We thank all reviewers for providing insightful comments and helpful suggestions. We have revised some parts of our paper accordingly to improve the clarity. The following are our responses to some specific topics.

Regarding adding a PAC-style analysis for VIEC using learned models (Reviewer 1). Including such an analysis is indeed good for the sake of completeness, but we feel that it also misses the main points of this paper: (1) an exploration strategy with PAC guarantees (R-MAX, MBIE, etc.) can still be far from optimal in terms of exploration cost; (2) explicit planning has remarkable potentials in improving it. Such potentials can only be shown by comparing the exploration costs of the best exploration scheme (the optimal one for the true MDP) and the ones actually taken by the existing algorithms. PAC-style analysis stresses the worst cases, and thus cannot be used to show such potentials. This is a little like guaranteeing that one can run 100m in at most 1 minute, but it does not say how (or even whether) one can achieve 9.6s (which is the main concern of this paper).

It is also very difficult (if not impossible) to use a PAC-style analysis to expose the weaknesses of algorithms such as distance / reward traps, because such an analysis usually occurs at an abstract level and does not show exactly when and how wrong decisions are made during exploration.

Therefore, while the PAC-style analysis is an interesting suggestion and can surely bring some more insight, we do not consider it essential for the purpose of this paper.

Regarding the reward traps, exploration vs exploitation, and the Six Arms example (Reviewer 1). Although algorithms like R-MAX and MBIE are conceptually designed for trading-off exploration and exploitation, in reality they achieve it by doing "pure" exploration first and then converging to some policy at certain stage. As Reviewer 1 mentioned in the Six Arms example, MBIE first explores and tries out different arms, then stops exploration and starts exploiting the best arm after it has sufficient evidence for confirming the optimality of that arm with a high confidence. This is a clear example of exploration-then-exploitation behaviour with exactly one phase change in the process.

As we have discussed in Section 3.1, the minimum sample size for confirming the optimality of a policy / action is given by the Hoeffding's or Chernoff's inequalities, which implicitly result in a fixed demand matrix for a given MDP. Given the same demand matrix, the optimal exploration scheme never chooses actions more than necessary (by definition), while MBIE can be distracted by immediate rewards and eventually choose suboptimal actions more often than needed (which we call reward traps). Therefore, no matter how the domain is designed, MBIE always needs more steps to fulfil a demand matrix (i.e. finish its exploration phase) than the optimal exploration scheme (unless the demand matrix happens to match the actual behaviour of MBIE exactly, in which case the costs are the same).

Even in the case where the total reward is of primary concern (and thus "balancing" exploration and exploitation is important), being lured to the reward traps only prolongs the exploration phase of MBIE, which leads to strictly less total reward in the long run. Therefore, what we called reward traps are still traps in such cases.

We understand that this part can be somewhat tricky (especially to readers accustomed to traditional concepts of exploration). More discussion is added to the final version to help clarify our ideas.

Regarding the computational cost (Reviewer 1). We agree that the computational cost for VIEC is too high to be of practical use. However, an efficient approximation is practical, since a large portion of the demand space is not crucial in deciding the (near-)optimal scheme. We have a simple implementation of the Monte-Carlo approximation that can achieve O(nmd) in tower MDPs like the optimal scheme (albeit with a worse constant factor) which computes much faster than VIEC. This, of course, is still not as fast as heuristic based algorithms and could benefit from further improvement. This will be our future work.

Regarding the lack of numerical experiments (Reviewer 2 & 3). We agree that including a section of empirical results might make the paper more complete. In fact, it is actually more difficult to visualise the exploration behaviours through empirical results than through the analysis done in Section 4. If we only provide the exploration cost curves, it will not provide more useful information than the theoretical results shown in Table 1 of Section 4. We consider that providing more details about how existing algorithms behave during exploration and why some decisions are bad is more helpful for readers to understand the weaknesses of those algorithms as well as the potential of explicit planning, hence the focus on the theoretical analysis in this paper.

Regarding the relationship with Bayesian RL (Reviewer 3). The resemblance between the two is a little superficial. Although both do planning in augmented MDPs, the objectives of planning, the augmented state spaces, and the definitions of optimality are all different. In the context of this paper, Bayesian RL is essentially yet another family of heuristic-based methods. It chooses actions that are (approximately) Bayesian optimal, which are not guaranteed to be optimal with respect to exploration costs. In fact, since Bayesian methods take immediate rewards into consideration, they are prone to reward traps just like the IE family elaborated in Section 4.5. The ones that use discounts in planning are also prone to distance traps.