by optimizing statistical information criteria, which are theoretical approximations to using the held-out data in a cross-validation pipeline (?). We describe how model selection in these two cases is performed.

6.1 Selection using an Interpretability Score

For a given interpretability test T, set of models $\mathcal{M}$, and data set D, we aim to find the models that achieve the best interpretability scores (IS).

$$M_S^* \in \text{argmax}_{M \in \mathcal{M}}(IS(M, D)) \quad (2)$$

Section 7 describes the design and results of a user study that is used to find the best model according to the interpretability score.

6.2 Model Selection using Statistical Theory

Ideally, the model parameters would be optimized on held-out data using predictive log-likelihood as the objective (?). However, the difficulty of collecting controlled sessions of student interaction in CW meant we had few data instances available. To address this challenge we use statistical information criteria as a theoretical approximation to the predictive accuracy of a model (?). Specifically, we use the following information criteria: the Deviance Information Criterion (DIC) and Watanabe-Akaike Information Criterion (WAIC).

Let $\log P(D \mid \hat{M})$ be the log-likelihood of the data given a model $\hat{M}$ with parameters set to the mean of the posterior estimates. The DIC is defined as follows:

$$DIC(M, D) = -2 \log p(D \mid \hat{M}) + 2 \cdot c_1 \quad (3)$$

where c_1 is a penalization term[6] that depends on the expectation of the log-likelihood of the data given M. In practice, this expectation is the average of $\log P(D \mid M)$, where M corresponds to parameters that are obtained via the posterior distribution samples from the Gibbs sampler.

The WAIC has the same structure, although it does not require $\hat{M}$:

$$WAIC(M, D) = -2 \log E[P(D \mid M)] + 2c_2 \quad (4)$$

here, the penalising factor c_2 subtracts $E[\log P(D \mid M)]$ from $\log(E[P(D \mid M)])$. Again, these expectations are computed using Monte Carlo estimates from the individual posterior samples from the Gibbs sampler.

Model selection is performed by assigning IC to be $WAIC$ or DIC respectively and minimizing Equation 5.

$$M_C^* \in \text{argmin}_{M \in \mathcal{M}}(IC(M, D)) \quad (5)$$

Figure 6 shows the two information criteria plotted as a function of the model (the random model has no notion of information criteria and so was not compared here). The data set comprised of both of the log files of students' interactions (8 minutes each). The optimal model for both DIC and WAIC is the MK_5 model but we note that MK_1, MK_5 and MK_{10} all perform close to this optimal setting. Notice that the fully Bayesian model (FB) is not optimal but it is in the top 5 models for both criteria.

[6]Exact formulae for the DIC and WAIC penalization terms can be found in **?** (?) in the section starting at pg.169.

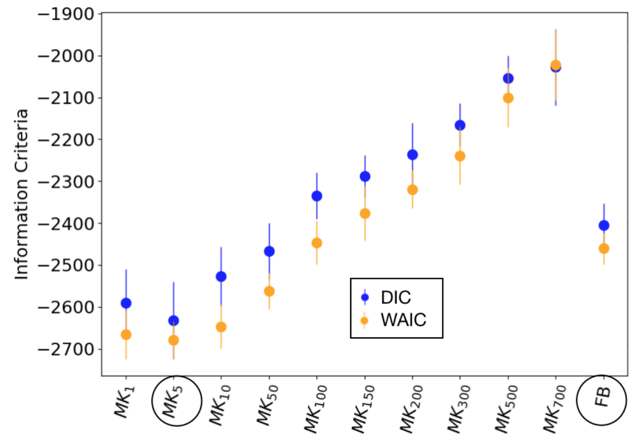

Figure 6: DIC and WAIC as a function of the model (lower on y-axis is better). The MK_5 model is optimal, the FB approach is in 5th place.

7 User Study

In this section we describe a user study that compares the interpretability of different models for describing the responses of the simulation to students' interactions in CW. The set of models includes the 12 CW models described in Section 5.

We recruited participants from two cohorts: undergraduate engineering students in a large public university and Mturk workers (with a total of 240 people who participated in the experimentation). For a given data instance, we randomly sampled a set of 12 time points, which remained constant across all model conditions. Each time point generated 2 experiment trials for each model, making $2 \times 12 \times 12 = 288$ trials per data instance. The reason for 2 trials per time point is to select both the forward and backward intruders (in time) for each selected candidate period. Each participant saw 20 randomly sampled experiment trials, with no more than 2 trials from any given model, to ensure a representative range of models. After making their choice, participants received brief visual feedback on whether or not their selection was correct.

All participants received a detailed tutorial about CW and the study, as well as a pre-study comprehension quiz[7]. Mturk workers were paid a base rate of $\$0.25$ for participating and a bonus structure of $\$0.1$ for each correct response.

We first describe results in terms of accuracy (the percent of correctly labelled test instances). The top performing model was MK_{200} with an accuracy of 83% on the Forward Simulation test and MK_{100} with an accuracy of 82% on the Binary Forced Choice test. The random baseline model performed consistently poorly with an average accuracy of 53% on both tests. The fully Bayesian model achieved an accuracy of 72% and 70% respectively on the two tests.

To control for ordering effects, chosen time periods, data instance used, and effects of individual participants, we applied an L2 regularized logistic regression for predicting the user specific success on the experiment trial, shown in

[7]Tutorial pdf slides are available in the supplementary material.

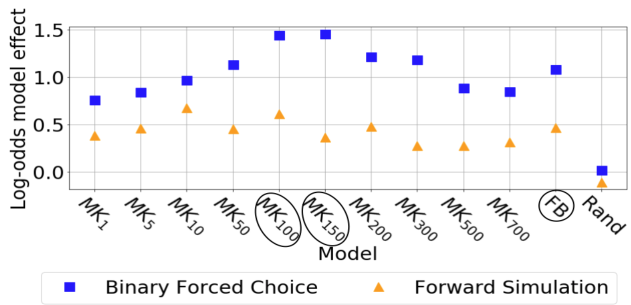

Figure 7: Effect of each model on the log-odds of a test evaluator selecting the correct response (controlling for the test evaluator, the experiment trial, log file and ordering effects).

Figure 7. The y-axis presents the improvement in log-odds that a model has on the expected response accuracy (higher is better). As shown by the figure, the Forward Simulation shows a high variance with no clear maximum. In contrast, the Binary Forced Choice test has a clear maximum in the region of MK_{100} and MK_{150}.

From both Figures 6 and 7 we can infer the following four conclusions. First, all of the models $(MK_1, \ldots, MK_{700}, FB)$ outperform the random baseline: participants are more likely to select the correct response from any of these models. This result suggests that periods of stable dynamics exist in the data and that it is possible to construct models, which describe these dynamics, that are interpretable to people.

Second, the Binary Forced Choice test is a preferable measure for interpretablity to the Forward Simulation test. Figure 7 shows that the Binary Forced Choice test exhibits a clear peak (around MK_{100} and MK_{150}) where interpretability of the model is maximized. These models also maximized the raw accuracy on the Binary Forced Choice test.

On the other hand, the Forward Simulation test has a greater variance across models and across data instances. Two possible causes for this higher variance are: (1) there is more room for error in the Forward Simulation test (5 choices vs. 2 choices in Binary Forced Choice); (2) sampling a single image to represent a period (as in Forward Simulation) presents less information to the user than sampling 4 images (as in Binary Forced Choice).

Third, the best κ settings vary for different tests and information criteria. Model interpretability grows steadily as the value of κ increases, with MK_{100} and MK_{150} being the optimal models, and then proceeds to decrease steadily. These models are not consistent with the model MK_5 that optimized the information criteria. Note that higher κ values are "sticky" - they bias the model towards longer periods, which condense too many activities to make sense to people. On the other end of the spectrum, lower κ values allow for shorter periods that capture too much of the noise in the system. In contrast, the κ value for models MK_{100} and MK_{150} represent a "sweet spot" in between these two extremes.

Finally, the fully Bayesian model (FB) performs consistently well on both information criteria and interpretability tests. It is interesting to note that while this model does not find the optimal setting (from neither the statistical information criteria nor from the human interpretability task) it does perform well across all tests, tasks and instances, and is fully automated (no human evaluation is required in order to choose an optimal parameter setting).

We conclude this section with mentioning the limitation that the user study was based on a small number $(n = 2)$ of instances. This was due to the difficulty in obtaining controlled sessions of student behavior in CW. Despite this issue, the differences between the models in Figure 7 are statistically significant, having being evaluated across 12 different time points for each instance and with hundreds of evaluators.

8 Conclusion & Future Work

With the growing prevalence of immersive simulations so arises the need for AI systems which help end-users gain insight into the activities of the participants. We have studied an environmental simulation where students learn about the causal effects of their actions. Our results show that algorithms can segment time series log data into periods that are coherent for people. However, selecting hyperparameters in these models is a challenge, especially when trying to optimize the representations for their interpretability. We have shown an example of how to select these hyperparameters from two tests that are grounded in the literature and we have further presented the fully Bayesian method as promising technique for implementing a model when human evaluations are not possible. Future work will apply these models to alternative domains and will work with teachers and experts "in the loop" such that we can target the goal of engaging the participants with insights drawn from their own experiences with such immersive simulations.

Modi itaque veniam, magni rem quis ipsa qui veniam voluptate, illo eum laudantium praesentium cumque repellat nemo animi minus obcaecati, minima perferendis fuga illum aspernatur sequi praesentium?Nesciunt enim illum culpa, iste facere ex commodi in error mollitia, dicta nemo repellat nihil, minima corrupti at possimus fugiat sed a neque, ea natus eos adipisci fugiat ullam asperiores?Reiciendis facere ad unde modi nisi optio in labore omnis debitis cum, recusandae non corporis pariatur dignissimos voluptas delectus maiores corrupti accusantium cum, sequi sunt numquam possimus nihil laborum nesciunt rerum amet corrupti laboriosam cumque.Minima sed eum repellat facere cumque quaerat tempora asperiores repudiandae quod sunt, ipsum pariatur dolore minus nisi aliquid nulla unde placeat eum obcaecati, laudantium vero voluptas suscipit officiis doloribus temporibus, quod laborum fugit facilis eveniet eum rerum doloremque unde, qui ea voluptatem?Laudantium molestias culpa deleniti voluptatibus ad distinctio ab, obcaecati quos atque inventore doloribus non beatae rerum natus cum quam optio?Ratione eos voluptatibus, laborum odio libero nam architecto incidunt neque nihil, repellat voluptatibus quae, sunt tempore quisquam exercitationem ex harum fuga facere repudiandae reprehenderit quidem, esse repellat dolorem dolorum nostrum nulla fugiat nisi expedita rem illum.