Thank you very much again for your most helpful thoughts and comments. We will reply to them here; and then work on improving the text along the lines of your concerns. We are rather constrained by the allowable length - but will try to use it more proficiently.

- My understanding of the central idea is that you extend the framework of Mattar and Daw (MD) from their simple EVB for observable states to the more general case of EVBs for belief states.

  I state it like this because it seems like a straightforward idea, whereas in contrast I found the exposition in the first few paragraphs a bit confusing.

That's exactly what we offer in the paper. The exposition in the first few paragraphs is twofold: to set the stage by briefly introducing M&D with all the necessary terminology, and to showcase the critical pitfall of their model which is the inability to explore — something that we subsequently address. We can't really talk about exploratory Gain & Need and the cost of exploration without introducing the key concepts of what replay is in M&D and the original Gain & Need.

I've presented some of this material to Marcelo's lab, and some people had difficulties grasping the idea of planning in belief states — hence the longer introduction to optimal planning and how it relates to hippocampal replay.

However, we will certainly try and make it less confusing.

> For example, figure 1 presents a straightforward example of an exploration vs exploitation tradeoff, but surely this is a shortcoming that would plague any control system at all. In other words, it's not clear why this is a particular shortcoming of MD or anything else to do with replay. Any system that optimizes a policy is going to be caught out by a change to the transition function that affects no-longer-visited states.

The main idea of this figure is to highlight that M&D's predictions for replay "break" as soon as the world changes. They "break" because M&D optimises a policy based on a certain knowledge of the environment. This of course only holds for deterministic environments, and so their normative theory for MDPs cannot make predictions for how replay contributes to learning in the presence of uncertainty. Once the knowledge is wrong, the optimising replay becomes detrimental (as the figure shows, since after the world change replay persists in reinforcing an objectively suboptimal policy). The figure also shows why this happens — no Gain in replaying exploratory actions, and no Need at those states.

We hinted at this in my previous paper which showed how forgetting and subjective state of knowledge significantly impact replay choices. Here we explicitly study the role of parametric uncertainty in a principled way by considering replay in belief states.

One reason to discuss this in the particular context of replay is that Sutton's DYNA, which is the parent of much of the work, including Mattar and Daw, was originally

concerned with arranging for exploration. Thus, it is notable that M&D doesn't actually do that.

> Also, the two paragraphs starting at line 70 are quite confusing. I didn't really understand them to be honest. Rereading it, I think you need to add some more explanations for what Exploratory Gain and Exploratory Need are, before pointing out the problems associated with them. I was able to pick up the thread using figures 1 and 2, so I'm not sure if these two paragraphs are doing much for the exposition.

We will rewrite these paragraphs, and perhaps add a (supplementary) figure illustrating the exploratory Gain example.

- I realized reading this that there might be something "clunky" about handling exploration in this way (or by using exploratory bonuses). Essentially, you are trying to engender exploration by tweaking the value function. You want to fool the agent into thinking that there is reward in places just to get it to explore those places. I imagine an ideal exploration process would be explicit, ie the agent would know it was deviating from the currently optimal policy in order to find out information. That way it can revert to the optimal policy if needs be. This intuition fits the idea of a temperature, which is only AFTER the policy and doesn't have to impact anything else, but ok you want to look at "directed" exploration. Is there no way to have directed exploration that is similarly explicit and does not involved messing up the agent's (currently optimal) model of the world?

As I think you noted in your last corrective comment: it's actually not true that we're tweaking the value function. We are really trying to approximate the *true* Bayesian value function, which exactly integrates exploration and exploitation. That is, Bayesian decision theorists don't actually make a distinction between these two modes - they just have a single value for a state that seamlessly integrates the expected benefits of exploitation with the expected benefits of exploration. Thus this is not really clunky at all! The two parts that are arguably more clunky are the need to have a stochastic policy (to fix Need) and the requirement for sequence replay - but these are a bit secondary.

There are algorithms that more explicitly separate exploration and exploitation - including Explicit-Explore-Exploit (E^3) and the recent Go-Explore algorithms. But they place quite strong demands on memory.

Anyhow, along with also the next point, we should evidently be clearer on the reasoning here about exploration - and the need for doing some hard computational work to explore well.

> I may have been thinking naively about it, but my own conception of the power of offline replay is that it releases an agent to explore explicitly during behavior, trying off-policy moves etc, because behavior is no longer a necessary part of the "value iteration" or "inverse model building" process -- replay instead handles that. However, you are now adding this indirect process of using (incorrect) values as a lever to push the agent into exploration, which is a mechanism that I suppose I didn't think was necessary.

You're right that replay does the inverse model building. However, it's important to *plan* exploration (and its integration with exploitation) - and replay is even more important for this. This was DYNA's original use for replay, as noted above. If you didn't do this, you would be stuck not knowing how much exploration is worthwhile in which circumstances.

- I thought the demonstration of figures 2E,F was very interesting. It made me wonder though if this is part of a larger critique of MD (and prioritized sweeping), that perhaps it only works under very special circumstances, when the values are acquired monotically from an initialization of all values to zero. Whereas change to a pre-existing model, as here, is the more common situation, and so it's a grave shortcoming if MD cannot deal with it. In other words, this feels like a bigger point about offline replay than just your particular model, and it might be worth expanding on that.

Indeed, M&D replay choices depend on the agent's current state of knowledge — i.e., its policy and MF values, as well as the values suggested by the model. This is what we try to address with sequence replay, with the critique being that M&D replay choices are myopic.

We'll think about how to expand on this. We're really short of space, which makes amplifying anything hard - particularly since you noted that the first paragraphs are confusing, and so will probably need expanding. Perhaps we can bring it out more in the discussion?

- The solution of figure 3 is appealing, capturing the intuition that replay effectively "chunks" sequences of states that rarely or never occur independently. However, compared to MD you have moved sequences from one side of the ledger to the other (from Asset to Liability!). Is this a problem? Generating sequences from first principles was one of the big (biggest?) results of MD, so how should we think about your model now that you are assuming it as a mechanism?

This is indeed a very interesting prediction from our modelling — which I would love to see experimentally tested (whether replay follows myopic or deep valuation)! Also, perhaps it's not very surprising given the great amount of effort in the AI/planning community to surpass the complexity of deep exploration — it's a very important question! The interesting bit for the hippocampally-inclined is the predictions for specific replay patterns to be expected.

It's certainly a very computationally demanding problem, much more so than M&D's sequence generation. However, the difference in this demand comes from the need for simultaneously evaluating a huge space of potential sequences. M&D still evaluate EVB for the sequences they generate, but by incrementally building them from shorter sequences.

This is of course a normative theory, much like M&D, which is very unlikely to be algorithmically credible.

- About the final section concerning unexplored space. I'm glad you avoided the most contentious of these claims! I confess though that in my opinion the ones you cite are still a bit weak (Redish observed in one rat, and we are dependent on the assertion that in weeks of pretraining the rat never once turned around where he shouldn't have; the Spiers paper is interesting but the data is quite underpowered unfortunately).

They are indeed weak and the hope is that our work will inspire more careful examination of similar experiments. The modelling of Olafsdottir et al. is part of a grand scheme of getting into Nature Neuro :)

> But you also have a funny situation in the model, where prior to experience your agent still has a model of the space that is almost the whole transition function! (S x S, not the full S x S x A).
>
> You can sellotape it altogether by claiming that's what a rat can get just by looking, but I wouldn't want to put any weight on it.

They unfortunately did not look at replay during the Run 1 period, since the amount of data collected during those 10 minutes was rather meagre for a satisfying replay analysis. Olafsdottir et al. only report the comparison between replay in Rest 1 (prior to experience) and Rest 2 (after the goal arm was cued). Which does indeed leave us in a funny situation where we have to resort to something like a 'proto-map' of unseen space.

They do, however, perform a comparison of the average firing rate of place cells which encode the cued arm during Run 1 (before it was cued) and Run 2 (after it was cued) based on bootstrapped data. They report that the average firing rate of those cells was significantly higher in Run 2 compared to Run 1 (below chance) – which is of course an indirect (proxy) measure of replay. So we can't use this for replay modelling (since they didn't measure replay in Run 1) – but we can speculate that it would've looked the same (comparison of replay in Rest 2 vs Rest 1 and Rest 2 vs Run 1), where in Run 1 the rats would have built the map just by looking through the see-through barrier.

- A final stylistic comment. Going back to #1, I see this as an attempt at generalizing MD to something broader (incorporating uncertainty in a principled way). However, the exposition feels a bit ad hoc with everything explained in the Tolman maze. The plus is that the maze made it easier for me to understand the concepts. However, it left me wondering which bits were specific to that maze versus being illustrative of broader principles.

We chose the Tolman maze mostly because it is a very simple yet powerful environment which allows us to demonstrate all the ideas and results (cost of exploration, blocked corridor, sequence vs greedy replay) in a single maze. Of course, we also appreciated its historical significance and relevance to the hippocampus community. There is one other example maze in the paper which shows the experiment of Olafsdottir, although it appears very briefly. In the supplementary, we also have a very extensive treatment of multi-arm bandits, which is there to i) show

how the theory works in a much simpler setting; and ii) give another example to demonstrate that the theory is very general.

Maybe we can be more explicit about the generality of the theory in the main text by referring to the bandit section more extensively? And by motivating more strongly the choice of the Tolman maze for most of the exposition.

- There were a few little things that seemed to need explanation although maybe it is in the Methods, apologies for not looking carefully enough. Why is the belief state a distribution of probabilities instead of a single probability (or two: P(open) and P(shut), which is the same as one)? Does it matter for the results in fig 2 that one barrier gets to be probabilistic and the other is stable?

The barriers need not be deterministic — i.e., they can more generally be stochastic, in which case it's important for animals/agents to learn the probability distributions associated with their states (open/closed). Also, we discretise belief states for tractability by restricting the posterior to be either P(open)=0 or P(open)=1, whereas optimal updates would result in continuous belief states in the range [0, 1].

It does matter (for some belief states) because, if the agent was sufficiently uncertain about the lower barrier, then the exploration bonus would propagate 'through' it and result in a different exploratory policy which would favour the exploration of both barriers through the shortest possible path, instead of going 'left' and the junction. This is in fact how the bonus propagates towards the start state in Fig 4.

In Fig 2 the other barrier is stable for simplicity, but in Fig S7 both barriers are probabilistic.

- Ok rereading my comments I have missed an important feature which is that there is a navigation problem in figuring out how to get to the bit of space you want to explore. That was also in the exploratory bonuses idea. So ok that is why you would want to fold it into the value function. Alright so then my point #2 becomes about reminding the dullards like me of this important point!

Yes this is exactly why we need it in the value function :) We will try and be clearer about this.