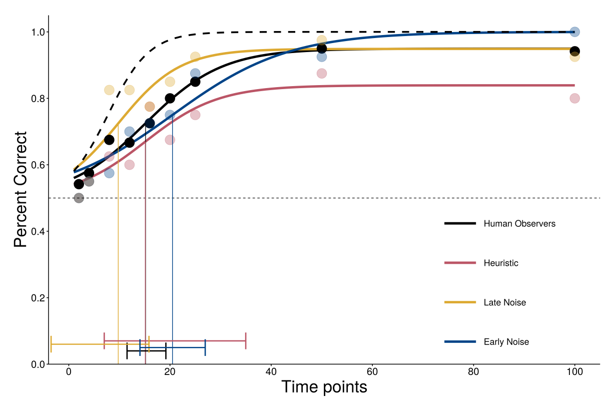
We would like to thank all reviewers for their valuable feedback and we very much appreciate their assessment of our work as “very high quality" (R1), “top 15% of accepted NeurIPS papers” (R2) and “well-thought experimental setup” (R3) in studying causal inference in humans. We will make sure to address all the minor issues raised by the reviewers in the final version of our paper. In the following we reply individually to the main issues raised:

R1: *I would like to see some explanation of how people solve the task using the models and ecological validity.* We agree that our visual system has not evolved to discriminate “in-time” from time-reversed events. However, we can learn a lot about the inner workings of a cognitive system by probing it with *appropriate* artificial—not ecologically valid–stimuli (RUST & MOVSHON, *Nature Neuroscience*, 2005; MARTINEZ-GARCIA ET AL, *Frontiers in Neuroscience*, 2019)—this is not to say we should *only* use simple, artificial stimuli, but there is a place for their use, particularly when studying less well known areas—such as the human visual system’s sensitivity to subtle temporal dependencies. In a predictive coding framework, e.g., it would be useful to know the exact temporal statistical structure of e.g. the motion of leaves and grass in the wind. An unusual motion pattern—e.g. having the “wrong” dependencies—may signal a hidden predator behind the foliage. Thus whilst we agree with reviewer R1 that the exact experiment of a moving single disk lacks ecological validity, the implications do not—we see this as the beginning of explorations into subtle temporal dependency structures in biological motion in general, causal inference abilities, leader-follower behaviour etc. Finally, we fitted three (!) more ecological valid models to the data, see discussion of R3’s comments and the figure. Obviously we will attempt to make our reasoning clearer in the revised version of the manuscript.

R1: *It’s not clear to me why the particular time-series and/or the noise parameterization are generalizable to new situations or special in some other manner*: It is only possible to classify time series in our setting with non-Gaussian noise. One measure for non-Gaussianty is kurtosis. Our parameterization yields Gaussian, platykurtic (bimodal) as well as leptokurtic (super-Gaussian) noise from the same equation. Thus, we belief we cover a broad range of non-Gaussian noise, and we can control the degree of non-Gaussianity.

R1: *I would have preferred to see a comparison to a simple recurrent neural network.* R1 is of course correct that the network gets as input the entire sequence which is different to humans. However, the network starts with a convolutional layer of size 10 which effectively slides over the time series. Thus we think that despite the seemingly very different nature of the inputs, de facto our scenario is somewhere in between a sequential RNN and a fully connected layer in which the network could use the evidence of the entire time series as a first step.

R2: Thank you very much for your review, we highly appreciate it.



R3: *Fitting a suboptimal Bayesian observer model to the data* : Thank you for pointing out this excellent paper to us. We agree that this is indeed a very promising way to extend our work. Based on the proposed paper, we test three additional strategies, shown in the figure on the right (dashed line is the ideal observer from the original submission). First, we fitted a noise term additive to the decision variable (Model 2 in Stengård and Berg). This corresponds to late noise in the visual pathway. Second, we fitted an additive noise term to the individual time series before calculating the decision of the ideal observer. This corresponds to noise in the early visual pathway and uncertainty about the exact location of the disk (e.g. micro-saccadic eye movements). Third, we have fitted a (ecological valid) heuristic to the data in spirit similar to those proposed by Stengård and Berg. This heuristic needs only a few lines of code, but it yields results close to that shown by our observers. We will add a discussion and comparison of the three new algorithms to our paper, in particular with an analysis of their consistency using frozen noise (Fig. 3 of the original submission).

R3: *Approximation used for the Bayesian likelihood distribution* We agree that it would be better to use the full likelihood. Estimating the stationary distribution $p(x_1)$ is good advice which we will of course implement. In addition we will assess if and how we can estimate or approximate the other 3 conditional distributions given that the time series values are not independent, we skip the first 400 terms and numerical estimation of the multidimensional (conditional) distributions could be challenging.

R3: *On a technical point, how did the authors obtain "frozen noise"?:* As you thought, we fixed the seed for the random number generator, but in addition—as a cautionary measure, as we worry the way you do—we saved all time series which were generated during the experiments. Thus we can confirm (ensure) that all observers and algorithms classified exactly the same time series.