Why this model makes sense:

* Just talking about habits of thought
* Tons of evidence that executive goals can be activated automatically by contextual stimuli
  + Implementation intentions
* Fits nicely with MF gating of working memory
* Could explain confounding of state & reward signals in striatum
* Also, supports hierarchical RL
  + How else could you choose goals?

What have people done?

* CS algorithm that incorporates the two (Dyna-2)
* Pruning trees

What are we doing that’s new?

* Most models portray the two systems as competitive
* Nobody’s actually formally modeled habitual goal selection
* And nobody has shown it using the precise behavioral experiments designed to tease apart MB from MF

What does this explain?

* Properties of goals: automatic contextual activation, subliminal reinforcement, intrusiveness, “functional fixedness”
* Bigger picture stuff
  + Acquiring cognitive skills
  + Addiction
    - Rather than just habitual action vs goals, the habit actually hijacks the goal-directed system
  + Implementation intentions
  + Doctrine of double effect

Future directions

* fMRI
* Realistic scenarios

**Objections**

* Isn’t it obvious *a priori* that goals depend on System 1? How else could they make value judgments about outcomes?
  + System 2 can represent value also
    - It just represents it as the value of outcomes, not of actions
  + Also, even if that were true, it’s a different kind of dependence
* Computational specificity… How exactly does our proposed MF on MB system work?
  + Really simple
  + Somehow has access to goal state (maybe in WM?)
    - Technically, we had it *infer* goal state from action
    - Could also have direct access (i.e. what if goal doesn’t lead to action?)
  + Predicts the value of each goal state
  + Calculates a prediction error after each round, updates value