We thank all of the reviewers for their useful and insightful feedback. We are encouraged they found our work interesting (R1, R2, R3, R4), novel (R2) and scientifically rigorous (R2, R3). We are happy all reviewers recognize the importance of the problem we address. We are glad they found our work impactful in the design of new algorithms (R1), that our empirical studies were convincing (R1, R2, R3) and that they proactively answer their questions (R3). We address reviewer comments below and will incorporate all feedback.

@R1*"clear examples to explain the concepts"*: Our work stands on the intuition that an agent might benefit from using its computation budget to spread information as fast as possible through the state space to all precursors of the current state, rather than just the one state and action that brought it to the current state on this occasion. An **analogy to human memory consolidation** may help: we are arguably good at retrospectively inferring (or inventing) the underlying causes of our experience and ruminating over them so as to improve our future predictions; ***we hypothesize agents could also use retrospective knowledge about the world (backward models) to consolidate (hindsight planning) their prospective knowledge (value functions)***. We will complement the **thought experiment** in Section 3 with this. Thank you for the suggestion. @R1*"not clear what an abstract model is"*: we define a model as being abstract when we remove inductive biases (e.g. structural constraints) in its construction and use, leaving it as a (learnable) black box. We will expand in the paper.

@R1,R2 *on related work*: The closest work to ours is that of **van Hasselt et. al.**, our results complement theirs: in the control experiment we extend by showing backward planning is *more robust in dealing with rare events (e.g. stochastic rewards) and different levels of stochasticity in the transitions*;@R2 van Hasselt et. al. compare against replay – a non-parametric model-based approach, NOT against model-free learning. **Goyal et. al.** use imitation learning on a generative model's outputs so as to improve exploration by incentivizing the agent towards the high value states on which the model was trained; in contrast, *we aim to formalize and tease apart the fundamental properties of* ***online*** *hindsight planning*. **Satija et. al.** use backward value functions to encode constraints for solving constrained MDPs (CMDPs) with safe policy improvements; though both our work and theirs employ some form of retrospective knowledge, *the contents, purposes and uses differ*. @R4 we will add the *missing citations* and discuss how they relate to our work. Thank you for spotting them.

@R3, R2 *on explanatory details*: **Details on experimental settings are in appendix D**. The prediction experiment is run on a leveled state space s.t. transition dynamics between states generate bipartite graphs; we vary the no. of states on each level and the no. of levels to generate different structural properties. We refer to **fan-in/fan-out** as *the no. of predecessors/successors a state might have in the state space*. Thumbnails depict the phenomena of transitioning from a larger no. of predecessors that "funnel" into a smaller no. of successors, and vice-versa. **Task** $\equiv$ *value prediction*. We will make the main paper more clear by adding more details from the appendix. Thank you for highlighting this.

@R2 *utility of forward planning when the future is predictable*: The claim rests on the prediction experiment; the deterministic control setting is more conflated, performance difference too marginal, task too easy and backward models are also in their best regime. @R2 *codebase*: Missing requirements can be installed from pip and the *import bw_PAML_qt* safely deleted. @R2, R3 *backward planning being less harmful*: Great question! An erroneous forward model that predicts an unreachable state will move the value of a real state towards the arbitrary value of the bootstrapped state; an erroneous backward model will harmlessly distribute the value of a bootstrapped real state to an unreachable arbitrary-value state (see concurrent work Jafferjee et. al. - *"Hallucinating value..."*). @R3*"tethered to a policy"*: Both $\overleftarrow{p}_\pi(x_t, a_t|x_{t+1})$ and $\overleftarrow{p}_\pi(x_t|a_t, x_{t+1})$ are backward models. Eq (3) shows how the latter still **depends on** π **through the stationary distributions** η_π. @R3*"using Bayes rule"* **Yes** (see appendix A). @R1,R2,R3 we appreciate pointing out the misplaced arrow in Fig.1's legend, detailed feedback on typos and suggestions on how to improve clarity.

@R1,R2,R3,R4*"deep RL experiments"*: Deep RL comes with **confounding factors**, e.g. the environment-dependent mixing time characterizing the correlated datastream forces the use of reply buffers to decorrelate experience and target networks to stabilize learning for incremental-update algorithms; this in turn pushes learning in the off-policy regime where convergence is only serendipitous; common testbeds are not informative of how different components of algorithms influence learning. We have left as future work extensions to function approximation and more complex problems, as we reckon (substantial) additional research is required in terms of testbeds and ablations to allow for **scientifically relevant hypotheses**. Backward models came with their own peculiarities and available choices in terms of estimation and use. Understanding and formalizing them was, in our view, **the first step in the right direction**, i.e. transferring to more complex problems in a principled way. **Part of our scientific contribution** was this analysis, revealing many options w.r.t model-learning objectives and planning-strategies (Section 3). The **goal** of our empirical studies was to disentangle core properties of these approaches so as to inform on the design of complex testbeds, where we can understand planning methods. We strongly believe that our contributions **set the stage for principled deep RL investigations**.

We can clarify these points in the paper and strengthen the discussion of existing literature. We feel the technical contribution is sound, surprising (to us at least) and potentially impactful to both scientific understanding and practice.