Type of Submission  
  
   Artificial Intelligence  
   Psychology  
  
The Review  
  
   Important question with a sensible experiment and clean data analysis. R2  
   & R3 both make a number of observations that I urge the authors to  
   address.  
  
  
------------------------ Submission 748, Review 1 ------------------------  
  
Title: Causal learning from interventions and dynamics in continuous time  
  
  
Type of Submission  
  
   Cognitive Science  
   Psychology  
  
The Review  
  
   This paper presents an new experimental paradigm and task which enables  
   the investigation of an exciting and (to my knowledge) previously  
   intractable field in causal reasoning: cyclical causal dependence graphs.  
   The new paradigm is continuous-time causal intervention; this lets  
   subjects learn about cyclic systems because subjects' interventions (and  
   thus the resulting effects) may be arbitrarily structured in time. The  
   paper compares human learning of both cyclic and acyclic systems, and  
   implements and evaluates several potential computational models against  
   subject data. Overall, excellent work.  
  
   Comments:  
  
   - I found it much easier to understand the task and its advantages by  
   trying the experiment myself; the authors might emphasize it more.  
  
  
------------------------ Submission 748, Review 2 ------------------------  
  
Title: Causal learning from interventions and dynamics in continuous time  
  
  
Type of Submission  
  
   Cognitive Science  
   Psychology  
  
The Review  
  
   This describes a study of causal learning in situations where people can  
   initiate events and there are stochastic delays between causes and their  
   effects. It considers both cyclic and acyclic causal systems.  
  
   The results suggest that in settings like the current experiment:  
  
   \* People tend to intervene preferentially on root causes/causes with more  
   effects (consistent with some past results).  
   \* People are reasonably good at identifying causal structure, but not as  
   good as a Bayesian ideal observer with correct priors.  
   \* Some participants' inferences respect the actual distributions of  
   delays between causes and their effects, rather than a recency bias in  
   causal attribution.  
  
   The experiments appear to be reasonable in their construction, the models  
   are easy to understand and shed light on participants' behavior,  
   the paper is well-written, and the results are new and interesting. I  
   think it would be appropriate for this work to be presented as a talk at  
   cogsci.  
  
   That said, if the paper is accepted, there are a few issues that would be  
   good to address in a final version:  
  
   \* I doubt I’m alone in being skeptical that the training is sufficient  
   to give participants a clear understanding of the distribution of delays  
   between causes and their effects. If some participants have strong  
   expectations that delays will be short even after the training, that  
   might explain the large number of participants who were best fit by the  
   "order-only" model, which is compatible with the inferences of a  
   time-sensitive learner who expects delays to be short.

**Sure but that’s not changing for the cogsci paper and doesn’t seem worth highlighting**

   \* Why not use a Bayesian ideal observer that incorporates uncertainty  
   about the delay distributions themselves? The only reason I can see is  
   that it would be computationally expensive.

**That’s why.**

   \* The statistics should be double-checked; I believe the p-value for  
   F(1,38)=4.6 is greater than .001  
   (second paragraph in "Accuracy" subsection).

**Yup, well spotted should have been 0.04!**

   \* There are several claims like "there was no relationship between  
   entropy and the variability of either  
   interval time. How were these non-relationships established? What were  
   the numbers?

**Changed to “no evidence for a” [I mean there wasn’t a significant relationship], there and in one other place**

   \* Is it possible that lack of motivation or attention is a common cause  
   of closely-spaced interventions and poor performance? Some participants  
   could exhaust their interventions for a trial, guess at a structure,  
   switch to a different HIT they’re doing simultaneously, and return some  
   time after the timer runs down.

**Well they were properly incentivised, but sure it may be part of what is going on, don’t think this means changing anything.**

   \* I'm not aware of likelihood ratio thresholds of 20/1 as being part of  
   "standard significance testing". Perhaps the authors could clarify what  
   they mean -- there's a risk that readers could think they're interpreting  
   p-values in an unconventional way.  
  
**Removed it.**

------------------------ Submission 748, Review 3 ------------------------  
  
Title: Causal learning from interventions and dynamics in continuous time  
  
  
Type of Submission  
  
   Cognitive Science  
   Psychology  
  
The Review  
  
   This work examined how humans effectively employ interventions to  
   construct and test causal structures in continuous time under dynamic  
   situations, in which delayed effects are involved. The research developed  
   a Bayesian model and found a discrepancy between human performance and  
   model predictions, and then moved to the comparisons with a few heuristic  
   inference models.  
  
   This is a well-written paper, and should be of general interest for the  
   broad cognitive science audience. Here are a few comments to hopefully  
   help improve the paper in the revision.  
  
   When comparing human performance with an ideal Bayesian model, the focus  
   is how accuracy varies as a function of the manipulated IVs  
   (reliable/unreliable delay, acyclic/cyclic) and some details on timing  
   of interventions and the preference for positive testing. However, for  
   heuristic models, the comparisons just focused on overall accuracy.  
   Although the average performance of these heuristic models is close to  
   that of humans, it is unclear if a similar pattern of results held for  
   comparisons across different experimental conditions. It would be good to  
   see a summary showing how model performance in the various conditions  
   matched/unmatched to human performance.

**>>Agreed, but mostly for the full paper. Managed to squeeze in 1 extra sentence (heuristic models also suffer under cyclicity like participants).**

   I also would like to see some model details, specifically how the  
   likelihood term is defined. A gamma distribution was used to model when  
   an effect could occur after activating the causal variable, and the  
   causal strength w was used to capture the probabilistic causal relation.  
   How these two distinct components are combined in assessing the  
   likelihood probability is unclear. Did you use the mixed distribution for  
   the likelihood calculation? Did you use noisy-or function to combine  
   causal influences?

**>>Hmm, no and no because they are point events. We already say”**

***The likelihood of the data given a specific path $\zz^\prime$, then, is the product of the (Gamma) likelihoods of the observed delays and causal strength $\ws$ combined with the likelihoods of (non-)events, the occurrence of which failed either due to the $1-\ws$ causal failure rate or due to the effect potentially occurring after $\tau$ (i.e., some time in the future).***

**Can’t see an easy addition without getting deep into it.**

   Finally, some discussions on the generalizability of the results would be  
   useful. The study used a fixed causal strength of .9, and no spontaneous  
   activations of any components. Would the findings generalize to more  
   ambiguous situations, such as lower causal strength in the presence of  
   non-zero spontaneous activations? In addition, what role would delayed  
   effects in time play in yielding the significant discrepancy between  
   human and Bayesian model performance? If there is no time delay, would  
   human performance be better accounted by a rationale model than heuristic  
   models?  
  
   Minor points:  
   (1)  In figure 1, it is unclear what white circles mean. The connection  
   between the left and right figures is also unclear. It would be helpful  
   to explicitly annotate the causal variables (A, B, C) of the causal model  
   in the right plot.

**Improved the caption and increased size of row labels (don’t wanna label the activation white circles: feel it would be misleading since components were unlabelled in the exp).**

   (2)  Would the model(s) predict the differences between figure 1A and 1B?

**Good question, will check for journal but no space in paper.**

   (3)  The first sentence in the section “Normative inference” is hard  
   to follow. I suggest breaking it into shorter sentences. For the second  
   sentence of the same paragraph, you may want to add the time for the  
   example: “performing interventions on components A and B at 100 ms and  
   1200 ms, respectively”.

**Done**  
   (4)  It seems that the Bayesian model did not predict any difference in  
   accuracy between acyclic and cyclic conditions. This is a bit surprising,  
   since the complexity of the two conditions would be quite different—any  
   intuition why?

**Yes: more evidence total, but more ambiguous too. I think we say this in GD already, couldn’t find space for an addition.**