it could have achieved. The setup is easily seen as a specific case of our meta-decision problem (see Fig. 2).

Kinny and Georgeff propose two families of intention reconsideration strategies: bold agents, who inflexibly replan after a fixed number of steps, and reactive agents, who respond to specific events in the environment. For us, a *bold* agent only reconsiders its intentions when it has reached the target hole; and a *reactive* agent is a bold agent that also replans when a hole closer than its current target appears, or when its target disappears.

In addition, we consider an *angelic* agent, who approximates the value of computation calculations that would allow always selecting think or act in an optimal way. It does so by recursively running a large number of simulations for the meta-level actions from a given state, approximating the expected value of both, and choosing the better. Because we are interested in the theoretically best policy, the angelic agent is not charged for any of this computation: time stops, and the agent can spend as much time as it needs to determine the best meta-level action (hence the term 'angelic').

[2]The source code is available on Github: `https://github.com/marcvanzee/mdp-plan-revision`. An example MDP visualization is depicted in Figure 1 of Appendix A.

3.2 Results

Graphs of the results can be found in Appendix A. In Figure 3 we compare the bold agent with the angelic planner with the same parameter settings as Kinny and Georgeff and a planning time of 2. Unsurprisingly, the angelic planner outperforms the bold agent. In Figure 4, we increase the planning time to 4, which increases the difference in performance between the angelic planner and the bold agent, while the reactive planner does equally well. However, in Figure 5, we see that when we change the parameters settings such that the world is significantly smaller and holes appear as quickly as they come, the angelic planner outperforms the reactive agent as well. Finally, in Figure 6 we consider a highly dynamic domain in which holes appear and disappear very fast. Here the bold agent outperforms the reactive strategy, and does nearly as well as the angelic agent. In such an environment, agents that replan too often never have a chance to make it toward their goals.

Intriguingly, even these very simple agents—bold agents and rudimentary reactive agents—come very close to ideal in certain environments. This suggests that if we fix a given environment, near-optimal intention/plan reconsideration can actually be done quite tractably. However, since these optimal meta-level strategies differ from environment to environment, this seems to be a natural setting in which meta-meta-level reasoning can be useful. One would like a method for determining which of a family of meta-level strategies one ought to use, given some (statistical or other) information about the current environment, its dynamics and the relative (opportunity) cost of planning.

4 Summary and Outlook

We have formalized and implemented intention reconsideration strategies as a specific case of meta-reasoning. We follow a long line of work in AI on this topic, where metareasoning is understood as involving approximate calculations of value of computation. There are at least two distinctive features of the work presented here. First, we focus on agents faced with the problem of whether to reconsider a plan/intention. Second—and this is what makes the first point most interesting—we focus on the interplay between different meta-level strategies for this problem and the *dynamicity* of the environment, captured by parameter γ. We believe that this angle is both worthwhile and of interest in itself, and that it may also lead to insights about the general metareasoning problem.

While the results presented here concern a rather specific case of the intention revision problem—in the Tileworld, which is not necessarily representative of other domains—the general framework concerns any sequential decision problem in a dynamic environment. Thus, in addition to exploring the possibility of meta-meta-level strategies for this particular domain, we are also currently exploring other settings, e.g., where states themselves may appear and disappear and probabilities may change. We would like as comprehensive an understanding of the general relation between these rational meta-level strategies and environmental parameters as possible, and we believe the results here mark a good first step.

Acknowledgments

M. van Zee is funded by National Research Fund (FNR), Luxembourg, RationalArchitecture project.

References

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

[2] D. N. Kinny and M. P. Georgeff. Commitment and effectiveness of situated agents. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence (IJCAI)*, 1991.

[3] F. Lieder and T. L. Griffiths. When to use which heuristic: A rational solution to the strategy selection problem. In *37th Annual Conference of the Cognitive Science Society*, 2015.

[4] C. H. Lin, A. Kolobov, E. Kamar, and E. Horvitz. Metareasoning for planning under uncertainty. *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*, 2015.

[5] S. Russell and E. Wefald. *Do the Right Thing. Studies in Limited Rationality*. MIT Press, 1991.

[6] M. C. Schut, M. Wooldridge, and S. Parsons. The theory and practice of intention reconsideration. *J. Exp. Theor. Artif. Intell.*, 16(4):261–293, 2004.

A Figures

In this Appendix, we present some illustrations of our simulation environments, and present graphs from some of our simulation results.

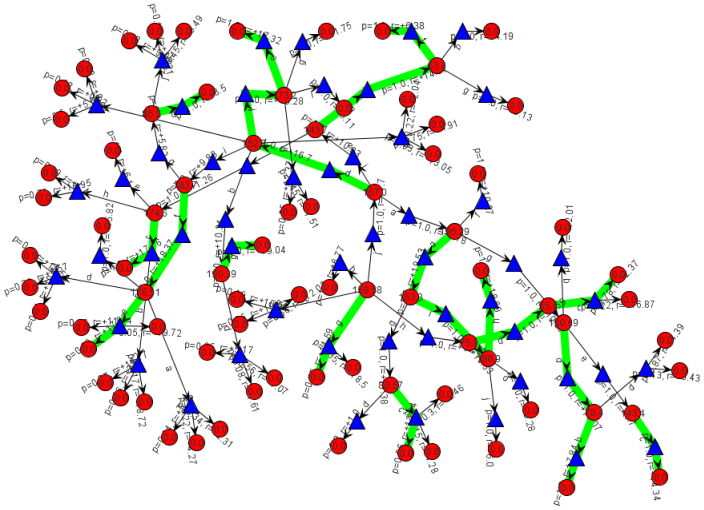

Figure 1: A simulated Markov Decision Process in our software. Red circles denote MDP states, blues triangles denote Q-states, and green arrows denote the optimal policy computed using value iteration. Rewards and probabilities are denotes respectively next to the states and arcs.

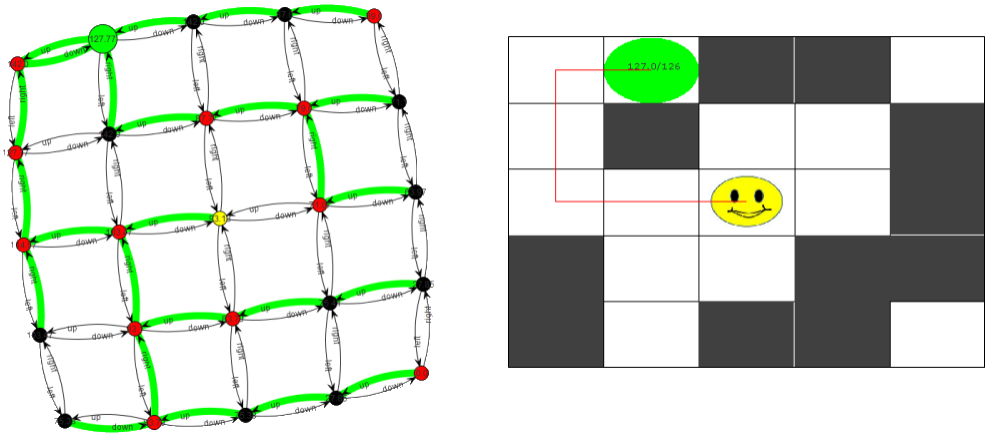

Figure 2: Tileworld representation in our software as an MDP (left), and in the more familiar Tileworld format (right), omitting Q-states (since all probabilities are 1).

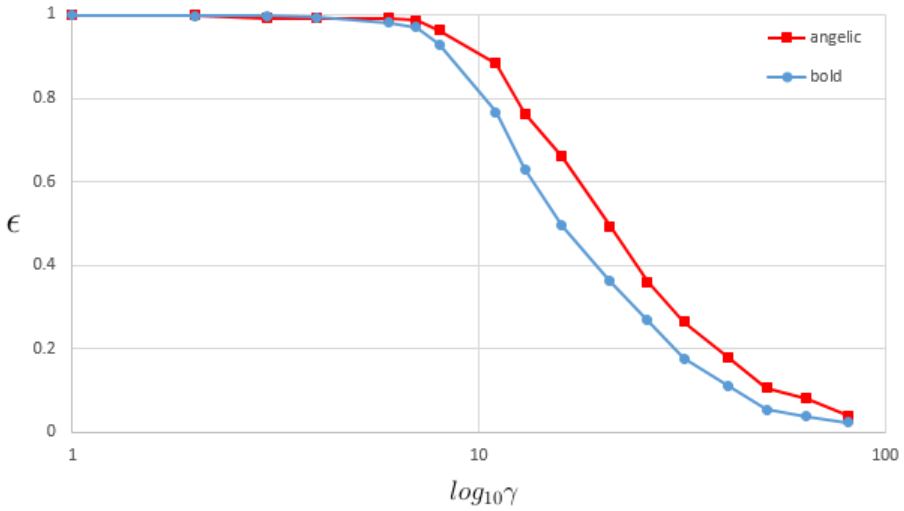

Figure 3: Angelic planner vs Bold agent ($p = 2$). Following Kinny and Georgeff, we plot the rate of the world change γ against the agent's effectiveness ϵ, and we plot values of γ in log_{10}.

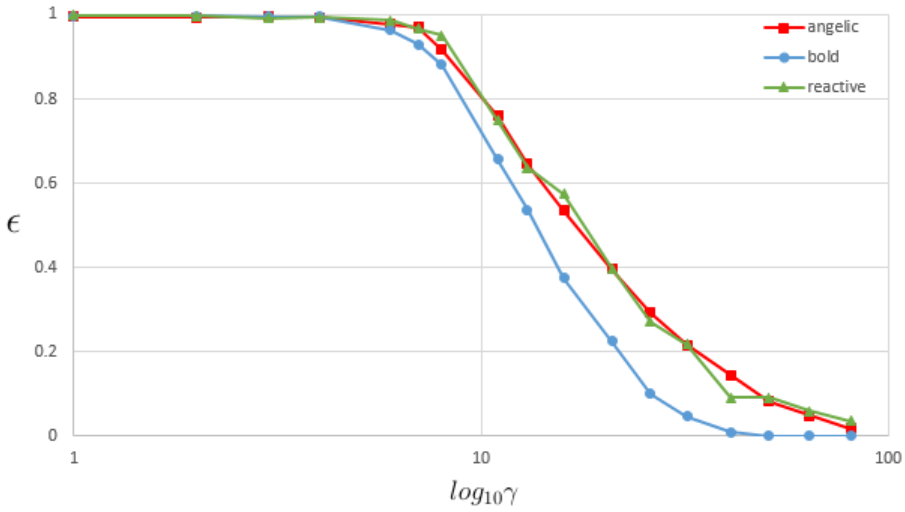

Figure 4: Angelic planner vs Bold agent vs Reactive agent ($p = 4$). The rate of the world change γ is plotted against the agent's effectiveness ϵ.

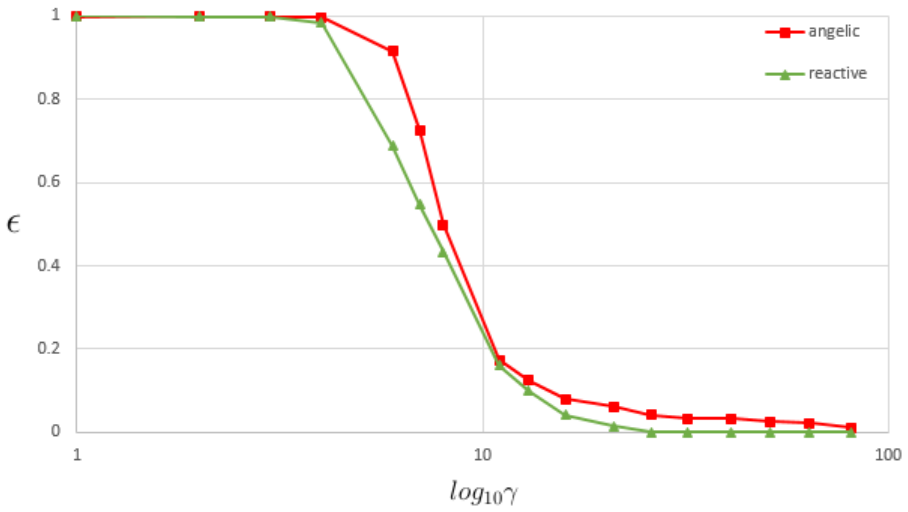

Figure 5: Angelic planner vs Reactive agent ($p = 2, w = 5 \times 5, g = [10, 20], l = [10, 20]$). The rate of the world change γ is plotted against the agent's effectiveness ϵ.

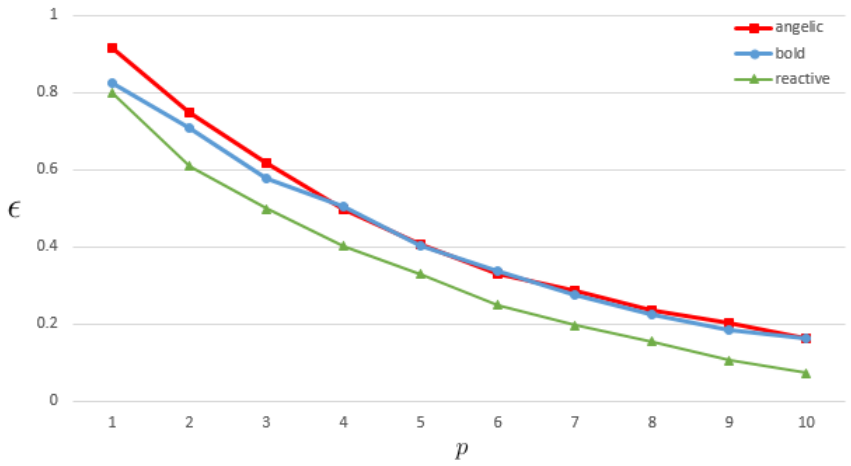

Figure 6: Angelic planner vs Bold agent vs Reactive agent ($p = 2, w = 5 \times 5, g = [3, 5], l = [5, 8]$). The planning time p is plotted against the agent's effectiveness ϵ.