Response to the UAI Reviews

We thank the reviewers for their helpful comments and constructive criticism. We would like to take this opportunity to correct some misunderstandings and explain how our camera ready version will address their concerns.

Reviewer 1

Thank you for your helpful feedback and suggestions! We will address them as outlined below. In addition, some parts of the paper appear to have confused you. To correct these misunderstandings, we would like to clarify that

1. Our method is not based on “augmenting the transition rule that is learned in an MDP with additional loss terms”. Our method does not learn any kind of transition model. Rather, it directly searches for a meta-level policy by optimizing the weights of a novel approximation to the value of computation.
2. The recursively blinkered policy not one of the problems on which we evaluated BMPS, but one of our method’s competitors. Concretely, it is a metalevel policy that we devised to extend the previously proposed blinkered policy (Hay et al. 2012) to planning problems with sequentially dependent actions.
3. You wrote that “in psychology and neuroscience it’s well understood that humans make suboptimal decisions (Wason selection task for example)”. While it is true that people’s decisions do not maximize expected utility, this might be simply because maximizing expected utility is computationally intractable. In fact, recent work has shown that the decision mechanisms of the human brain make near-optimal use of people’s finite time and bounded computational resources (see e.g.,Gershman, Horvitz, & Tenenbaum, 2015; Lieder, Griffiths, & Hsu, 2018; Lieder & Griffiths, 2017). AI systems may outperform humans in games like chess and Go, but it takes them at least 10 million times more computation and training data to achieve human level performance (Lake, Ulman, Tenenbaum, & Gershman, 2017). Thus, rather than being suboptimal, people might be optimally balancing the cost of computation with the quality of their decisions in a way that current AI systems can’t. Furthermore, even for the specific example of the Wason card selection task, there is a rational analysis suggesting that people’s strategy corresponds to near-optimal active learning in the natural environment that human cognition is optimized for (Oaksford & Chater, 1994).

The camera ready version of our paper will address your concerns as follows:

1. We will rewrite Sections 3 and 4 to make them more accessible and clearer.
2. We agree that the equations defining the objective function that is being optimized to compute the BMPS policy are difficult to follow, and we have concrete plans to improve

them in a final revision. By “loss function” you may be referring to the expected metalevel return (the objective), which unfortunately is not defined in a display equation. Similarly to a traditional MDP, the expected return is $E[sum_t r_meta(b_t, a_t)]$. Alternatively, you may be referring to Equation 10 which defines our approximation to the Value of Computation. Due to the complexity of the terms involved, it is impractical to write out this equation from the fundamental units. That being said, we are working on simplifying notation in order to make Section 3 easier to follow.

3. You are right that all of their experiments use simulated computation times. This is a valid criticism, and a necessary drawback of synthetic problems. However, we found that the benefit of our method is very robust to our assumptions about how long the object-level computations might take. In fact, the Section 5 of our paper and the Supplementary Material demonstrate that BMPS leads to significant savings in computation time for all plausible durations that each weather simulation might take (and for a wide range of implausible ones as well).
4. You are right that "myopic value of information" is a technical term that will be unfamiliar to most readers. We therefore define this term in Equation 7, and we will italicize it when it is first introduced. Since, the “myopic value of computation” (VOC_1) is a more commonly used term, we will add a sentence explaining that the myopic value of information (VOI_1) is simply the myopic value of computation minus the cost of the optimal sequence of computations. If you have a suggestion for improving the terminology, we are happy to consider it.

Reviewer 2

Thank you for your helpful feedback and suggestions! We will address them as follows:

1. Thank you for suggesting using a running example! In the camera ready version, we will use a concrete sequential decision problem as a running example to explain the features of our VOC approximation and the notion of macro actions.
2. While it is true that Bayesian optimization has been used for policy search before, our application of this general idea to solving the problem of meta-reasoning is very novel in that it represents the first example of a meta-level reinforcement learning algorithm. Thus, the novelty of BMPS goes beyond the introduction of a new parametric class of meta-level policies that is based on insights about the structure of the value of computation.
3. Thank you for pointing us to the interesting work by Kandasamy et al. (2016). We were not previously aware of this paper. We found that our method differs from their MF-GB-UCB method in some important ways that make our method more suitable for certain applications, and we will add a paragraph discussing these differences. I brief, the most important differences are that:

- The Kandasamy et al method, MF-GP-UCB, aims to minimize cumulative regret for all possible computational budgets. In contrast, BMPS determines the optimal amount of computation to perform given the costliness of computation.
- MF-GP-UCB is a method for global optimization of a continuous function. BMPS can be applied to arbitrary decision problems (although, we have so far limited our attention to the discrete action case).
- MF-GP-UCB assumes that a computation directly provides an estimate of the objective function (with varying amounts of noise), whereas BMPS
- Like other UCB algorithms, MF-GP-UCB has a hyper-parameter β_t which controls the size of the confidence bound, or the degree of optimism under uncertainty. BMPS has a similar parameter, w_2 in Eq. 10 (the weight on VPI). However, this parameter is learned (BMPS has no hyper-parameters).

Reviewer 3

Thank you very much for your helpful feedback and suggestions! We will address them as follows:

1. We agree that your question “Can't we simply learn when to select deliberation actions directly without the indirection of VOC computations?” should be explicitly addressed. Meta-MDPs are challenging for traditional approaches because they have extremely large (belief-)state spaces and highly delayed rewards. The large state space precludes exact dynamic programming solutions for all but the smallest metareasoning problems, and necessitates strong generalization for an RL approach. We have found that off-the-shelf neural network policy gradient methods to perform quite poorly, likely because of the temporal credit assignment problem caused by delayed rewards. As you suggest, in the final version, we will include a formal experiment demonstrating the advantage of our approach over off-the-shelf methods.
2. Thank you for suggesting a table of notation definitions. We will add one in the camera ready version.
3. We apologize for the change in notation in Equation 9; this was unnecessarily confusing. Since you asked “how can we go from reasoning about policies under a utility function to dropping max_\pi and reasoning simply about the max utility directly?”, we would like to clarify that the max-operator in Equation 9 is still taken over all policies that the computation c is informative about. This is implicit in our notation because each element of the set $U_c^{(\theta)}$ is the expected return of one of the policies that the computation c is informative about, and each element of the set $\hat{U}_{\neg c}^{(b)}$. Thus, the $\max$ in Equation 9 is in fact equivalent to $\max_\pi$ in Equations 7 and 8. It is very unfortunate that the submitted version obfuscated this equivalence by an accidental change in notation from π to a, when in fact both symbols mean exactly the same. We will carefully rewrite Section 2.3 and Section 3 to convey our

approximation to the VOC and the definition of the features that it uses as clear and as accessible as possible.

4. We apologize that Section 3 was not as clear as it should have been. In particular, we would like to clarify that we suggested that Monte Carlo integration could be used to approximate the main features of our VOC approximation (namely the VOI_1, the VPI_all, and the VPI_a) but we did not mean to suggest that it could be used to approximate the VOC directly. To answer your question “why is a linear interpolation the "right" thing do in light of the preceding Monte Carlo claims that lead in to this equation?”, let us explain why we 1) cannot use Monte Carlo integration to approximate the VOC directly, and 2) linearly interpolate between features (that may be approximated via Monte Carlo Integration). First, Monte Carlo integration can be used to approximate the VOI_1, the VPI_all and the VPI_a features because they are the expected meta-level returns of known sequences of computations, but it cannot be used to approximate the VOC because it pertains to the unknown optimal sequence of computations. Second, the primary motivation for the linear interpolation is the observation that true VOI (which is the VOC without the cost of computation) always lies between the myopic value of information (VOI_1) and the value of perfect information (VPI). The VPI_A term is included because it can provide a tighter bound than VPI in some cases. The cost term is included to capture the expected cost of future computations. Linear interpolation is the simplest way to combine these features into an estimate of the VOC, so it was a natural starting point and we found that it was sufficient to achieve a satisfactory approximation and good performance.