We thank the reviewers for their comments.

Reviewer #1:▷ **Role of each term**.

We actually showed the unique effect of the separation loss over the other terms in suppl. Fig. S1 and S2.

Fig. S1 shows the results without the separation loss; Fig. S2 shows how the results change after the introduction of separation loss during training. We, therefore, respectfully disagree with the reviewer's comment that "it is unclear how much 3 contributes over 2". This is also pointed out in the main text in lines 177 and 178. We could move this material to the main text in light of the reviewer's comments. The other two terms in the loss function are similar to the standard terms taken from Tolstikhin et al, 2017 cited in the paper, which is why we didn't establish their usefulness here (the KL term experimentally improves optimization, but we do not consider it as our contribution here). Further, RNNs without hyper-networks are studied elsewhere (Dezfouli et al 2018a) in the same BD dataset that we used here, so their relative performance is known (they showed important aspects of decision-making remained uncaptured by typical computational models and even their enhanced variants, but were captured by RNNs automatically).

▷ **Comparison with other works**. Without the separation loss, the autoencoder framework is equivalent to Tolstikhin et al, 2017 (without RNN and hyper-net), which as we showed (Fig. S1; S2) does not disentangle effectively. Therefore we are indeed comparing our framework with this previous work.

Please also note that since the latent space has only two dimensions, we were able to directly visualize/report one-to-one relationships between each latent variable and each factor of variation in the data (Fig 2a, S1a). Following the reviewer's comment, we calculated the disentanglement metrics reported in disentanglement_lib (Locatello et al, 2019) based on the results in the synthetic dataset obtained by including the separation loss, and without including the separation loss. In all the metrics the separation loss improved disentanglement (with separation loss > without separation loss): MIG: 0.29>0.11, DCI: 0.19>0.03, SAP: 0.15>0.06, β-VAE score: 1>0.99, factor-VAE score: 0.94>0.68, modularity: 0.99>0.87. Please refer to Locatello et al, 2019 for the meaning of each metric.

▷ **Quadratic cost w.r.t number of latent dimensions**. The utility of our method depends on the number of latent *not* growing with the number of subjects. Current experiments are using >1000 sequences, which is considered to be on the high-end of the number of subjects in psychological/neuroscience studies.

Reviewer #2:▷ **Limited novelty compared to previous works**.

Our aim is NOT to represent or classify behavioural trajectories as a generic time series (#1, for short), but to characterize differences in the (typically causal) processes underlying reinforced choice (#2).

These aims are very different. From the references cited, the reviewer might be under the misapprehension that we are solving #1. For instance, Johnson et al 2016 build an interpretable representation of movements of a mouse in an experiment (# 1). This framework is NOT able to extract how the individual differences in such movements can be explained in a low dimensional space (#2). The same applies to Fox el al, 2011. Indeed, these two frameworks are functionally equivalent to the learning network in the current architecture. As such, we believe that the aims and the architecture of the current framework are quite different from the previous models on disentanglement on general time-series. We can, of course, discuss theses references in the paper and explain their differences with our model.

▷ **Usefulness of RNNs for modelling human learning processes**. On the encoder side, the reviewer suggests using PCA instead of an RNN to extract the features of the learning processes. Even on the very same dataset (BD), it is shown that both linear models and more complex cognitive characterizations fail to learn the complex patterns in human behaviour (Dezfouli et al 2018a; cited in the paper, which has now been published in PLOS Computational Biology). Given the constraints of disentanglement, we fail to see the merit of employing a weaker technique.

▷ **Ultimate aim is classification**. We are working in an unsupervised setting with the aim of characterizing individual differences. Psychiatric classification is notoriously crude; we just used them in the BD dataset for coarse validation.

▷ **Comparison with previous work**.

We do show that the previous autoencoder architecture (Tolstikhin et al 2017) fails to produce desirable disentanglement results here and our new separation loss is required. Please see Figure S1 and S2.

Reviewer #3: Apologies for the short response.

▷ **Separation loss for options**. In principle the separation loss can be used over the space of options instead of actions, which could be interesting since it allows analyzing high-level strategies.

▷ **Intuition behind equation 10**. The tightness of equation 10 is intuitively related to whether the direction of the effect of z_1 and z_2 on behaviour depends on t. For example, in the simulations here changing z_1, z_2 affects action probabilities in the same direction (increase or decrease) across different time steps, which makes the approximation more accurate. We plan to derive more formal results about this approximation in future works.