Human-Grounded XAI: An Evaluation of Explanation Faithfulness and Intelligibility For Interpretable Neural Networks

Juan D. Pinto[1,*], Luc Paquette[1]

[1]University of Illinois Urbana-Champaign, Urbana, IL, USA

Abstract

This paper presents a human-grounded evaluation of a constraints-based approach to fully interpretable neural networks for detecting learner behaviors. We designed and administered a test consisting of forward simulation and counterfactual simulation tasks. Participants achieved high accuracy on both tasks, independent of their level of experience with machine learning, suggesting that the model's explanations are both faithful and intelligible. We discuss our evaluation design, some challenges we faced, and implications for explainable AI in education.

Keywords

Explainable AI, explainability evaluation, interpretable neural networks, model transparency

1. Introduction

In an influential keynote address, Baker [1] presented a series of six critical challenges facing researchers in learning analytics, educational data mining, and AI in education. He proposed one possible way to address the third challenge—the "challenge of interpretability"—by having participants with different levels of expertise attempt to predict the behavior of a complex model based on its explanations. High agreement on this task, he argued, would indicate that the explanations are effective at conveying the model's reasoning, thus making it interpretable.

In this paper, we present our successful attempt to meet this challenge. We derive explanations directly from the parameters of a convolutional neural network (CNN) trained to detect gaming-the-system (GTS) behavior in an educational environment and evaluate them using two different human-grounded tasks. An analysis of our results shows that participants were able to accurately predict and strategically alter the model's behavior. This is an important step towards robust evaluations of eXplainable AI (XAI) in education.

2. Background

The pursuit of explainability has become increasingly critical, particularly in sensitive domains such as education where decisions made by models can have significant impacts on learners. While complex machine learning models offer powerful predictive capabilities, their inherent opacity often raises concerns regarding fairness, accountability, and the potential for pedagogical insights to be obscured. This has spurred research into developing educational models that are understandable by various human stakeholders [2].

2.1. Evaluating explainability

The development of explainable models necessitates robust methods for evaluating their explainability. While model accuracy has well-established metrics, assessing explainability is a less mature task that often depends on the context and needs of the target end-users [3].

Towards this end, Doshi-Velez & Kim [4] proposed a framework categorizing evaluation methodologies for explainability into three categories. *Application-grounded evaluations* test explanations in real-world tasks with end-users, offering high fidelity but often being costly. *Functionally grounded evaluations* use proxy tasks without human involvement, measuring aspects like model sparsity or explanation simplicity as proxies for intelligibility. This paper focuses on *human-grounded evaluation*, which involves real humans performing simplified tasks to assess how well they understand and can use the model's explanations.

Central to evaluating explanations are the criteria of *faithfulness* and *intelligibility*. Faithfulness refers to how accurately an explanation reflects the model's internal reasoning, while intelligibility refers to the ease with which a human can understand the explanation [5]. Depending on the use case, a high level of both can typically be considered a prerequisite for an explanation to be truly useful.

2.2. A constraints-based approach to interpretable models

With the goal of achieving faithful and intelligible explanations, we developed a novel *constraints-based approach* to creating fully interpretable neural networks [6]. Our methodology departs from traditional post-hoc explanation methods by integrating interpretability directly into the model design process.

The model was specifically trained to detect GTS behavior, a form of student disengagement where learners exploit system properties to achieve success without genuine understanding. This choice was strategic, leveraging existing expert-defined features and models for GTS detection [7].

Our approach aligns with the theoretical framework for AI explanations proposed by [8], which defines explanations as the output of an explicit interpretation function performed on evidences derived from a model's parameters. Our model is designed such that its convolutional filters serve as direct evidence, with an inherent interpretation as sequential behavioral patterns. This design ensures high *explanatory potential*, meaning that the extracted evidence (the filters) fully accounts for the model's predictions at inference-time [6].

While we previously conducted a functionally grounded

HEXED'25: 2nd Human-Centric eXplainable AI in Education Workshop, 20 July, 2025, Palermo, Italy

*Corresponding author.

jdpinto2@illinois.edu (J. D. Pinto); lpaq@illinois.edu (L. Paquette)

https://jdpinto.com (J. D. Pinto)

0000-0002-2972-485X (J. D. Pinto); 0000-0002-2738-3190 (L. Paquette)

evaluation of our model's interpretability [6], this paper presents a human-grounded evaluation, which is crucial for more accurately assessing the faithfulness and intelligibility of the explanations generated by our constraints-based approach.

3. Methods

3.1. Questionnaire design

We designed a series of problems to evaluate how well participants were able to understand our model's inner workings using our explanations. We implemented these problems in the form of a questionnaire that participants could complete online.

The structure of the questionnaire is as follows: (1) consent form, (2) demographic questions, (3) description of the model, its explanations, and the tasks, (4) one practice problem with debriefing, (5) forward simulation task, comprising of five problems, (6) counterfactual simulation task, comprising of another five problems, and (7) an optional form to receive compensation for completion. We estimated that the entire questionnaire would take approximately 20–60 minutes to complete.

The questionnaire was designed to be accessible to a wide audience, including those with no prior experience in machine learning or AI. When presenting a brief textual introduction to GTS behavior and our model, we also introduced participants to the visualization format used in our explanations, which consists of a grid representing each of our model's convolutional filters. Because our model only has a single convolutional layer, and due to the nature of our data, each column in the grid corresponds to an action step (a point in time at which data was collected), and each row corresponds to a feature of the model. Our model purposely uses a small set of expert-defined features and convolutional weights are constrained to be binary, culminating in a grid that is easy to interpret.

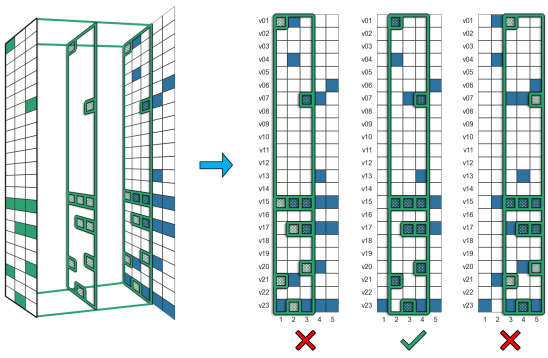


Figure 1: Figure from the questionnaire explaining the nature of the model's convolutional filters (blue) and inputs (green).

We also represented the model's inputs in the same format, overlaying the input grid on top of the filter grid to help participants understand how the model's predictions are made. If all positive features in the input (green squares) are also positive on any of the filters (blue squares), then the model will give a positive prediction. If none of the filters matches the input, then the model's prediction will be negative. Figure 1 shows an example included in our questionnaire, demonstrating the sliding nature of the convolutional filters. The entire text of our questionnaire's introduction to the explanation format is included in Appendix A.

We asked participants to complete two tasks: a forward simulation task and a counterfactual simulation task. The forward simulation asks participants to predict the model's output given a specific input—this was the example proposed by [1]. For each of the five forward simulation problems, we provided participants with visualizations of the model's filters and a single input visualization. Each filter had a number assigned. Participants were asked to select whether the model would predict "Not GTS" (no match) or "GTS" and the specific number of the matching filter from a drop-down menu. To make the visual inspection less cumbersome, we also provided an interactive explanation with a slider that allowed participants to cycle through each filter, one at a time, and to move the overlaid input grid left and right, simulating the sliding nature of the convolutional filters (see Figure 2).

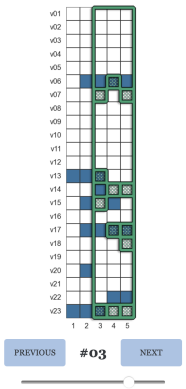


Figure 2: Interactive explanation as presented to participants.

Following the forward simulation problems, participants were presented with the counterfactual simulation task, consisting of five problems in which participants were asked to identify a specific change to the input that would alter the model's output. The way the explanation was presented remained the same, but participants were given the model's prediction (either "GTS" or "Not GTS") and the drop-down menu consisted of five possible changes to the input, such as "remove v15 at action 5", "add v21 at action 4", "remove v7 at action 2", "remove v23 at action 4", and "add v12 at action 1".

For each of the two tasks, we progressively increased the number of filters presented in each problem, starting with 8 filters in the first problem, increasing to 16 filters in the second and third problems, and finally increasing to 32 filters in the last two problem. While the task itself remained the same, this design allowed us to assess the impact of increasing apparent complexity on participants' performance. We say "apparent" complexity because we expect this change to primarily impact the psychological perception of being presented with more possibilities, as well as the time required to go through the entire set. In actuality, the task remains the same. In other words, both tasks are sequential (go through each filter one at a time), so while the length of the sequence may increase, the cognitive load remains constant with no additional parallel processing required.

In addition to participants' answer to each problem, we also asked them to rate their level of confidence in their

answer. We used a 4-point Likert scale for this, with the options: "not at all confident", "not very confident", "somewhat confident", and "very confident".

3.2. Data collection and analysis

We administered the questionnaire online via Qualtrics, a web-based survey platform. We recruited participants through three venues: the International Educational Data Mining Society mailing list, the International Artificial Intelligence in Education Society mailing list, and the Learning Engineering Google Group. We also provided a small monetary compensation for participation.

We collected a total of 222 complete responses. Unfortunately, we did not anticipate the amount of obviously fake responses, with participants completing the questionnaire multiple times very quickly (in one particular case, many dozens of times). The irony of this outcome for a study on GTS behavior was not lost on us.

To address this, we analyzed the distribution of time taken to respond to each problem. We identified a gap in the median distribution at around one minute and set this as a conservative minimum threshold for valid responses. Realistically, we estimated that participants who were genuinely trying to solve the problems accurately would generally take longer than this, but we wanted to err on the side of caution. We then removed all responses that did not meet this threshold, resulting in a final sample size of 36 "serious" participants.

Out of these, 13 participants were graduate students or postdocs, 6 were non-university researchers, and 5 were tenure-track faculty, with the remaining 12 falling in other smaller categories. 20 participants were female, 14 were male, 1 was non-binary, and 1 preferred not to say. The majority of participants were from North America (19), followed by Europe (9), with other continents barely represented. Most relevant to our study, 15 had substantial experience with machine learning, 17 had an introductory understanding, and 4 had little to no knowledge.

For our analyses, we first calculated the mean accuracy and 95% confidence intervals (CIs) across all problems. We also calculated mean accuracy and participant confidence for each problem separately. To assess the impact of task type (forward vs. counterfactual simulation), we used the Wilcoxon signed-rank test due to the non-normal distribution of scores. For the analysis of difficulty levels based on the number of filters, we used the Friedman test, and for experience level comparisons, we used the Kruskal-Wallis H-test. Finally, we calculated Spearman's rank correlation coefficient and Kendall's tau-b coefficient to assess the relationship between accuracy and confidence.

4. Results

Across all 10 problems, our subset of 36 "serious" participants achieved a mean accuracy of 0.775 (standard deviation = 0.310, 95% CI = [0.656, 0.861]) and a median accuracy of 0.900 (95% CI = [0.800, 1.00]). 16 participants achieved a perfect score, as shown in the score distribution in Figure 3. Results for each individual problem are shown in Table 1.

When analyzing the two tasks separately, we found that participants performed slightly better on the forward simulation problems (mean accuracy = 0.806, standard deviation = 0.319, 95% CI = [0.678, 0.889]) compared to the counterfactual simulation problems (mean accuracy = 0.744, standard deviation = 0.360, 95% CI = [0.611, 0.850]). However, we found these differences to not be statistically significant. For this we used a Wilcoxon signed-rank test due to the non-normal distribution of the differences between means (Shapiro-Wilk test, $W = 0.759$, $p < 0.001$), which yielded a statistic of 36.500 and a non-significant p-value of 0.180.

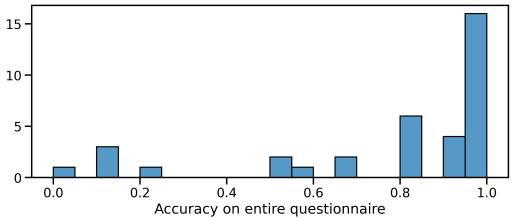


Figure 3: Histogram of participant accuracy across all problems.

Table 1
Per-problem accuracy and confidence statistics. Median accuracy for all problems was 1.00. FS = forward simulation, CS = counterfactual simulation.

	Accuracy		Confidence	
Problem	Mean	Std. dev.	Mean	Std. dev.
FS.1	0.778	0.422	3.556	0.809
FS.2	0.889	0.319	3.167	0.941
FS.3	0.778	0.422	3.694	0.668
FS.4	0.722	0.454	3.611	0.766
FS.5	0.861	0.351	3.222	0.898
CS.1	0.639	0.487	3.611	0.688
CS.2	0.861	0.351	3.389	0.964
CS.3	0.694	0.467	3.250	1.052
CS.4	0.750	0.439	3.361	0.990
CS.5	0.778	0.422	3.528	0.941

We also analyzed the impact of the number of filters presented in each problem. We grouped the problems into three difficulty levels based on the number of filters: easy (8 filters), medium (16 filters), and hard (32 filters). For the forward simulation task, we found no significant difference in accuracy across difficulty levels (Friedman test, $\chi^2 = 1.167$, $p = 0.558$). For the counterfactual simulation task, we also found no significant difference in accuracy across difficulty levels (Friedman test, $\chi^2 = 3.950$, $p = 0.139$). This suggests that the number of filters did not significantly impact participants' performance on either task.

Since participants' level of experience with machine learning varied, we evaluated the impact of their background on their task performance. Due to the non-normality and non-homogenous variances of the scores in each category, we used the Kruskal-Wallis H-test to compare the mean accuracy across the four groups of experience levels. The results showed no significant difference between groups (H-statistic = 0.819, $p = 0.845$). The mean accuracy and standard deviation for each group are shown in Table 2.

We explored the relationship between participants' accuracy and their confidence in their answers. We calculated Spearman's rank correlation coefficient and Kendall's tau-b coefficient to assess the monotonic relationship between correctness and confidence. The results showed a moderate positive correlation, with Spearman's rho = 0.449 ($p < 0.001$) and Kendall's tau-b = 0.430 ($p < 0.001$). This indicates

Table 2
Accuracy by experience level.

Experience level	Mean	Std. dev.	N
Little to no knowledge	0.550	0.520	4
Intro-level college course	0.876	0.139	17
Substantial real-life experience	0.745	0.336	11
Regularly develops models	0.650	0.473	4

that as participants' accuracy increased, their confidence in their answers also tended to increase. Aggregate statistics (problem-level) for these two measures are shown in Figure 4.

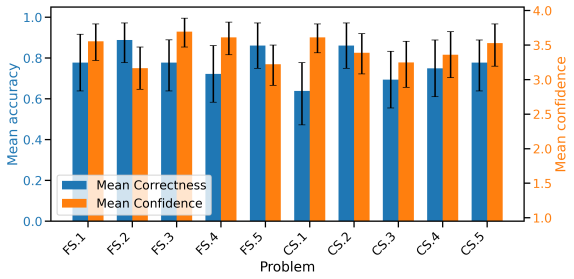


Figure 4: Mean accuracy and confidence by problem (with 95% CI). Confidence is measured on a 4-point Likert scale.

One thing to note is that the number of multiple-choice options in each problem varied widely, ranging from 5 (all counterfactual simulation problems) to 33 options (the two forward simulation problems with 32 filters). We reasoned that this could potentially skew the results, as participants might find it easier to guess correctly when there are fewer options. To address this, we adjusted the "correctness" of each answer by assigning a score of 1 for correct answers and a negative score of $-\frac{1}{k-1}$ for incorrect answers, where k is the number of options in the problem. This adjustment ensures that the penalty for incorrect answers is proportional to the number of options available, making it fairer across problems with different numbers of choices.

We applied adjustment and re-ran all of the analyses presented above for comparison. While this led to small changes in the various statistics, all significance test results remained unchanged. The accuracy difference between tasks remained non-significant (Wilcoxon signed-rank statistic = 63.000, $p = 0.117$), the difference based on number of filters remained non-significant for both tasks (Friedman test for forward simulation: $\chi^2 = 1.444$, $p = 0.486$; for counterfactual simulation: $\chi^2 = 3.950$, $p = 0.139$), and the experience level comparison also remained non-significant (Kruskal-Wallis H-statistic = 0.834, $p = 0.841$). Finally, the correlation between accuracy and confidence remained significant (Spearman's rho = 0.450, $p < 0.001$; Kendall's tau-b = 0.419, $p < 0.001$).

5. Discussion

Our finding that participants were able to accurately predict the model's behavior in both the forward simulation and counterfactual simulation tasks suggests that our model's explanations are both faithful and intelligible, as participants could effectively understand and utilize the information provided.

The correlation between accuracy and confidence further supports the notion that participants were not only able to predict the model's behavior but also felt confident in their understanding of the explanations. This further suggests that participants could grasp the underlying reasoning behind the model's predictions, as well as their own sense of understanding.

Though a bit counterintuitive, the lack of impact of the number of filters on participants' performance validates our intuition that our explanations allow for the tasks to be performed sequentially, with no additional cognitive load introduced by the increased number of filters. This is particularly important considering the large number of filters in our final model (132), and the essentially limitless number of potential patterns indicative of GTS behavior.

Though we are also encouraged by the lack of significant difference in performance based on participants' experience level, it is worth noting that this may in part be due to our small sample size (after removing responses that were clearly completed too quickly to have been attempted seriously). This sample size limitation would be good to address in future work.

Another point to note is that we did not present participants with the meaning of our model's features. While participants were generally able to complete the tasks successfully using our explanations, those explanations left the specific features vague. We did this intentionally. Although each feature was carefully engineered and imbued with meaning through the process of cognitive task analysis with a GTS expert [7], we reasoned (accurately, as evidenced by our findings) that for our particular simulation tasks, such understanding was not necessary. Even understanding GTS behavior or that this model was designed to be used in educational settings was ultimately irrelevant to the task. This is one difference between some human-grounded evaluations—such as the one we present here—and application-grounded evaluations, in which details pertaining to the domain and context become important.

This partly blind design makes it impossible to evaluate our explanations' plausibility, which refers to the degree to which an explanation aligns with human intuition [9]. This somewhat limits how much our evaluation encompasses the full scope of intelligibility [5]. Our participants clearly understood how to use the explanations, but they did not have sufficient information to be able to apply it to a real-world scenario.

This leads us to consider the implications of our findings. The results of our human-grounded evaluation suggest that our constraints-based approach to interpretable neural networks can indeed produce explanations that are demonstrably faithful, but also intelligible to a large degree. This is a significant step towards creating models that can be effectively used in educational settings, where understanding learner behavior is crucial.

6. Conclusion

In this paper, we presented a human-grounded evaluation of a constraints-based approach to fully interpretable neural networks for detecting learner behaviors. We performed this evaluation by designing and administering a questionnaire that included forward simulation and counterfactual simulation tasks. Our results on both tasks indicate that participants were able to accurately predict the model's

behavior, achieving an accuracy with mean = 0.775 and median = 0.900 across all problems. This suggests that the model's explanations are both faithful and intelligible, as participants could effectively understand and utilize the information provided.

Acknowledgments

This study is supported by the National Science Foundation under Award #1942962. Any conclusions expressed in this material do not necessarily reflect the views of the NSF.

References

[1] R. S. Baker, Challenges for the future of educational data mining: The baker learning analytics prizes, Journal of Educational Data Mining 11 (2019).

[2] Q. Liu, J. D. Pinto, L. Paquette, Applications of explainable AI (XAI) in education, in: D. Kourkoulou, A.-O. Tzirides, B. Cope, M. Kalantzis (Eds.), Trust and Inclusion in AI-Mediated Education: Where Human Learning Meets Learning Machines, Springer Nature Switzerland, Cham, 2024, pp. 93–109. doi:10.1007/978-3-031-64487-0_5.

[3] H. Suresh, S. R. Gomez, K. K. Nam, A. Satyanarayan, Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 1–16. doi:10.1145/3411764.3445088.

[4] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, 2017. arXiv:1702.08608.

[5] J. D. Pinto, L. Paquette, Towards a unified framework for evaluating explanations, in: Joint Proceedings of the Human-Centric eXplainable AI in Education and the Leveraging Large Language Models for Next Generation Educational Technologies Workshops (HEXED-L3MNGET 2024), volume 3840, CEUR-WS, Atlanta, Georgia, USA, 2024.

[6] J. D. Pinto, L. Paquette, A constraints-based approach to fully interpretable neural networks for detecting learner behaviors, in: Proceedings of the 18th International Conference on Educational Data Mining (EDM 2025), Palermo, Italy, 2025.

[7] L. Paquette, A. M. de Carvalho, R. S. Baker, Towards understanding expert coding of student disengagement in online learning, in: Proceedings of the 36th Annual Cognitive Science Conference, 2014, pp. 1126–1131.

[8] M. Rizzo, A. Veneri, A. Albarelli, C. Lucchese, M. Nobile, C. Conati, A theoretical framework for AI models explainability with application in biomedicine, in: 2023 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), IEEE, Eindhoven, Netherlands, 2023, pp. 1–9. doi:10.1109/CIBCB56990.2023.10264877.

[9] A. Jacovi, Y. Goldberg, Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4198–4205. doi:10.18653/v1/2020.acl-main.386.

A. Questionnaire: Task information

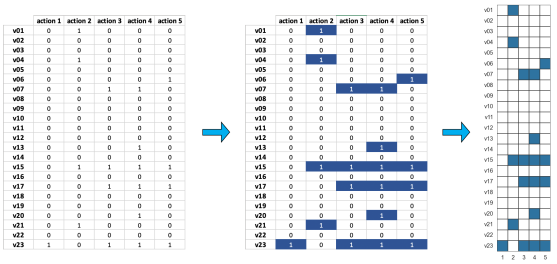
You will be presented with two tasks that involve visually matching patterns and selecting the correct response to a series of questions. These patterns have been extracted from a machine learning model designed to detect the student behavior known as *gaming the system* (GTS) in data of students' interactions with an intelligent tutoring system.

You do NOT need to understand the intricate details of the model or have any previous knowledge of GTS behavior in order to correctly answer the questions throughout this questionnaire. Your tasks will simply involve matching patterns to decide when the variables fed into the model will result in the model detecting GTS (positive) or not GTS (negative). If you find this interesting and wish to better understand how the model works (preferably after you complete this questionnaire), you can find some preliminary findings for this project in Pinto et al. (2023) or read the overarching big-picture proposal here.

For the sake of simplicity, each of the 23 variables that are fed into the model (inputs) have been labeled v01 to v23. Each is a binary variable, meaning it can have one of two values, 0 or 1. A value of 1 indicates that the variable was present for that action. As an example, one student's sequence of 5 actions on a particular problem may look like this:

	v01	v02	v03	v04	v05	v06	v07	v08	v09	v10	v11	v12	v13	v14	v15	v16	v17	v18	v19	v20	v21	v22	v23
action 1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
action 2	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0
action 3	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1
action 4	0	0	0	0	0	0	1	0	0	0	0	0	1	0	1	0	1	0	0	1	0	0	1
action 5	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1

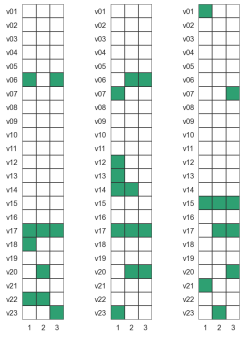
The patterns from the model will be presented to you in a transposed (vertical) format, so let's transpose this data now. Notice that the actions (indicative of time passing) now progress from left to right. To make it easier to read this at a glance, we'll also convert all values to colors: white for 0 and blue for 1.



This visualization makes it easy to see that this student took the following actions: **first**, they performed v23; **second**, they performed v01, v04, v15, and v21; **third**, they followed this with v07, v15, v17, and v23; their **fourth** action consisted of v07, v13, v15, v17, v20, and v23; finally, their **fifth** action in this sequence included v06, v15, v17, and v23.

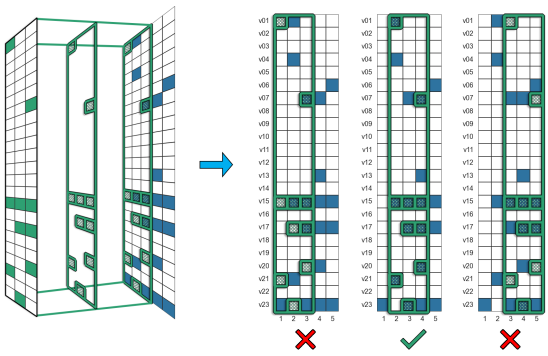
Now you will see some of the patterns automatically learned by the model. Notice that these patterns are all three actions long.

You can think of these learned patterns as a set of windows that "slide" across students' actions, one at a time. If

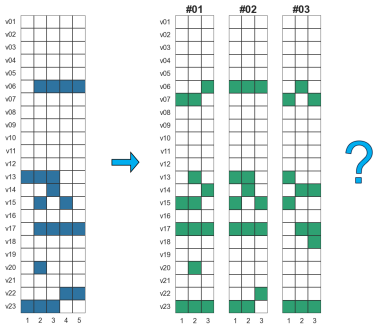


a sequence of student actions matches any of the model's learned patterns, then the model will label that input sequence as *GTS*. Otherwise, it will label it as *not GTS*.

As you can see below, the third pattern above matches the student data we've been looking at. Note that student actions can include additional positive (blue) variables that are not positive in the model's pattern. As long as the positive variables in the pattern are also positive in at least one set of three consecutive student actions, then the entire sequence is labeled *GTS*.

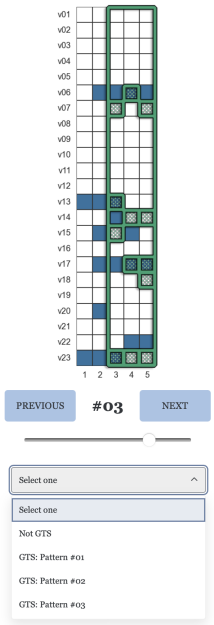


Time for you to test your understanding. Using the learned patterns below, would the model label the input sequence of actions (blue) as *GTS* or *not GTS*?



To make this process a bit more intuitive, we've included a second way to view each question, with a slider and the ability to cycle through each pattern. This interactive overlay view has all the same patterns as the static view above. Try it out below and feel free to use the view that you prefer in the real questions.

If you selected *GTS: Pattern #02*, you're right! (If you got it wrong, feel free to use the back button below to try to understand why.) This pattern is the only one that matches. Of course, the real model uses more than just three patterns,



and they're all patterns that it learned on its own given lots of training data.

Now that you're familiar with the visualizations of the model patterns and student actions, you'll get the chance to answer some similar questions. Remember that this is all to help us evaluate how well the inner workings of this machine learning model can be interpreted by flesh-and-blood humans—like you!