Comments:   
In three experiments with sequential decision tasks, Cushman and Morris investigate choice with subgoals in humans. Using an elegant design that examines trial-trial adjustments in preference on a set of probe trials that are carefully balanced to isolate the comparison of interest, they show that an intermediate goal is reinforced (ie more likely to be targeted again) when it is chosen and rewarded, but crucially even on trials when that subgoal wasn't actually achieved -- and therefore, according to the logic of the task, the reward should formally have been uninformative about the subgoal's value.   
  
The questions of where goals come from, how "model-based" planning might be simplified, and how habitual and goal directed processes interact are all timely, open and important ones. Moreover, the effect itself is crisp and twice replicated. So I am definitely an enthusiastic supporter of this study overall. But though I am wholly convinced as to the reality of the phenomenon reported -- and that it is likely important -- despite a lot of struggling I am not really sure what is its broader interpretation or how I should expect it to generalize. I believe with further development and revision the article can be clearer conceptually (and perhaps ultimately, experimentally) about how this bias arises and why. Below I offer some thoughts which, despite my admitted confusion, I hope will be helpful.   
  
In setting up and motivating the question, the authors cast a pretty broad net around different notions of planning and hierarchy, and I struggled to discern the specific hypothesis. As I understand it, the hypothesis is that model-based planning (to attain a particular goal) might be nested inside model-free choice, of which goal to pursue. This interrelationship can be motivated by the shortcomings of either system alone. Model-based search can be simplified or pruned to focus on particular subtrees, by identifying goals or subgoals, as in the Boyan backward search idea (and perhaps also Botvinick's saltatory model-based RL using options), where the goal choice itself might be learned by model-free reinforcement. Conversely, a core model-free choice mechanism can be improved by adding what amount to options to its choice set, where these options are identified with subgoals and incorporate model-based planning internally to reach them. The results show that part of the  
reinforcing effect of a large (vs small) reward is mediated by the chosen goal state (in that it transfers to subsequent choice of other actions leading there), and also that the effect is not mediated by the value of the goal state itself, in that it occurs even when that goal state was not attained. These two aspects of the results are taken to support the model-based and model-free halves of the hypothesis, respectively.   
  
As far as I can see, the second conclusion is not demonstrated by the results. While it is true that a "flat" model-based learner mediates goal choice by the value of the goal, from the perspective of the goal-level chooser at the top of the hierarchy, this task amounts to a single-step bandit task -- choose a goal, get a reward (with the intermediate action choices and goal attainment encapsulated within the lower-level planner). Viewed from this level of abstraction, model-based and model-free valuation coincide. To distinguish whether this reinforcement is really model-free, I think a subsequent level of choices would have to follow the initial goal outcome, much as in the original two-step task logic, augmenting the task "from below" rather than "from above" as in Expt 3.   
  
I also thought the computational modeling exercise was a little underwhelming, even for didactic purposes. Instead of embodying and clarifying (what I take to be) the core hypothesis about how MB and MF systems interact, it tacks a third choice system on top of those as previously described. I'd have expected either to see goal-focused pruning incorporated within the modeled MB system, or goal-focused planning incorporated within the MF system, or indeed, both systems to be replaced with one hybrid, nested system, and this would have helped me a lot to understand what the authors were getting at.   
  
This relates to the results and analysis, which don't especially shed light on the question whether this additional goal bias is a third thing, relatively how strong it is, or if it coincides with or exhaustively captures the contribution of either MB or MF processes vs. being a bias within them. To some extent that relates to the parameters w1 and w2 in the authors' model, though I'm not sure that's the most informative way to think about it. Ideally one would like to see full model fits that answer this question by inferring parameters of this sort from all the data -- though I take it the model as written probably has too many moving parts to make this viable. But though I really like how the analysis focuses on the probe trials, which are carefully balanced with respect to revealing this bias, the same mechanisms should be operating throughout the full dataset and it should be possible to verify that they help to account for the ongoing choices as well. At least it would be  
interesting to know how much bigger is the reward effect from the probe trials, vs the facilitation you would see in otherwise the same circumstances if the chosen goal had been actually attained or if the same actions were offered again instead of the equivalent ones. (I think these conditions might indicate the additional effect, if any, of regular, action- rather than goal-focused MB or TD-1 processes?)   
  
A few minor thoughts:   
  
If I understand the concern motivating the third experiment -- i.e. that the seemingly model-based nature of the planning to goal might actually arise via some sort of associative spreading to the chosen goal's other associated actions, followed by MF reinforcement -- I don't understand why the experiment would address this. Couldn't the family of preceding actions A/B/C/D also be activated in the same way to the same effect? Although I don't see this as a fatal concern, I also don't see that this question is addressable behaviorally (though Wimmer and Shohamy and Doll et al have gone after it with neuroimaging), since it ultimately comes down to a question about when, prior to the actual choice, does the computation happen.   
  
I don't really understand the analysis partialling out MB and MF value from the test of interest. I'm not too concerned about this -- since the probe trial design should balance out all competing factors anyway -- but the analysis as given is not explained well and not especially convincing. Contrary to the computational model, these are apparently based on a single preceding trial, but "discounted" by delay which is both unclear mathematically (what is the form of this?) and undermotivated empirically (shouldnt the amount of this discount be estimated?) Relatedly the key result in Expt 2 is the interaction, not the comparison of one effect being significant and the other not, but this test is only presented for a subset of the analyses I think.   
  
I wonder if part of the seemingly goal related effect in this study relates to subjects just ignoring the bottom level state identities once they have learned the transition rules. This is another reason why it might be interesting to know if the effect persisted if they had to pay attention to where they ended up (eg because they had to make another state dependent choice there), and again whether it might be interesting to know if there is positive evidence to what extent subjects are indeed assigning value to the goal states in the usual Markovian way.