

METACONTROL FOR ADAPTIVE IMAGINATION-BASED OPTIMIZATION

Jessica B. Hamrick
UC Berkeley & DeepMind
jhamrick@berkeley.edu

Andrew J. Ballard
DeepMind
aybd@google.com

Razvan Pascanu
DeepMind
razp@google.com

Oriol Vinyals
DeepMind
vinyals@google.com

Nicolas Heess
DeepMind
heess@google.com

Peter W. Battaglia
DeepMind
peterbattaglia@google.com

ABSTRACT

Many machine learning systems are built to solve the hardest examples of a particular task, which often makes them large and expensive to run—especially with respect to the easier examples, which might require much less computation. For an agent with a limited computational budget, this “one-size-fits-all” approach may result in the agent wasting valuable computation on easy examples, while not spending enough on hard examples. Rather than learning a single, fixed policy for solving all instances of a task, we introduce a *metacontroller* which learns to optimize a sequence of “imagined” internal simulations over predictive models of the world in order to construct a more informed, and more economical, solution. The metacontroller component is a model-free reinforcement learning agent, which decides both *how many* iterations of the optimization procedure to run, as well as *which* model to consult on each iteration. The models (which we call “experts”) can be state transition models, action-value functions, or any other mechanism that provides information useful for solving the task, and can be learned on-policy or off-policy in parallel with the metacontroller. When the metacontroller, controller, and experts were trained with “interaction networks” (Battaglia et al., 2016) as expert models, our approach was able to solve a challenging decision-making problem under complex non-linear dynamics. The metacontroller learned to adapt the amount of computation it performed to the difficulty of the task, and learned how to choose which experts to consult by factoring in both their reliability and individual computational resource costs. This allowed the metacontroller to achieve a lower overall cost (task loss plus computational cost) than more traditional fixed policy approaches. These results demonstrate that our approach is a powerful framework for using rich forward models for efficient model-based reinforcement learning.

1 INTRODUCTION

While there have been significant recent advances in deep reinforcement learning (Mnih et al., 2015; Silver et al., 2016) and control (Lillicrap et al., 2015; Levine et al., 2016), most efforts train a network that performs a fixed sequence of computations. Here we introduce an alternative in which an agent uses a *metacontroller* to choose which, and how many, computations to perform. It “imagines” the consequences of potential actions proposed by an actor module, and refines them internally, before executing them in the world. The metacontroller adaptively decides which expert models to use to evaluate candidate actions, and when it is time to stop imagining and act. The learned experts may be state transition models, action-value functions, or any other function that is relevant to the task, and can vary in their accuracy and computational costs. Our metacontroller’s learned policy can exploit the diversity of its pool of experts by trading off between their costs and reliability, allowing it to automatically identify which expert is most worthwhile.

We draw inspiration from research in cognitive science and neuroscience which has studied how people use a meta-level of reasoning in order to control the use of their internal models and allocation of their computational resources. Evidence suggests that humans rely on rich generative models of the world for planning (Gläscher et al., 2010), control (Wolpert & Kawato, 1998), and reasoning (Hegarty, 2004; Johnson-Laird, 2010; Battaglia et al., 2013), that they adapt the amount of computation they perform with their model to the demands of the task (Hamrick et al., 2015), and that they trade off between multiple strategies of varying quality (Lee et al., 2014; Lieder et al., 2014; Lieder & Griffiths, in revision; Kool et al., in press).

Our imagination-based optimization approach is related to classic artificial intelligence research on bounded-rational *metareasoning* (Horvitz, 1988; Russell & Wefald, 1991; Hay et al., 2012), which formulates a meta-level MDP for selecting computations to perform, where the computations have a known cost. We also build on classic work by Schmidhuber (1990a;b), which used an RL controller with a recurrent neural network (RNN) world model to evaluate and improve upon candidate controls online.

Recently Andrychowicz et al. (2016) used a fully differentiable deep network to learn to perform gradient descent optimization, and Tamar et al. (2016) used a convolutional neural network for performing value iteration online in a deep learning setting. In other similar work, Fragkiadaki et al. (2015) made use of “visual imaginations” for action planning. Our work is also related to recent notions of “conditional computation” (Bengio, 2013; Bengio et al., 2015), which adaptively modifies network structure online, and “adaptive computation time” (Graves, 2016) which allows for variable numbers of internal “pondering” iterations to optimize computational cost.

Our work’s key contribution is a framework for *learning to optimize* via a metacontroller which manages an adaptive, imagination-based optimization loop. This represents a hybrid RL system where a model-free metacontroller constructs its decisions using an actor policy to manage model-free and model-based experts. Our experimental results demonstrate that a metacontroller can flexibly allocate its computational resources on a case-by-case basis to achieve greater performance than more rigid fixed policy approaches, using more computation when it is required by a more difficult task.

2 MODEL

We consider a class of fully observed, one-shot decision-making tasks (i.e., continuous, contextual bandits). The performance objective is to find a *control* $c \in \mathcal{C}$ which, given an initial state $x \in \mathcal{X}$, minimizes some loss function \mathcal{L} between a known future goal state x^* and the result of a forward process, $f(x, c)$. The *performance loss* L_P is the (negative) utility of executing the control in the world, and is related to the optimal solution $c^* \in \mathcal{C}$ as follows:

$$L_P(x^*, x, c) = \mathcal{L}(x^*, f(x, c)), \tag{1}$$

$$c^* = \arg \min_c L_P(x^*, x, c). \tag{2}$$

However, (2) defines only the optimal solution—not how to achieve it.

2.1 OPTIMIZING PERFORMANCE

We consider an iterative optimization procedure that takes x^* and x as input and returns an approximation of c^* in order to minimize (1). The optimization procedure consists of a controller, which iteratively proposes controls, and an expert, which evaluates how good those controls are. On the n^{th} iteration, the controller $\pi^C : \mathcal{X} \times \mathcal{X} \times \mathcal{H} \rightarrow \mathcal{C}$ takes as input, x^* , x , and information about the history of previously proposed controls and evaluations $h_{n-1} \in \mathcal{H}$, and returns a proposed control c_n that aims to improve on previously proposed controls. An expert $E : \mathcal{X} \times \mathcal{X} \times \mathcal{C} \rightarrow \mathcal{E}$ takes the proposed control and provides some information $e_n \in \mathcal{E}$ about the quality of the control, which we call an *opinion*. This opinion is added to the history, which is passed back to the controller, and the loop continues for N steps, after which a final control c_N is proposed.

Standard optimization methods use principled heuristics for proposing controls. In gradient descent, for example, controls are proposed by adjusting c_n in the direction of the gradient of the reward with respect to the control. In Bayesian optimization, controls are proposed based on selection criteria such as “probability of improvement”, or a meta-selection criterion for choosing among

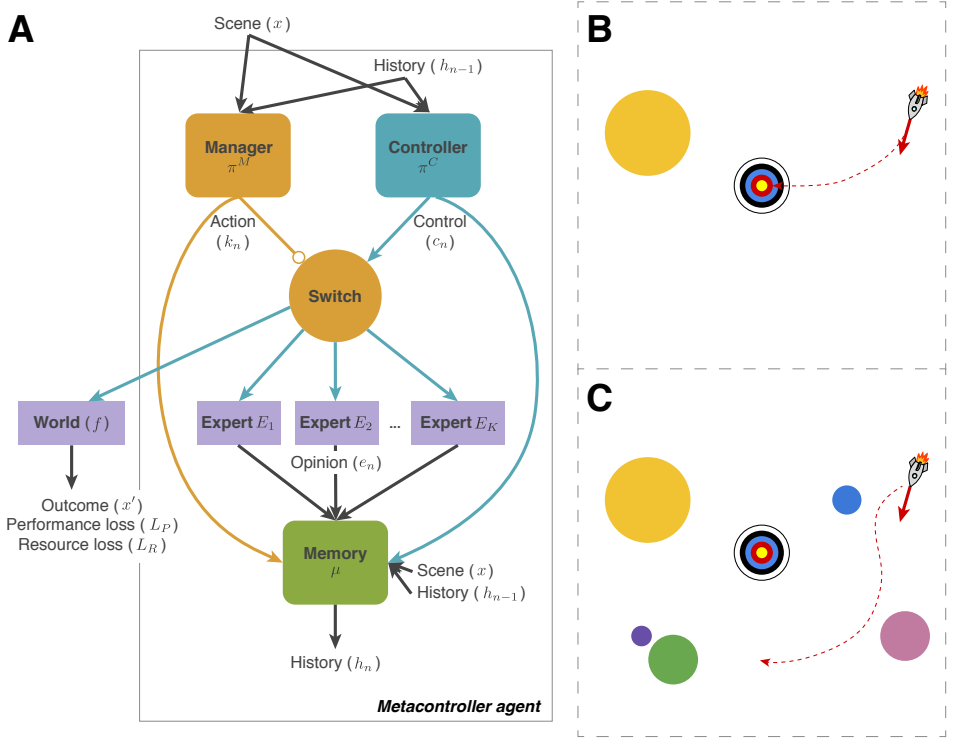


Figure 1: **Metacontroller architecture and task.** **A:** All components are part of the metacontroller agent (box) except the scene and the world, which are part of the agent’s environment. The *manager* takes the scene and history and determines which action to take (i.e., whether to execute or ponder, and with what expert to ponder with), denoted by the orange lines. The *controller* takes the scene and history and computes a control (e.g., the force to apply to a spaceship), denoted by the blue lines. The orange line ending with a circle at the switch reflects the fact that the manager’s action affects the behavior of the switch, which routes the controller’s control to either an *expert* (e.g., a simulation model of the spaceship’s trajectory, an action-value function, etc.) or the *world*. The outcome and reward from the expert, along with the history, action, and control, are fed into the *memory*, which produces the next history. The history is fed back to the controller on the next iteration in order to allow it to propose controls based on what it has already tried. **B-C:** Scenes consisted of a number of planets (depicted here by colored circles) of different masses as well as a spaceship (also with a variable mass). The task was to apply a force to the spaceship for one time step of simulation (depicted here as a solid red arrow) such that the resulting trajectory (dotted red arrow) would put the spaceship at a target (bullseye) after 11 steps of simulation. The white ring of the bullseye corresponds to a performance loss of 0.12-0.15, the black ring to a loss of 0.09-0.12, the blue ring to a loss of 0.06-0.09, the red ring to a loss of 0.03-0.06, and the yellow center to a loss of 0.03 or less. B depicts an easy, 1-planet scene, while C depicts a very difficult 5-planet scene.

several basic selection criteria Hoffman et al. (2011); Shahriari et al. (2014). Rather than choosing one of several controllers, our work learns a single controller and instead focuses on selecting from multiple experts (see Sec. 2.2). In some cases f is known and inexpensive to compute, and thus the optimization procedure sets $E \equiv f$. However, in many real-world settings, f is expensive or non-stationary and so it can be advantageous to use an approximation of f (e.g., a state transition model), L_P (e.g., an action-value function), or any other quantity that gives some information about f or L_P .

2.2 OPTIMIZING COMPUTATIONAL COST

Given a controller and one or more experts, there are two important decisions to be made. First, how many optimization iterations should be performed? The approximate solution usually improves

with more iterations, but each iteration costs computational resources. However, most traditional optimizers either ignore the cost of computation or select the number of iterations using simple heuristics. Because they do not balance the cost of computation against the performance loss, the overall effectiveness of these approaches is subject to the skill and preferences of the practitioners who use them. Second, which expert should be used on each step of the optimization? Some experts may be accurate but expensive to compute in terms of time, energy and/or money, while others may be crude, yet cheap. Moreover, the reliability of the experts may not be known *a priori*, further limiting the effectiveness of the optimization procedure. Our use of a metacontroller address these issues by jointly optimizing over the choices of how many steps to take and which experts to use.

We consider a family of optimizers which use the same controller, π^C , but vary in their expert evaluators, $\{E_1, \dots, E_K\}$. Assuming that the controller and experts are deterministic functions, the number of iterations N and the sequences of experts $\mathbf{k} = (k_1, \dots, k_{N-1})$ exactly determine the final control and performance loss L_P . This means we have transformed the performance optimization over c into an optimization over N and \mathbf{k} : $(N, \mathbf{k})^* = \arg \min_{k,n} L_P(x^*, x, c(N, \mathbf{k}, x, x^*))$, where the notation $c(N, \mathbf{k}, x, x^*)$ is used to emphasize that the control is a function N , \mathbf{k} , x , and x^* .

If each optimizer has an associated computational cost τ_k , then N and \mathbf{k} also exactly determine the computational *resource loss* of the optimization run, $L_R(N, \mathbf{k}) = \sum_{n=1}^{N-1} \tau_{k_n}$. The total loss is then the sum of L_P and L_R , each of which are functions of N and \mathbf{k} ,

$$L_T(x^*, x, N, \mathbf{k}) = L_P(x^*, x, c(N, \mathbf{k}, x, x^*)) + L_R(N, \mathbf{k}) \quad (3)$$

$$= \mathcal{L}(x^*, f(x, \pi^C(x^*, x, h_{N-1}))) + \sum_{n=1}^{N-1} \tau_{k_n}, \quad (4)$$

and the optimal solution is defined as $(N, \mathbf{k})^* = \arg \min_{N, \mathbf{k}} L_T(x^*, x, N, \mathbf{k})$. Optimizing L_T is difficult because of the recursive dependency on the history, h_{N-1} , and because the discrete choices of N and \mathbf{k} mean L_T is not differentiable.

To optimize L_T we recast it as an RL problem where the objective is to jointly optimize task performance and computational cost. As shown in Figure 1a, the **metacontroller agent** a^M is comprised of a controller π^C , a pool of experts $\{E_1, \dots, E_K\}$, a manager π^M , and a memory μ . The *manager* is a meta-level policy (Russell & Wefald, 1991; Hay et al., 2012) over actions indexed by k , which determine whether to terminate the optimization procedure ($k = 0$) or to perform another iteration of the optimization procedure with the k^{th} expert. Specifically, on the n^{th} iteration the controller produces a new control c_n based on the history of controls, experts, and evaluations. The manager, also relying on this history, independently decides whether to end the optimization procedure (i.e., to execute the control in the world) or to perform another iteration and evaluate the proposed control with the k_n^{th} expert (i.e., to *ponder*, after Graves (2016)). The memory then updates the history h_n by concatenating k , c_n , and e_n with the previous history h_{n-1} . Coming back to the notion of imagination-based optimization, we suggest that this iterative optimization process is analogous to imagining what will happen (using one or more approximate world models) before actually executing that action in the world. For further details, see Appendix A, and for an algorithmic illustration of the metacontroller agent, see Algorithm 1 in the appendix.

We also define two special cases of the metacontroller for baseline comparisons. The **iterative agent** a^I does not have a manager and uses only a single expert. Its number of iterations are pre-set to a single N . The **reactive agent**, a^0 , is a special case of the iterative agent, where the number of iterations is fixed to $N = 0$. This implies that proposed controls are executed immediately in the world, and are not evaluated by an expert. For algorithmic illustrations of the iterative and reactive agents, see Algorithms 2 and 3 in the appendix.

2.3 NEURAL NETWORK IMPLEMENTATION

We use standard deep learning building blocks, e.g., multi-layer perceptrons (MLPs), RNNs, etc., to implement the controller, experts, manager, and memory, because they are effective at approximating complex functions via gradient-based and reinforcement learning, but other approaches could be used as well. In particular, we constructed our implementation to be able to make control decisions in complex dynamical systems, such as controlling the movement of a spaceship (Figure 1b-c), though we note that our approach is not limited to such physical reasoning tasks. Here we used mean-squared error (MSE) for our \mathcal{L} and Adam (Kingma & Ba, 2014) as the training optimizer.

Experts We implemented the experts as MLPs and “interaction networks” (INs) (Battaglia et al., 2016), which are well-suited to predicting complex dynamical systems like those in our experiments below. Each expert has parameters θ^{E_k} , i.e. $e_n = E_k(x^*, x, c_n; \theta^{E_k})$, and may be trained either on-policy using the outputs of the controller (as is the case in this paper), or off-policy by any data that pairs states and controls with future states or reward outcomes. The objective L_{E_k} for each expert may be different depending on what the expert outputs. For example, the objective could be the loss between the goal and future states, $L_{E_k} = \mathcal{L}(f(x, c), E_k(x^*, x, c; \theta^{E_k}))$, which is what we use in our experiments. Or, it could be the loss between L_P and an action-value function that predicts L_P directly, $L_{E_k} = \mathcal{L}(L_P(x^*, x, c), E_k(x^*, x, c; \theta^{E_k}))$. See Appendix B.1 for details.

Controller and Memory We implemented the controller as an MLP with parameters θ^C , i.e. $c_n = \pi^C(x^*, x, h_{n-1}; \theta^C)$, and we implemented the memory as a Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) with parameters θ^μ . The memory embeds the history as a fixed-length vector, i.e. $h_n = \mu(h_{n-1}, k_n, c_n, E_{k_n}(x^*, x, c_n); \theta^\mu)$. The controller and memory were trained jointly to optimize (1). However, this objective includes f , which is often unknown or not differentiable. We overcame this by approximating L_P with a differentiable critic analogous to those used in policy gradient methods (e.g. Silver et al., 2014; Lillicrap et al., 2015; Heess et al., 2015). See Appendices B.2 and B.3 for details.

Manager We implemented the manager as a stochastic policy that samples from a categorical distribution whose weights are produced by an MLP with parameters θ^M , i.e. $k_n \sim \text{Categorical}(k; \pi^M(x^*, x, h_{n-1}; \theta^M))$. We trained the manager to minimize (3) using REINFORCE (Williams, 1992), but other deep RL algorithms could be used instead. See Appendix B.4 for details.

3 EXPERIMENTS

To evaluate our metacontroller agent, we measured its ability to learn to solve a class of physics-based tasks that are surprisingly challenging. Each episode consisted of a scene which contained a spaceship and multiple planets (Figure 1b-c). The spaceship’s goal was to rendezvous with its mothership near the center of the system in exactly 11 time steps, but it only had enough fuel to fire its thrusters once. The planets were static but the gravitational force they exerted on the spacecraft induced complex non-linear dynamics on the motion over the 11 steps. The spacecraft’s action space was continuous, up to some maximum magnitude, and represented the instantaneous Cartesian velocity vector imparted by its thrusters. Further details are in Appendix C.

We trained the reactive, iterative, and metacontroller agents on five versions of the spaceship task involving different numbers of planets.¹ The iterative agent was trained to take anywhere from zero (i.e., the reactive agent) to ten ponder steps. The metacontroller was allowed to take a maximum of ten ponder steps. We considered three different experts which were all differentiable: an *MLP expert* which used an MLP to predict the final location of the spaceship, an *IN expert* which used an interaction network (Battaglia et al., 2016) to predict the full trajectory of the spaceship, and a *true simulation expert* which was the same as the world model. In some conditions the metacontroller could use exactly one expert and in others it was allowed to select between the MLP and IN experts. For experiments with the true simulation expert, we used it to backpropagate gradients to the controller and memory. For experiments with an MLP as the only expert, we used a learned IN as the critic. For experiments with an IN as one of its experts, the critic was an IN with shared parameters. We trained the metacontroller on a range of different ponder costs, τ_k , for the different experts. Further details of the training procedure are available in Appendix D.

3.1 REACTIVE AND ITERATIVE AGENTS

Figure 2 shows the performance on the test set of the reactive and iterative agents for different numbers of ponder steps. The reactive agent performed poorly on the task, especially when the task was more difficult. With the five planets dataset, it was only able to achieve a performance loss of 0.583 on average (see Figure 1 for a depiction of the magnitude of the loss). In contrast, the iterative agent with the true simulation expert performed much better, reaching ceiling performance on the

¹Available from: https://www.github.com/deepmind/spaceship_dataset

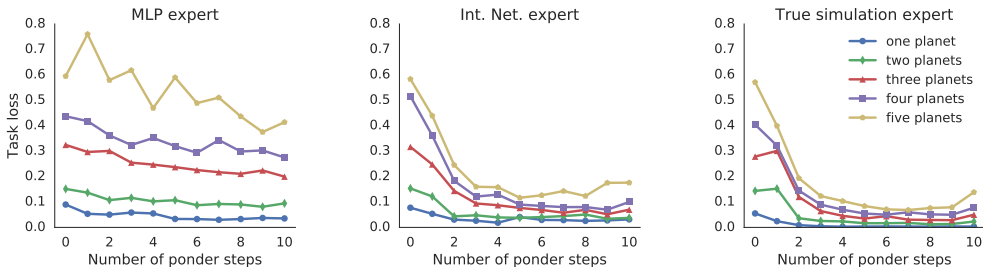


Figure 2: **Test performance of the reactive and iterative agents.** Each line corresponds to the performance of an iterative agent (either the true simulation expert, the MLP expert, or the interaction net expert) trained for a fixed number of ponder steps on one of the five datasets; the line color indicates which dataset the controller was trained on. In all cases, performance refers to the performance loss, L_P . **Left:** the MLP expert struggles with the task due to its limited expressivity, but still benefits from pondering. **Middle:** the IN expert performs almost as well as the true simulation expert, even though it is not a perfect model. **Right:** The true simulation expert does quite well on the task, especially with multiple ponder steps.

datasets with one and two planets, and achieving a performance loss of 0.0683 on the five planets dataset. The IN and MLP experts also improve over the reactive agent, with a minimum performance loss of 0.117 and 0.375 on the five planets dataset, respectively.

Figure 2 also highlights how important the choice of expert is. When using the true simulation and IN experts, the iterative agent performs well. With the MLP expert, however, performance is substantially diminished. But despite the poor performance of the MLP expert, there is still some benefit of pondering with it. With even just a few steps, the MLP iterative agent outperforms its reactive counterpart. However comparing the reactive agent with the $N = 1$ iterative agent is somewhat unfair because the iterative agent has more parameters due to the expert and the memory. However, given that there tends to *also* be an increase in performance between one and two ponder steps (and beyond), it is clear that pondering—even with a highly inaccurate model—can still lead to better performance than a model-free reactive approach.

3.2 METACONTROLLER WITH ONE EXPERT

Though the iterative agents achieve impressive results, they expend more computation than necessary. For example, in the one and two planet conditions, the performances of the IN and true simulation iterative agents received little performance benefit from pondering more than two or three steps, while for the four and five planet conditions they required at least five to eight steps before their performance converged. When computational resources have no cost, the number of steps are of no concern, but when they have some cost it is important to be economical.

Because the metacontroller learns to choose its number of pondering steps, it can balance its performance loss against the cost of computation. Figure 3 (top row, middle and right subplots) shows that the IN and true simulation expert metacontroller take fewer ponder steps as τ increases, tracking closely the minimum of the iterative agent’s cost curve (i.e., the metacontroller points are always near the iterative agent curves’ minima). This adaptive behavior emerges automatically from the manager’s learned policy, and avoids the need to perform a hyperparameter search to find the best number of iterations for a given τ .

The metacontroller does not simply choose an average number of ponder steps to take per episode: it actually tailors this choice to the difficulty of each episode. Figure 4 shows how the number of ponder steps the IN metacontroller chooses in each episode depends on that episode’s difficulty, as measured by the episode’s loss under the reactive agent. For more difficult episodes, the metacontroller tends to take more ponder steps, as indicated by the positive slopes of the best fit lines, and this proportionality persists across the different levels of τ in each subplot.

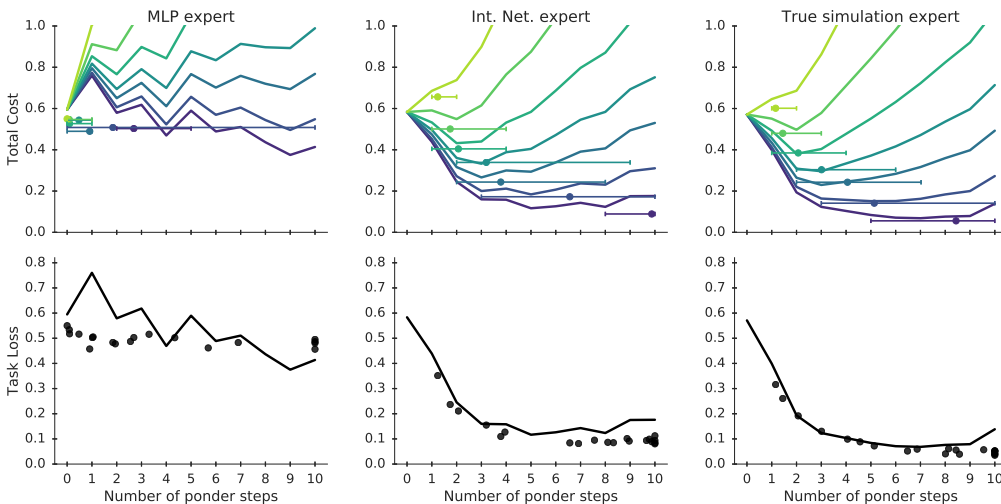


Figure 3: **Test performance of the metacontroller with a single expert on the five planets dataset.** Each column corresponds to a different experts. The lines indicate the performance of the iterative agents for different numbers of ponder steps. The points indicate the performance of the metacontroller, with each point corresponding to a different value of τ . The x -coordinate of each point is an average across the number of ponder steps, and the y -coordinate is the average loss. **Top row:** Here we show total cost rather than just performance on the task (i.e., including computation cost). Different colors show the result for different τ , with the different lines showing the cost for the same iterative controller under different values of τ . The error bars (for the metacontroller) indicate 2.5% and 97.5% confidence intervals. When the point is below its corresponding curve, it means that the metacontroller was able to achieve a better speed-accuracy trade-off than that achievable by the iterative agent. Line colors of increasing brightness correspond to increasing τ , with τ values taken from $[0, 0.0134, 0.0354, 0.0576, 0.0934, 0.152, 0.246]$. **Bottom row:** Here we show just the performance loss (i.e., without computational cost). Each point corresponds to a different value of τ . The fact that the points are below the curve means the metacontroller agent learns to perform better than the iterative agent with the equivalent number of ponder steps.

The ability to adapt its choice of number of ponder steps on a per-episode basis is very valuable because it allows the metacontroller to spend additional computation only on those episodes which require it. The total costs of the IN and true simulation metacontrollers’ are 11% and 15% lower (median) than the *best achievable* costs of their corresponding iterative agents, respectively, across the range of τ values we tested (see Figure 7 in the Appendix for details).

There can even be a benefit to using a metacontroller when there are no computational resource costs. Consider the rightmost points in Figure 3 (bottom row, middle and right subplots), which show the performance loss for the IN and true simulation metacontrollers when τ is low. Remarkably, these points still outperform the best achievable iterative agents. This suggests that there can be an advantage to stopping pondering once a good solution is found, and more generally demonstrates that the metacontroller’s learning process can lead to strategies that are superior to those available to less flexible agents.

The metacontroller with the MLP expert had very poor average performance and high variance on the five planet condition (Figure 3, top left subplot), which is why we restricted our focus in this section to how the metacontrollers with IN and true simulation experts behaved. The MLP’s poor performance is crucial, however, for the following section (3.3) which analyzes how a multiple-expert metacontroller manages experts which vary greater in their reliability.

3.3 METACONTROLLER WITH TWO EXPERTS

When we allow the manager to additionally choose between two experts, rather than only relying on a single expert, we find a similar pattern of results in terms of the number of ponder steps (Figure 5, left). Additionally, the metacontroller is successfully able to identify the more reliable IN network and consequently uses it a majority of the time, except in a few cases where the cost of the IN network is extremely high relative to the cost of the MLP network (Figure 5, right). This pattern of results makes sense given the good performance (described in the previous section) of the metacontroller with the IN expert compared to the poor performance of the metacontroller with the MLP expert. The manager *should not* generally rely on the MLP expert because it is simply not a reliable source of information.

However, the metacontroller has more difficulty finding an optimal balance between the two experts on a step-by-step basis: the addition of a second expert did not yield much of an improvement over the single-expert metacontroller, with only 9% of the different versions (trained with different τ values for the two experts) achieving a lower loss than the best iterative controller. We believe the mixed performance of the metacontroller with multiple experts is partially due to an entropy term which we used to encourage the manager’s policy to be non-deterministic (see Appendix B.4). In particular, for high values of τ , the optimal thing to do is to always execute immediately without pondering. However, because of the entropy term, the manager is encourage to have a non-deterministic policy and therefore is likely to ponder more than it should—and to use experts that are more unreliable—even when this is suboptimal in terms of the total loss (3).

Despite the fact that the metacontroller with multiple experts does not result in a substantial improvement over that which uses a single expert, we emphasize that the manager is able to identify and use the more reliable expert the majority of the time. And, it is still able to choose a variable number of steps according to how difficult the task is (Figure 5, left). This, in and of itself, is an improvement over more traditional optimization methods which would require that the expert is hand-picked ahead of time and that the number of steps are determined heuristically.

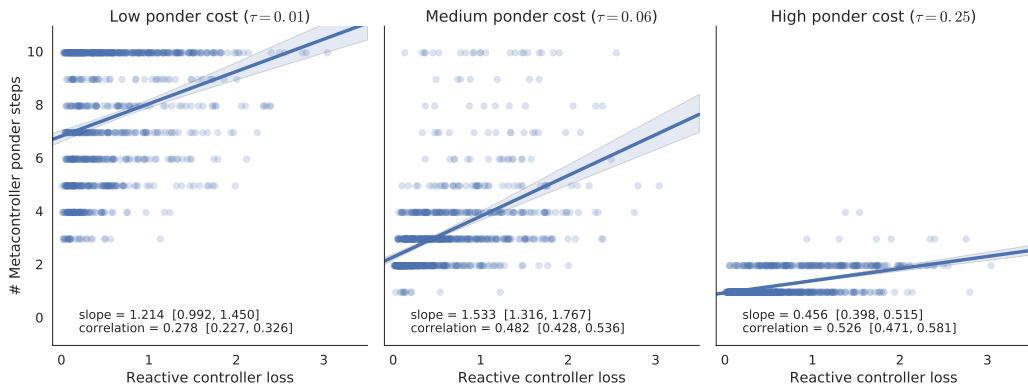


Figure 4: **Relationship between the number of ponder steps and per-episode difficulty for the IN metacontroller.** Each subplot’s x -axis represents the episode difficulty, as measured by the reactive controller’s loss. Each y -axis represents the number of ponder steps the metacontroller took. The points are individual episodes, and the line is the best fit regression line and 95% confidence intervals. The different subplots show different values of τ (labeled in the title). In each case, there is a clear positive relationship between the difficulty of the task and the number of ponder steps, suggesting that the metacontroller learns to spend more time on hard problems and less time on easier problems. At the bottom of each plot are the fitted slope and correlation coefficient values, along with their 95% confidence intervals in brackets.

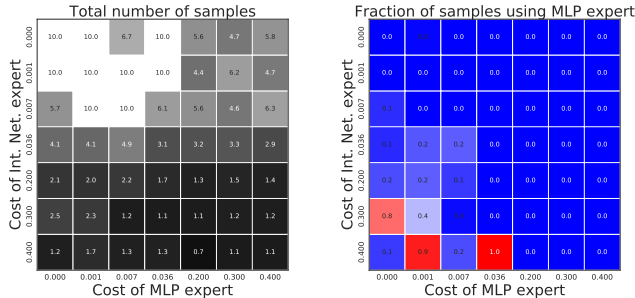


Figure 5: **Test performance of the metacontroller with multiple experts on the five planets dataset.** **Left:** The average number of total ponder steps, for different values of τ . As with the single-expert metacontrollers, fewer ponder steps are taken when the cost is very high, and more are taken when the cost is low. **Right:** The fraction of ponder steps taken by the MLP expert relative to the IN expert. In the majority of cases, the metacontroller favors using the IN expert as it is much more reliable. The few exceptions (red squares) are cases when the cost of the IN expert is much higher relative to the cost of the MLP expert.

4 DISCUSSION

In this paper, we have presented an approach to adaptive, imagination-based optimization in neural networks. Our approach is able to flexibly choose *which* computations to perform as well as *how many* computations need to be performed, approximately solving a speed-accuracy trade-off that depends on the difficulty of the task. In this way, our approach learns to rely on whatever source of information is most useful *and* most efficient. Additionally, by consulting the experts on-the-fly, our approach allows agents to test out actions to ensure that their consequences are not disastrous before actually executing them.

While the experiments in this paper involve a one-shot decision task, our approach lays a foundation that can be built upon to support more complex situations. For example, rather than applying a force only on the first time step, we could turn the problem into one of trajectory optimization for continuous control by asking the controller to produce a sequence of forces. In the case of planning, our approach could potentially be combined with methods like Monte Carlo Tree-Search (MCTS) (Coulom, 2006), where our experts would be akin to having several different rollout policies to choose from, and our controller would be akin to the tree policy. While most MCTS implementations will run rollouts until a fixed amount of time has passed, our approach would allow the manager to adaptively choose the number of rollouts to perform and which policies to perform the rollouts with. Our method could also be used to naturally augment existing model-free approaches such as DQN (Mnih et al., 2015) with online model-based optimization by using the model-free policy as a controller and adding additional experts in the form of state-transition models. An interesting extension would be to compare our metacontroller architecture with a naïve model-based controller that performs gradient-based optimization to produce the final control. We expect our metacontroller architecture might require fewer model evaluations and to be more robust to model inaccuracies compared to the gradient-based method, because our method has access to the full history of proposed controls and evaluations whereas traditional gradient-based methods do not.

Although we rely on differentiable experts in our metacontroller architecture, we do not utilize the gradient information from these experts. An interesting extension to our work would be to pass this gradient information through to the manager and controller (as in Andrychowicz et al. (2016)), which would likely improve performance further, especially in the more complex situations discussed here. Another possibility is to train some or all of the experts inline with the controller and metacontroller, rather than independently, which could allow their learned functionality to be more tightly integrated with the rest of the optimization loop, at the expense of their generality and ability to be repurposed for other uses.

To conclude, we have demonstrated how neural network-based agents can use metareasoning to adaptively choose what to think about, how to think about it, and for how long to think for. Our

method is directly inspired by human cognition and suggests a way to make agents much more flexible and adaptive than they currently are, both in decision making tasks such as the one described here, as well as in planning and control settings more broadly.

ACKNOWLEDGMENTS

We would like to thank Matt Hoffman, Andrea Tacchetti, Tom Erez, Nando de Freitas, Guillaume Desjardins, Joseph Modayil, Hubert Soyer, Alex Graves, David Reichert, Theo Weber, Jon Scholz, Will Dabney, and others on the DeepMind team for helpful discussions and feedback.

REFERENCES

- Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, and Nando de Freitas. Learning to learn by gradient descent by gradient descent. *arXiv:1606.04474*, 2016.
- Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, and Koray Kavukcuoglu. Interaction networks for learning about objects, relations and physics. *Advances in Neural Information Processing Systems*, 2016.
- Peter W. Battaglia, Jessica B. Hamrick, and Joshua B. Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45):18327–18332, 2013.
- Emmanuel Bengio, Pierre-Luc Bacon, Joelle Pineau, and Doina Precup. Conditional computation in neural networks for faster models. *arXiv:1511.06297*, 2015.
- Yoshua Bengio. Deep learning of representations: Looking forward. *arXiv:1305.0445*, 2013.
- Rémi Coulom. Efficient selectivity and backup operators in monte-carlo tree search. In *International Conference on Computers and Games*, pp. 72–83. Springer, 2006.
- Katerina Fragkiadaki, Pulkit Agrawal, Sergey Levine, and Jitendra Malik. Learning Visual Predictive Models of Physics for Playing Billiards. *Proceedings of the International Conference on Learning Representations (ICLR 2016)*, pp. 1–12, 2015. URL <http://arxiv.org/abs/1511.07404>.
- Jan Gläscher, Nathaniel Daw, Peter Dayan, and John P. O’Doherty. States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, 66(4):585 – 595, 2010.
- Alex Graves. Adaptive computation time for recurrent neural networks. *arXiv:1603.08983*, 2016.
- Jessica B. Hamrick, Kevin A. Smith, Thomas L. Griffiths, and Edward Vul. Think again? the amount of mental simulation tracks uncertainty in the outcome. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, 2015.
- Nicholas Hay, Stuart J. Russell, David Tolpin, and Solomon Eyal Shimony. Selecting computations: Theory and applications. *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, 2012.
- Nicolas Heess, Gregory Wayne, David Silver, Tim Lillicrap, Tom Erez, and Yuval Tassa. Learning continuous control policies by stochastic value gradients. *Advances in Neural Information Processing Systems*, 2015.
- Mary Hegarty. Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences*, 8(6):280 – 285, 2004.
- Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- Matthew W Hoffman, Eric Brochu, and Nando de Freitas. Portfolio allocation for Bayesian optimization. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, pp. 327–336, 2011.
- Eric J. Horvitz. Reasoning about beliefs and actions under computational resource constraints. In *Uncertainty in Artificial Intelligence, Vol. 3*, 1988.
- Philip N Johnson-Laird. Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, 107(43):18243–18250, 2010.
- Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.
- Wouter Kool, Fiery A. Cushman, and Samuel J. Gershman. When does model-based control pay off? *PLOS Computational Biology*, in press.
- Sang Wan Lee, Shinsuke Shimojo, and John P. O’Doherty. Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, 81:687–699, 2014.
- Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17:1–40, 2016.
- Falk Lieder and Thomas L. Griffiths. Strategy selection as rational metareasoning. in revision.
- Falk Lieder, Dillon Plunkett, Jessica B. Hamrick, Stuart J. Russell, Nicholas J. Hay, and Thomas L. Griffiths. Algorithm selection by rational metareasoning as a model of human strategy selection. 27:2870–2878, 2014.

- Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv:1509.02971*, 2015.
- Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.
- Stuart Russell and Eric Wefald. Principles of metareasoning. *Artificial Intelligence*, 49(1):361 – 395, 1991.
- Jürgen Schmidhuber. An on-line algorithm for dynamic reinforcement learning and planning in reactive environments. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 1990a.
- Jürgen Schmidhuber. Reinforcement learning in Markovian and non-Markovian environments. *Advances in Neural Information Processing Systems*, 1990b.
- Bobak Shahriari, Ziyu Wang, Matthew W Hoffman, Alexandre Bouchard-Côté, and Nando de Freitas. An entropy search portfolio for Bayesian optimization. *arXiv:1406.4625*, 2014.
- David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. *Proceedings of the 31st International Conference on Machine Learning*, 2014.
- David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.
- Aviv Tamar, Sergey Levine, and Pieter Abbeel. Value Iteration Networks. *Advances in Neural Information Processing Systems*, 2016. URL <http://arxiv.org/abs/1602.02867>.
- Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4):229–256, 1992.
- D.M. Wolpert and M. Kawato. Multiple paired forward and inverse models for motor control. *Neural Networks*, 11(78):1317 – 1329, 1998.

A METACONTROLLER DETAILS

Here, we give the precise definitions of the metacontroller agent. As described in the main text, the iterative and reactive agents are special cases of the metacontroller agent, and are therefore not discussed here.

The metacontroller agent a^M is comprised of the following components:

- A *history-sensitive controller*, $\pi^C : \mathcal{X} \times \mathcal{X} \times \mathcal{H} \rightarrow \mathcal{C}$, which is a policy that maps goal and initial states, and a history, $h \in \mathcal{H}$, to controls, whose aim is to minimize (1).
- A *pool of experts* $\{E_1, \dots, E_K\}$. Each expert $E : \mathcal{X} \times \mathcal{X} \times \mathcal{C} \rightarrow \mathcal{E}$ maps goal states, input states, and actions to *opinions*. Opinions can be either states-only ($\mathcal{E} = \mathcal{X}$), states and rewards ($\mathcal{E} = \mathcal{X} \times \mathbb{R}$), or rewards-only ($\mathcal{E} = \mathbb{R}$). The expert corresponds to the evaluator for the optimization routine, i.e., an approximation of the forward process f .
- A *manager*, $\pi^M : \mathcal{X} \times \mathcal{X} \times \mathcal{H}_n \rightarrow \{0, \dots, K\}$, which is a policy which decides whether to send a proposed control to the world ($k = 0$) or to the k^{th} expert for evaluation, in order to minimize (3). This formulation is based on that used by metareasoning systems (Russell & Wefald, 1991; Hay et al., 2012). Details on the corresponding MDP are given in Appendix A.1.
- A *memory*, $\mu : \mathcal{H}_{n-1} \times \mathcal{Z} \rightarrow \mathcal{H}_n$, which is a function that maps the prior history $h_{n-1} \in \mathcal{H}_{n-1}$, as well as the most recent manager choice, proposed control, and expert evaluation $(k, c, e) \in \{0, \dots, K\} \times \mathcal{C} \times \mathcal{E} = \mathcal{Z}$, to an updated history $h_n \in \mathcal{H}_n$, which is then made available to the manager and controller on subsequent iterations. The history at step n is a recursively defined tuple which is the concatenation of the prior history with the most recently proposed control, expert evaluation, and expert identity: $h_n = h_{n-1} \cap ((k_n, c_n, E_{k_n}(x^*, x, c_n))) = ((k_1, c_1, E_{k_1}(x^*, x, c_1)), \dots, (k_n, c_n, E_{k_n}(x^*, x, c_n)))$ where $h_0 = ()$ represents an empty initial history. Similarly, the finite set of histories up to step n is: $\mathcal{H}_n = \mathcal{H}_{n-1} \times \mathcal{Z} = \mathcal{Z}^n$ where $\mathcal{H}_0 = \{()\}$.

The metacontroller produces:

$$a^M(x^*, x) = \pi^C(x^*, x, h_{N-1}) = c_N \quad (5)$$

where $N = n$ s.t. $k_n = 0$. This function is summarized in Algorithm 1. The other agents (iterative and reactive), as mentioned in the main text, are simpler versions of the metacontroller agent and are summarized in Algorithms 2 and 3.

A.1 META-LEVEL MDP

To implement the manager for the metacontroller agent, we draw inspiration from the metareasoning literature (Russell & Wefald, 1991; Hay et al., 2012) and formulate the problem as a finite-horizon Markov Decision Process (MDP) $\langle \mathcal{S}, \mathcal{A}, P, R \rangle$ over the decision of whether to perform another iteration of the optimization procedure or to execute a control in the world.

- The state space \mathcal{S} consists of goal states, external states, and internal histories, $\mathcal{S} = \mathcal{X} \times \mathcal{X} \times \mathcal{H}$.
- The action space \mathcal{A} contains $K + 1$ discrete actions, $\{0, \dots, K\}$, which correspond to *execute* ($k = 0$) and *ponder* ($k \in \{1, \dots, K\}$), where *ponder* (after Graves (2016)) refers to performing an iteration of the optimization procedure with the k^{th} expert.
- The (deterministic) state transition model $P : \mathcal{S} \times \mathcal{C} \times \mathcal{S} \rightarrow [0, 1]$ is,

$$P(x', h_n | x^*, x, h_{n-1}, k) = \begin{cases} P(x' | x^*, x, h_{n-1}, k) & \text{if } k = 0 \\ P(h_n | x^*, x, h_{n-1}, k) & \text{otherwise} \end{cases}$$

where $x' = f(x, c)$ and $c = \pi^C(x^*, x, h_{n-1})$ and,

$$P(x' | x^*, x, h_{n-1}, k) = \begin{cases} 1 & \text{if } x' = f(x, c) \\ 0 & \text{otherwise} \end{cases}$$

$$P(h_n | x^*, x, h_{n-1}, k) = \begin{cases} 1 & \text{if } h_n = h_{n-1} \cup \{(k, c, E_k(x^*, x, c))\} \\ 0 & \text{otherwise} \end{cases}$$

Algorithm 1 Metacontroller agent. x is the scene and x^* is the target.

```

1: function  $a^M(x, x^*)$ 
2:    $h_0 \leftarrow ()$                                 ▷ Initial empty history
3:    $k_0 \leftarrow \pi^M(x, x^*, h_0)$                 ▷ Get an action from the manager
4:    $c_0 \leftarrow \pi^C(x, x^*, h_0)$                 ▷ Propose a control with the controller
5:    $n \leftarrow 0$ 
6:   while  $k_n \neq 0$  do                            ▷ When  $k \neq 0$ , ponder with an expert
7:      $e_n \leftarrow E_{k_n}(x, x^*, c_n)$             ▷ Get an expert's opinion
8:      $h_{n+1} \leftarrow \mu(h_n, k_n, c_n, e_n)$       ▷ Update the history
9:      $n \leftarrow n + 1$ 
10:     $k_n \leftarrow \pi^M(x, x^*, h_n)$               ▷ Choose the next action
11:     $c_n \leftarrow \pi^C(x, x^*, h_n)$               ▷ Propose the next control
12:  end while
13:  return  $c_n$ 
14: end function

```

Algorithm 2 Iterative agent. x is the scene, x^* is the target, and N is the number of ponder steps.

```

1: function  $a^I(x, x^*, N)$ 
2:    $h_0 \leftarrow ()$                                 ▷ Initial empty history
3:    $c_0 \leftarrow \pi^C(x, x^*, h_0)$                 ▷ Propose a control with the controller
4:    $n \leftarrow 0$ 
5:   while  $n < N$  do                                ▷ Ponder with an expert for  $N$  steps
6:      $e_n \leftarrow E(x, x^*, c_n)$                 ▷ Get the expert's opinion
7:      $h_{n+1} \leftarrow \mu(h_n, k_n, c_n, e_n)$       ▷ Update the history
8:      $n \leftarrow n + 1$ 
9:      $c_n \leftarrow \pi^C(x, x^*, h_n)$               ▷ Propose the next control
10:  end while
11:  return  $c_n$ 
12: end function

```

Algorithm 3 Reactive agent. x is the scene and x^* is the target.

```

1: function  $a^0(x, x^*)$ 
2:    $c_0 \leftarrow \pi^C(x, x^*, ())$                 ▷ Propose a control with the controller
3:   return  $c_0$ 
4: end function

```

- The (deterministic) reward function $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$ maps the current state, current action, and next state to real-valued loss:

$$R(x^*, x, h_{n-1}, k, x') = \begin{cases} \mathcal{L}(x^*, x') & \text{if } k = 0 \text{ (see Eq. 1)} \\ \tau_k & \text{otherwise (see Eq. 3)} \end{cases}$$

where $x' = f(x, \pi^C(x^*, x, h_{n-1}))$.

We approximate the solution to this MDP with a stochastic manager policy π^M . The manager chooses actions proportional to the immediate reward for taking action k in state s_n plus the expected sum of future rewards. This construction imposes a trade-off between accuracy and resources, incentivizing the agent to ponder longer and with more accurate (and potentially expensive) experts when the problem is harder.

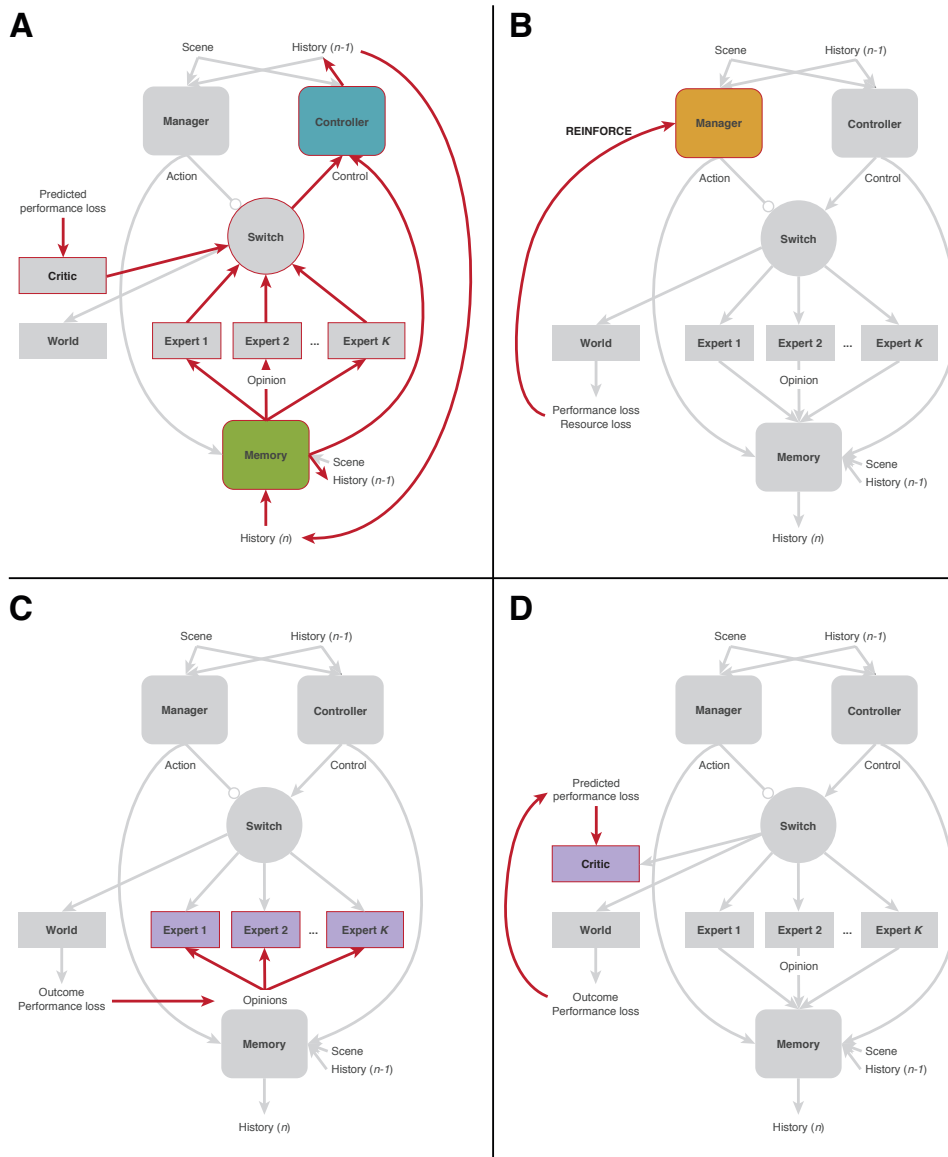


Figure 6: **Training each part of the network.** In each subplot, red arrows depict gradients. Dotted arrows indicate backward connections that are not part of the forward pass. Colored nodes indicate weights that are being updated. All backpropagation occurs at the very end of a full forward pass (i.e., after the control has been executed in the world). **A:** Training the controller and memory with backpropagation-through-time (BPTT), beginning with the critic, and flowing to the controller, through the memory, through the relevant expert, through the controller again, and so on. **B:** Training the manager using REINFORCE (Williams, 1992). **C:** Training the experts (note that each expert may have a different loss with respect to the outcome from the world). **D:** Training the critic.

B GRADIENTS

B.1 EXPERTS

Training the experts is a straightforward supervised learning problem (Figure 6c). The gradient is:

$$\frac{\partial \mathcal{L}^{E_k}}{\partial \theta^{E_k}} = \frac{\partial \mathcal{L}^{E_k}}{\partial E_k} \frac{\partial E_k}{\partial \theta^{E_k}}, \quad (6)$$

where E_k is the k^{th} expert and \mathcal{L}^{E_k} is the loss function for the k^{th} expert. For example, in the case of an action-value function expert, this loss function might be $\mathcal{L}^{E_k}(f, E_k) = \|\mathcal{L}(x^*, f(x, c)) - E_k(x^*, x, c; \theta^{E_k})\|_2$. In the case of an expert that predicts the final state using a model of the system dynamics, it might be $\mathcal{L}^{E_k}(f, E_k) = \|f(x, c) - E_k(x^*, x, c; \theta^{E_k})\|_2$.

B.2 CRITIC

The critic, \hat{L}_P , is an approximate model of the performance loss, L_P , (1), which is used to back-propagate gradients to the controller and memory. This means the critic can either be an action-value function, which approximates $\hat{L}_P = E_0 \approx L_P$ directly, or a model of the system dynamics composed with a known loss function between the goal and future states, $\hat{L}_P = \mathcal{L} \circ E_0 \approx \mathcal{L} \circ f$. We train the critic, $E_0 : \mathcal{X} \times \mathcal{X} \times \mathcal{C} \rightarrow \mathbb{R}$, using the same procedure as the experts are trained (Figure 6d). A good expert may even be used as the critic.

B.3 CONTROLLER AND MEMORY

As shown in Figure 6a, we trained the controller and memory using backpropagation through time (BPTT) with an actor-critic architecture. Specifically, rather than assuming f is known and differentiable, we use a critic and backpropagate through it (Heess et al., 2015):

$$\frac{\partial \mathcal{L}}{\partial \theta^C} = \frac{\partial \mathcal{L}}{\partial E_*} \frac{\partial E_*}{\partial \pi_n^C} \frac{\partial \pi_n^C}{\partial \mu_n} \frac{\partial^+ \mu_n}{\partial \pi_{n-1}^C} \dots \frac{\partial \pi_0^C}{\partial \theta^C}, \quad \frac{\partial \mathcal{L}}{\partial \theta^\mu} = \frac{\partial \mathcal{L}}{\partial E_*} \frac{\partial E_*}{\partial \pi_n^C} \frac{\partial \pi_n^C}{\partial \mu_n} \frac{\partial^+ \mu_n}{\partial \mu_{n-1}} \dots \frac{\partial \mu_0}{\partial \theta^\mu} \quad (7)$$

where E_* is the critic, n is the maximum number of iterations the controller can use, and:

$$\frac{\partial^+ \mu_n}{\partial \pi_{n-1}^C} = \frac{\partial \mu_n}{\partial E_{k_{n-1}}} \frac{\partial E_{k_{n-1}}}{\partial \pi_{n-1}^C} + \frac{\partial \mu_n}{\partial \pi_{n-1}^C}, \quad \frac{\partial^+ \mu_n}{\partial \mu_{n-1}} = \frac{\partial^+ \mu_n}{\partial \pi_{n-1}^C} + \frac{\partial \mu_n}{\partial \mu_{n-1}} \quad (8)$$

where we are using the ∂^+ notation to indicate summed gradients, following Pascanu et al. (2013). Since k_n has already been produced by the manager it can be treated as a constant and will produce an unbiased estimate of the gradient. This is convenient because it allows for training the controller and manager separately, or testing the controller’s behavior with arbitrary actions post-training.

B.4 MANAGER

As discussed in the main text, we used the REINFORCE algorithm Williams (1992) to train the manager (Figure 6b). One potential issue, however, is that when training the controller and manager simultaneously, the controller will result in high cost early on in training and thus the manager will learn to always choose the *execute* action. To discourage the manager from learning what is an essentially deterministic policy, we included a regularization term based on the entropy, L_H (Williams & Peng, 1991; Mnih et al., 2016):

$$L_H(\cdot; \theta^M) = \lambda \mathbb{E}_{\pi^M} [\log \pi^M(\cdot; \theta^M)] \\ \frac{\partial \mathbb{E}_{\pi^M} [r]}{\partial \theta^M} = (r - L_H(\cdot; \theta^M)) \frac{\partial}{\partial \theta^M} \log \pi^M(\cdot; \theta^M),$$

r is the full return given by (3) and λ is the strength of the regularization term.

C SPACESHIP TASK

C.1 DATASETS

We generated five datasets, each containing scenes with a different number of planets (ranging from a single planet to five planets). Each dataset consisted of 100,000 training scenes and 1,000 testing scenes. The target in each scene was always located at the origin, and each scene always had a sun with a mass of 100 units. The sun was located between 100 and 200 distance units away from the target, with this distance sampled uniformly at random. The other planets had a mass between 20 and 50 units, and were located 100 to 250 distance units away from the target, sampled uniformly at random. The spaceship had a mass between 1 and 9 units, and was located 150 to 250 distance units away from the target. The planets were always fixed (i.e., they could not move), and the spaceship always started at the beginning of each episode with zero velocity.

C.2 ENVIRONMENT

We simulated our scenes using a physical simulation of gravitational dynamics. The planets were always stationary (i.e., they were not acted upon by any of the objects in the scene) but acted upon the spaceship with a force of:

$$\mathbf{F}_p = G \frac{m_p m_s}{r^3} (\mathbf{x}_p - \mathbf{x}_s), \quad (9)$$

where \mathbf{F}_p is the force vector of the planet on the spaceship, $G = 1000000$ is a gravitational constant, m_p is the mass of the planet, m_s is the mass of the spaceship, r is the distance between the centers of masses of the planet and the spaceship, \mathbf{x}_p is the location of the planet, and \mathbf{x}_s is the location of the spaceship. We simulated this environment using the Euler method, i.e.:

$$\mathbf{a}_s = \frac{(\sum_p \mathbf{F}_p) - d\mathbf{v}_s + \mathbf{c}}{m_s} \quad \mathbf{x}'_s = \mathbf{x}_s + \epsilon \mathbf{v}_s \quad \mathbf{v}'_s = \mathbf{v}_s + \epsilon \mathbf{a}_s \quad (10)$$

where \mathbf{a}_s , \mathbf{v}_s , and \mathbf{x}_s are the acceleration, velocity, and position of the spaceship, respectively; $d = 0.1$ is a damping constant; \mathbf{c} is the control force applied to the spaceship; and ϵ is the step size. Note that we set \mathbf{c} to zero for all timesteps except the first.

D IMPLEMENTATION DETAILS

We used TensorFlow (Abadi et al., 2015) to implement and train all versions of the model.

D.1 ARCHITECTURE

In our implementation of the controller, we used a two-layer MLP each with 100 units. The first layer used ReLU activations and the second layer used a multiplicative interaction similar to van den Oord et al. (2016), which we found to work better in practice. In our implementation of the memory, we used a single LSTM layer of size 100. In our implementation of the manager, we used a MLP of two fully connected layers of 100 units each, with ReLU nonlinearities.

We constructed three different experts to test the various controllers. The *true simulation* expert was the same as the world model, and consisted of a simulation for 11 timesteps with $\epsilon = 0.05$ (see Appendix C). The IN expert was an interaction network (Battaglia et al., 2016), which has previously been shown to be able to learn to predict n -body dynamics accurately for simple systems. The IN consists of a relational module and an object module. In our case, the relational module was composed of 4 hidden layers of 150 nodes each, outputting “effects” encodings of size 100. These effects, together with the relational model input are then used as input to the object model, which contained a single hidden layer of 100 nodes. The object model outputs the velocity of the spaceship and we trained it to predict the velocity on every timestep of the spaceship’s trajectory. The *MLP* expert was a MLP that predicted the final location of the spaceship and had the same architecture as the controller.

As discussed in Appendix B, we used a critic to train the controller and memory. We always used the IN expert as the critic, except in the case when the true simulation expert was used, in which case we also used the true simulation as the critic.

D.2 TRAINING PROCEDURE

All weights were initialized uniformly at random between 0 and 0.01. An iteration of training consisted of gradient updates over a minibatch of size 1000; in total, we ran training for 100,000 iterations. We additionally used a waterfall schedule for each of the learning rates during training, such that after 1000 iterations, if the loss was not decreasing, we would decay the step size by 5%.

We trained the controller and memory together using the Adam optimizer (Kingma & Ba, 2014) with gradients clipped to a maximum global norm of 10 (Pascanu et al., 2013). The manager was trained simultaneously, but using a different learning rate than the controller and memory. The IN and MLP experts were also trained simultaneously, but again with different learning rates. Learning rates were determined using a grid search over a small number of values, and are given in Table 1 for the iterative agent, in Table 2 for the metacontroller with one expert, and in Table 3 for the metacontroller with two experts.

The iterative agent was trained to take a fixed number of ponder steps, ranging from 0 (i.e., the reactive agent) to 10. The metacontrollers were allowed to take a variable number of ponder steps up to a maximum of 10. For the metacontroller with a single expert, we trained the manager using $\tau = 0$ and 20 additional values of τ spaced logarithmically between 0.00004 and 0.4 (inclusive). For the metacontroller with multiple experts, we trained the manager on a grid of pairs of τ values, where each expert could have $\tau = 0$ or one of 6 values spaced logarithmically between 0.00004 and 0.2 (inclusive). In all cases, the entropy penalty for the metacontroller was $\lambda = 0.2$.

D.3 CONVERGENCE

Reactive agent. Training for the reactive agents was straightforward and converged reliably on all datasets.

Iterative agent. For the iterative agent with the interaction network or true simulation experts, convergence was also reliable for small numbers of ponder steps. Convergence was somewhat less reliable for larger numbers of ponder steps. We believe this is because for some scenes, a larger number of ponder steps was more than necessary to solve the task (as is evidenced by the plateauing performance in Figure 2). So, the iterative agent had to effectively “remember” what the best control was while it took the last few ponder steps, which is a more complicated and difficult task to perform.

For the iterative agent with the MLP expert, convergence was more variable especially when the task was harder, as can be seen in the variable performance on the five planets dataset in Figure 2 (left). We believe this is because the MLP agent was so poor, and that convergence would have been more reliable with a better agent.

Metacontroller with a single expert. The metacontroller agent with a single expert converged more reliably than the corresponding iterative agent (see the bottom row of Figure 3). As mentioned in the previous paragraph, the iterative agent had to take more steps than actually necessary, causing it to perform less well for larger numbers of ponder steps, whereas the metacontroller agent had the flexibility of stopping when it had found a good control. On the other hand, we found that the metacontroller agent sometimes performed too many ponder steps for large values of τ (see Figures 3 and 7). We believe this is due to the entropy term (λ) added to the REINFORCE loss. This is because when the ponder cost is very high, the optimal thing to do is to behave deterministically and always execute (never ponder); however, the entropy term encouraged the policy to be nondeterministic. We plan to explore different training regimes in future work to alleviate this problem, for example by annealing the entropy term to zero over the course of training.

Metacontroller with multiple experts. The metacontroller agent with multiple experts was somewhat more difficult to train, especially for high ponder cost of the interaction network expert. For example, note how the proportion of steps using the MLP expert does not decrease monotonically in Figure 5 (right) with increasing cost for the MLP expert. We believe this is also an unexpected result of using the entropy term: in all of these cases, the optimal thing to do actually is to rely on the MLP expert 100% of the time, yet the entropy term encourages the policy to be non-deterministic. Future work will explore these difficulties further by using experts that complement each other better (i.e., so there is not one that is wholly better than the other).

Experts. The experts themselves always converged quickly and reliably, and trained much faster than the rest of the network.

REFERENCES

- Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL <http://tensorflow.org/>. Software available from tensorflow.org.
- Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, and Koray Kavukcuoglu. Interaction networks for learning about objects, relations and physics. *Advances in Neural Information Processing Systems*, 2016.
- Alex Graves. Adaptive computation time for recurrent neural networks. *arXiv:1603.08983*, 2016.

Dataset	# Ponder Steps	True sim.		MLP		IN	
		α_c	α_c	$\alpha_{E_{IN}}$	$\alpha_{E_{MLP}}$	α_c	$\alpha_{E_{IN}}$
one planet	0	1e-03	1e-03	3e-03	5e-04	1e-03	1e-03
one planet	1	1e-03	1e-03	3e-03	1e-03	1e-03	1e-03
one planet	2	1e-03	1e-03	3e-03	5e-04	1e-03	1e-03
one planet	3	1e-03	1e-03	3e-03	5e-04	1e-03	1e-03
one planet	4	1e-03	1e-03	3e-03	1e-03	1e-03	1e-03
one planet	5	1e-03	1e-03	3e-03	5e-04	5e-04	1e-03
one planet	6	1e-03	1e-03	3e-03	5e-04	1e-03	1e-03
one planet	7	1e-03	1e-03	3e-03	5e-04	1e-03	1e-03
one planet	8	1e-03	1e-03	3e-03	1e-03	1e-03	1e-03
one planet	9	5e-04	1e-03	3e-03	5e-04	5e-04	1e-03
one planet	10	1e-03	1e-03	3e-03	5e-04	1e-03	1e-03
two planets	0	1e-03	1e-03	3e-03	1e-03	3e-03	3e-03
two planets	1	1e-03	1e-03	3e-03	5e-04	1e-03	1e-03
two planets	2	1e-03	1e-03	3e-03	5e-04	1e-03	1e-03
two planets	3	1e-03	1e-03	3e-03	5e-04	1e-03	1e-03
two planets	4	1e-03	1e-03	3e-03	1e-03	1e-03	1e-03
two planets	5	1e-03	1e-03	1e-03	1e-03	1e-03	1e-03
two planets	6	1e-03	1e-03	3e-03	1e-03	1e-03	1e-03
two planets	7	5e-04	1e-03	3e-03	5e-04	5e-04	1e-03
two planets	8	1e-03	1e-03	3e-03	5e-04	5e-04	1e-03
two planets	9	1e-03	1e-03	3e-03	5e-04	3e-03	3e-03
two planets	10	5e-04	1e-03	3e-03	1e-03	5e-04	1e-03
three planets	0	1e-03	1e-03	3e-03	1e-03	1e-03	3e-03
three planets	1	1e-03	1e-03	3e-03	1e-03	1e-03	1e-03
three planets	2	1e-03	5e-04	3e-03	1e-03	1e-03	1e-03
three planets	3	1e-03	1e-03	1e-03	5e-04	1e-03	1e-03
three planets	4	1e-03	1e-03	3e-03	5e-04	1e-03	1e-03
three planets	5	1e-03	1e-03	1e-03	5e-04	5e-04	1e-03
three planets	6	1e-03	5e-04	3e-03	5e-04	1e-03	1e-03
three planets	7	1e-03	1e-03	3e-03	1e-03	1e-03	1e-03
three planets	8	1e-03	1e-03	3e-03	1e-03	5e-04	1e-03
three planets	9	1e-03	1e-03	3e-03	5e-04	1e-03	1e-03
three planets	10	1e-03	5e-04	3e-03	1e-03	1e-03	1e-03
four planets	0	1e-03	5e-04	3e-03	5e-04	1e-03	1e-03
four planets	1	1e-03	5e-04	3e-03	1e-03	1e-03	1e-03
four planets	2	1e-03	5e-04	3e-03	1e-03	1e-03	1e-03
four planets	3	1e-03	1e-03	3e-03	5e-04	1e-03	1e-03
four planets	4	1e-03	5e-04	3e-03	1e-03	1e-03	1e-03
four planets	5	1e-03	1e-03	3e-03	1e-03	1e-03	1e-03
four planets	6	1e-03	1e-03	3e-03	1e-03	1e-03	1e-03
four planets	7	5e-04	1e-03	1e-03	1e-03	1e-03	1e-03
four planets	8	5e-04	1e-03	3e-03	1e-03	1e-03	1e-03
four planets	9	1e-03	1e-03	3e-03	1e-03	5e-04	1e-03
four planets	10	1e-03	1e-03	3e-03	1e-03	5e-04	1e-03
five planets	0	1e-03	1e-03	3e-03	5e-04	1e-03	3e-03
five planets	1	1e-03	1e-03	3e-03	5e-04	1e-03	1e-03
five planets	2	5e-04	1e-03	3e-03	5e-04	1e-03	1e-03
five planets	3	1e-03	1e-03	3e-03	1e-03	1e-03	1e-03
five planets	4	5e-04	1e-03	3e-03	5e-04	1e-03	1e-03
five planets	5	1e-03	5e-04	3e-03	1e-03	1e-03	1e-03
five planets	6	1e-03	1e-03	3e-03	1e-03	1e-03	1e-03
five planets	7	1e-03	1e-03	3e-03	1e-03	1e-03	3e-03
five planets	8	5e-04	1e-03	3e-03	1e-03	1e-03	3e-03
five planets	9	1e-03	1e-03	3e-03	1e-03	1e-03	1e-03
five planets	10	1e-03	1e-03	3e-03	5e-04	1e-03	1e-03

Table 1: Hyperparameter values for the *iterative* controller. α_c refers to the learning rate for the controller and memory, while $\alpha_{E_{IN}}$ refers to the learning rate for the IN expert, and $\alpha_{E_{MLP}}$ refers to the learning rate for the MLP expert.

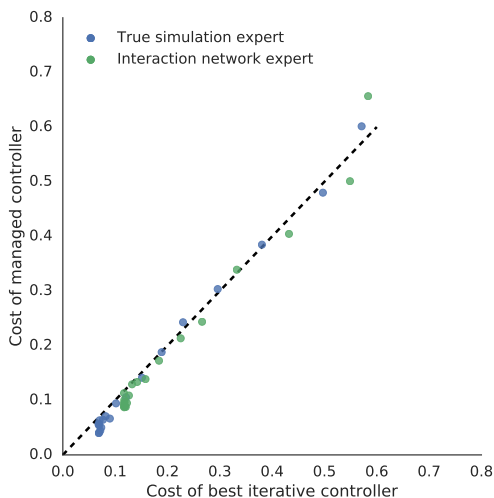


Figure 7: **Cost of the best iterative controller compared to the managed controller.** Each point represents the total cost of the best iterative agent under a particular value of τ (x -axis) versus the total cost achieved by the metacontroller trained with the same value of τ (y -axis). The best iterative agent was chosen by computing the cost for all the different number of ponder steps, and then choosing the whichever number of ponder steps yielded the lowest cost (i.e., finding the minimum of the curves in Figure 3, top row). In almost all cases, the managed controller achieves a lower loss than the iterative controller: for the metacontroller with the IN expert, the cost is 11% lower than the iterative controller on average, and for the metacontroller with the true simulation expert, it is 15% lower on average.

- Nicholas Hay, Stuart J. Russell, David Tolpin, and Solomon Eyal Shimony. Selecting computations: Theory and applications. *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, 2012.
- Nicolas Heess, Gregory Wayne, David Silver, Tim Lillicrap, Tom Erez, and Yuval Tassa. Learning continuous control policies by stochastic value gradients. *Advances in Neural Information Processing Systems*, 2015.
- Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.
- Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. *Proceedings of the 33rd International Conference on Machine Learning*, 2016.
- Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. *Proceedings of the 27th International Conference on Machine Learning*, pp. 1310–1318, 2013.
- Stuart Russell and Eric Wefald. Principles of metareasoning. *Artificial Intelligence*, 49(1):361 – 395, 1991.
- Aäron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. Conditional image generation with PixelCNN decoders. *arXiv:1606.05328*, 2016.
- Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4):229–256, 1992.
- Ronald J. Williams and Jing Peng. Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, 3(3):241–268, 1991.

τ	True sim.		MLP				IN		
	α_c	α_m	α_c	α_m	$\alpha_{E_{IN}}$	$\alpha_{E_{MLP}}$	α_c	α_m	$\alpha_{E_{IN}}$
0.00000	5e-04	5e-04	5e-04	1e-03	3e-03	1e-03	5e-04	1e-04	1e-03
0.00004	1e-03	1e-04	1e-03	5e-05	3e-03	5e-04	1e-03	1e-03	1e-03
0.00006	5e-04	5e-05	1e-03	5e-04	3e-03	1e-03	5e-04	5e-05	1e-03
0.00011	1e-03	1e-04	1e-03	1e-04	3e-03	1e-03	5e-04	5e-04	1e-03
0.00017	5e-04	1e-04	1e-03	1e-03	3e-03	1e-03	1e-03	5e-05	1e-03
0.00028	1e-03	1e-03	1e-03	1e-03	3e-03	1e-03	5e-04	5e-05	1e-03
0.00045	1e-03	1e-03	5e-04	1e-04	3e-03	1e-03	1e-03	5e-05	1e-03
0.00073	1e-03	1e-04	1e-03	1e-04	3e-03	1e-03	1e-03	5e-05	1e-03
0.00119	1e-03	5e-05	1e-03	1e-04	5e-04	1e-03	5e-04	5e-04	1e-03
0.00193	1e-03	5e-05	1e-03	5e-05	3e-03	5e-04	1e-03	5e-05	1e-03
0.00314	1e-03	1e-04	1e-03	1e-04	3e-03	5e-04	1e-03	1e-04	1e-03
0.00510	1e-03	5e-05	1e-03	5e-05	3e-03	1e-03	1e-03	5e-05	1e-03
0.00828	1e-03	5e-04	1e-03	5e-04	3e-03	5e-04	1e-03	1e-03	1e-03
0.01344	1e-03	5e-05	1e-03	5e-05	3e-03	5e-04	5e-04	5e-05	1e-03
0.02182	1e-03	1e-04	1e-03	1e-04	3e-03	5e-04	1e-03	1e-04	1e-03
0.03543	1e-03	1e-04	1e-03	1e-04	3e-03	1e-03	1e-03	1e-04	1e-03
0.05754	1e-03	5e-04	1e-03	5e-04	3e-03	5e-04	1e-03	1e-04	1e-03
0.09343	1e-03	5e-05	1e-03	5e-05	3e-03	1e-03	1e-03	1e-04	1e-03
0.15171	1e-03	1e-04	1e-03	5e-04	3e-03	5e-04	1e-03	1e-04	1e-03
0.24634	1e-03	5e-05	1e-03	1e-03	3e-03	1e-03	1e-03	1e-03	1e-03
0.40000	1e-03	1e-03	1e-03	1e-03	3e-03	5e-04	1e-03	1e-03	1e-03

Table 2: Hyperparameter values for the metacontroller with a single expert. τ refers to the ponder cost, α_c refers to the learning rate for the controller and memory, α_m refers to the learning rate for the manager, $\alpha_{E_{IN}}$ refers to the learning rate for the IN expert, and $\alpha_{E_{MLP}}$ refers to the learning rate for the MLP expert.

τ_{IN}	τ_{MLP}	IN + MLP			
		α_c	α_m	$\alpha_{E_{IN}}$	$\alpha_{E_{MLP}}$
0.00000	0.00000	1e-03	5e-05	1e-03	1e-03
0.00000	0.00121	1e-03	5e-04	1e-03	1e-03
0.00000	0.00663	1e-03	1e-03	1e-03	1e-03
0.00000	0.03641	1e-03	5e-05	1e-03	1e-03
0.00000	0.20000	1e-03	5e-05	1e-03	1e-03
0.00000	0.30000	5e-04	1e-04	1e-03	1e-03
0.00000	0.40000	5e-04	5e-05	1e-03	1e-03
0.00121	0.00000	1e-03	1e-04	1e-03	1e-03
0.00121	0.00121	1e-03	5e-05	1e-03	1e-03
0.00121	0.00663	1e-03	1e-03	1e-03	1e-03
0.00121	0.03641	1e-03	1e-04	1e-03	1e-03
0.00121	0.20000	1e-03	5e-04	1e-03	1e-03
0.00121	0.30000	5e-04	5e-05	1e-03	1e-03
0.00121	0.40000	1e-03	1e-04	1e-03	1e-03
0.00663	0.00000	1e-03	1e-03	1e-03	1e-03
0.00663	0.00121	5e-04	5e-05	1e-03	1e-03
0.00663	0.00663	5e-04	1e-04	1e-03	1e-03
0.00663	0.03641	1e-03	1e-04	1e-03	1e-03
0.00663	0.20000	5e-04	5e-04	1e-03	1e-03
0.00663	0.30000	5e-04	1e-03	1e-03	1e-03
0.00663	0.40000	5e-04	1e-04	1e-03	1e-03
0.03641	0.00000	1e-03	5e-04	1e-03	1e-03
0.03641	0.00121	1e-03	5e-04	1e-03	1e-03
0.03641	0.00663	1e-03	1e-03	1e-03	1e-03
0.03641	0.03641	1e-03	5e-04	1e-03	1e-03
0.03641	0.20000	1e-03	1e-04	1e-03	1e-03
0.03641	0.30000	1e-03	5e-05	1e-03	1e-03
0.03641	0.40000	1e-03	1e-04	1e-03	1e-03
0.20000	0.00000	1e-03	5e-04	1e-03	1e-03
0.20000	0.00121	1e-03	5e-04	1e-03	1e-03
0.20000	0.00663	1e-03	5e-04	1e-03	1e-03
0.20000	0.03641	1e-03	1e-04	1e-03	1e-03
0.20000	0.20000	5e-04	1e-03	1e-03	1e-03
0.20000	0.30000	1e-03	5e-05	1e-03	1e-03
0.20000	0.40000	1e-03	5e-04	1e-03	1e-03
0.30000	0.00000	5e-04	1e-04	1e-03	1e-03
0.30000	0.00121	5e-04	1e-03	1e-03	1e-03
0.30000	0.00663	1e-03	1e-03	1e-03	1e-03
0.30000	0.03641	1e-03	5e-04	1e-03	1e-03
0.30000	0.20000	1e-03	1e-03	1e-03	1e-03
0.30000	0.30000	1e-03	1e-04	1e-03	1e-03
0.30000	0.40000	1e-03	5e-05	1e-03	1e-03
0.40000	0.00000	1e-03	1e-03	1e-03	1e-03
0.40000	0.00121	5e-04	1e-03	1e-03	1e-03
0.40000	0.00663	1e-03	5e-04	1e-03	1e-03
0.40000	0.03641	5e-04	1e-04	1e-03	1e-03
0.40000	0.20000	1e-03	1e-03	1e-03	1e-03
0.40000	0.30000	5e-04	1e-03	1e-03	1e-03
0.40000	0.40000	5e-04	5e-04	1e-03	1e-03

Table 3: Hyperparameter values for the metacontroller with two experts. τ_{IN} refers to the ponder cost for the interaction network expert, τ_{MLP} refers to the ponder cost for the MLP expert, α_c refers to the learning rate for the controller and memory, α_m refers to the learning rate for the manager, $\alpha_{E_{IN}}$ refers to the learning rate for the IN expert, and $\alpha_{E_{MLP}}$ refers to the learning rate for the MLP expert.