2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

# Quantitative Operator Strategy Comparisons across Human Supervisory Control Scenarios

Haibei Zhu, Rong Xu and Mary L. Cummings

*Abstract*— **Human-automation collaborations, like automated driving assistance and piloting drones, have become prevalent as these technologies become more commonplace. Designers need tools that help them understand how and why design interventions may change the strategies of operators in such complex human supervisory control systems. To this end, we demonstrate that when the divergence metric is applied to Hidden Markov Model (HMM) comparisons, it can accurately capture statistical differences between operator strategies for interfaces that embody different tasks. However, the use of such an approach is problematic when used to compare HMM strategy models with non-equivalent observations. To address this limitation, we developed an observation reduction approach and conducted a sensitivity analysis to assess the impact of this approach. Our results show that when comparing two non-equivalent interfaces, our observation reduction approach does not fundamentally change the divergence metric, thus allowing for direct model comparison. The results further show that HMMs from different interfaces produce a much higher divergence metric than model comparison from the same people who repeatedly use the same interface. Future work will examine if this method can detect differences in models with different tasks or modified interfaces.**

## I. INTRODUCTION

Human supervisory control (HSC) is a commonly-utilized control scheme in HRI applications where operators remotely manage an automated or autonomous system via control interfaces [1], [2]. In such a scheme, many factors can influence operators' performances and problem-solving strategies, including individual differences, different interface designs, and varying levels of autonomy [3]–[5]. Other than task performance, which can be directly observed and measured, operators' strategies are not directly observable and comparable. Thus, strategy models are needed for both comparing strategies and investigating those factors that influence operators' strategies.

One important application of such models is using them to understand whether a specific design has had its intended effect on human performance. For example, there are many poor designs in Unmanned Aerial Vehicle (UAV) control stations that have led to accidents [6]. A designer can use a strategy model to objectively compare the behaviors of UAV operators before and after a specific interface change, like adding a new decision support system, to determine whether the change resulted in improved strategies and overall better system performance.

This work was supported by the ONR under agreements number N00014-17-1-2012 and N00014-17-1-2504.

The authors are with the Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708, USA (haibei.zhu, rong.xu, m.cummings)@duke.edu

Hidden Markov models (HMMs) have been successfully used to model operators and systems in HSC scenarios like UAV control because of its two-layer structure and the ability to diagnose human-automation interactions [7]–[12]. While HMM models can be qualitatively compared to understand how before and after hidden states and transition probabilities influence an operator's problem-solving strategy, a more objective, quantitative approach is needed to determine whether one model is statistically different from another model. To this end, we have used the divergence as a statistical measure of model similarity between two HMM models [13], [14], and have shown that such a metric is stable when applied in similar settings [15].

Comparing HMMs can be difficult if there are non-equivalent observations. For example, one UAV interface may produce a different set of tasks to be performed when compared to a different UAV control station. In order to make comparisons between models with non-equivalent observations, we propose an observation reduction approach in which we realign observation types in the model with the greatest number of states by collapsing observations.

To demonstrate both the use of divergence metrics and the observation reduction approach, we develop and compare strategy models from four human-in-the-loop experimental sessions where operators control simulated multiple UAVs. Using the divergence distance as a measure of similarity between HMM models [13], the resulting divergence values quantitatively illustrate the impacts on operators' strategies from the differences across the experimental scenarios. Such differences include different participants, modified interfaces, and systems with increased autonomy.

This paper is organized such that Section II provides the background of related work and the HMM divergence measure. Section III describes experiment sessions with the model development process. Section IV shows the first two model comparisons. Section V presents the observation reduction approach, while comprehensive model comparisons are shown in Section VI. Section VII concludes this paper with a detailed discussion.

## II. BACKGROUND

Comparing operators' performance is a simple and direct metric for analyzing factor changes in human-automation interaction scenarios [3], [16], [17]. Such metrics only provide an aggregate summary of performance, so analyzing strategies is important to understanding the causes of good and bad performance. Some previous studies have demonstrated the development and comparison of operators' cognitive

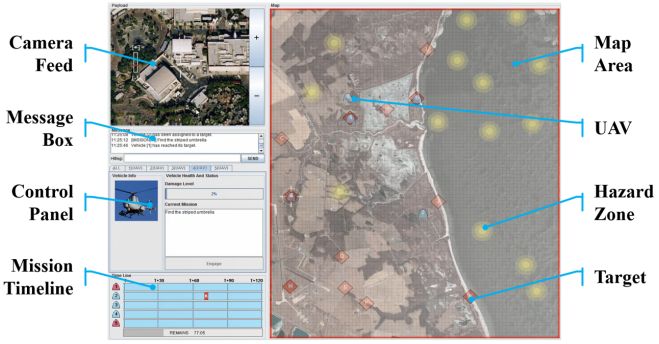


Fig. 1. The RESCHU interface - Interface 1

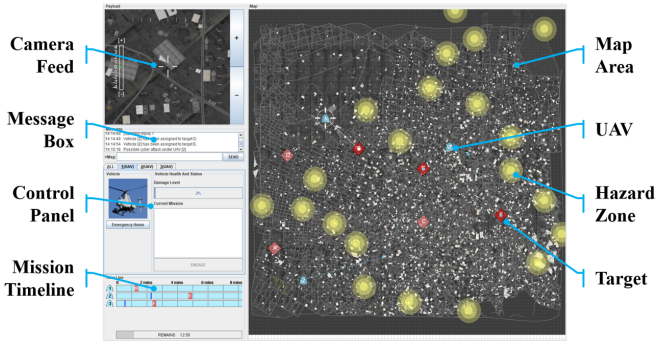


Fig. 2. The RESCHU-SA interface - Interface 2

models in the human information processing level [18]–[21]. However, such approaches have limited access to operators' strategies in high-level tasks. Thus, we focus on a higher modeling level of investigating strategies by utilizing HMMs.

HMMs have been utilized as strategy models because the two-layer structure, a hidden state layer and an observation layer [22], [23], approximates physical actions that can be seen and mental states that cannot be seen. We can consider a weighted cluster of observable interactions between operators and automation systems to be an abstract behavioral group, which is represented by a hidden state. Thus, hidden states, determined by clustered observations, and transitions among hidden states can describe operators' strategies [24], [25].

The HMM structure can be described as a tuple [22], [26], $\lambda = \{S, O, A, B\}$, in which, $S = \{S_1, S_2, \cdots, S_N\}$ represents $N$ different hidden states, and $O = \{O_1, O_2, \cdots, O_M\}$ represents $M$ different observation types. $A = \{a_{ij}\}$ is a $N \times N$ transition probability matrix, in which $a_{ij} = P\{S_j^{t+1} | S_i^t\}$, and $B = \{b_{ik}\}$ is a $N \times M$ emission matrix, in which $b_{ik} = P\{O_k | S_i\}$, $i, j \in [1, N]$, $k \in [1, M]$. The transition and emission matrices connect all hidden states and observations of an HMM.

Many HMM model comparison methods have been proposed and utilized for various applications [27]–[30]. However, these approaches compare HMMs with fixed model structures and observations. Based on the primary concept of quantitatively measuring model fitting for model comparisons as presented in these works, we further expanded this concept to a more comprehensive metric by evaluating all possible model structures. Also, since little work has focused on HMM comparisons with non-equivalent observations, we propose an observation reduction approach to accomplish such comparisons.

Thus, for this effort, we focus on the divergence measure approach, which can provide a distinguishable range of model difference measure, for quantitatively comparing HMM models [13]. The calculation of the divergence measure is defined as:

$$D(\lambda_1 \| \lambda_2) = \frac{1}{num} |\log(P(O_{all}|\lambda_1)) - \log(P(O_{all}|\lambda_2))| \quad (1)$$

In this equation, $\lambda_1$ is the first HMM model and $\lambda_2$ is the second model. Assume

$$\lambda_1 = \{S_1, O_1, A_1, B_1\};\ \lambda_2 = \{S_2, O_2, A_2, B_2\}$$

$\lambda_1$ contains $N_1$ different hidden states in $S_1$, and $M_1$ types of observations in $O_1$. $\lambda_2$ contains $N_2$ different hidden states in $S_2$, and $M_2$ types of observations in $O_2$. $O_{all}$ represents all observation sequences, and *num* is the total number of observations, or data points, in $O_{all}$. Also, $P(O_{all}|\lambda)$ is the likelihood value of an HMM model fitting on the evaluation dataset. For convenience and calculation efficiency, we usually take the logarithmic value $\log(P(O_{all}|\lambda))$ to represent the likelihood. Thus, the divergence approach measures the likelihood difference between two HMM models applying on a given evaluation dataset. Generally, a lower divergence value indicates a higher model similarity level [13].

The evaluation dataset $O_{all}$ is combined from the training datasets of the two HMM models to be compared, $O_{\lambda_1}$ and $O_{\lambda_2}$, to increase the confidence of the divergence measure. Understanding that slightly different tasks within one interface or slight modifications to an existing interface may produce different observations even if the overall interface is generally the same, HMM models developed from such datasets may contain different types and numbers of observations. In another word, if $\lambda_1$ and $\lambda_2$ are trained on datasets collected from different interfaces, it would cause $M_1 \neq M_2$. Then, both HMM models, $\lambda_1$ and $\lambda_2$, will be applied on the combined dataset $O_{all}$ to calculate the divergence distance. Since $M_1 \neq M_2$, then one of $P(O_{all}|\lambda_1)$ and $P(O_{all}|\lambda_2)$ makes this approach fail because the emission matrix of the model with less observation types cannot cover all observations in the dataset with a larger number of observations. To solve this non-equivalent observation issue in the divergence calculation, we investigated an observation reduction approach, discussed in the following sections.

## III. EXPERIMENTAL SESSIONS AND MODEL DEVELOPMENT PROCESS

### A. Experimental Sessions

We collected operator interaction data from four human-in-the-loop experimental sessions where participants controlled multiple UAVs to conduct high-level tasks using two different interfaces [15], [31]–[33]. The first interface used was the Research Environment for Supervisory Control of Heterogeneous Unmanned Vehicles (RESCHU) platform [34]. Shown as Interface 1 in Figure 1, RESCHU is a simulation-based platform that allows a single operator to

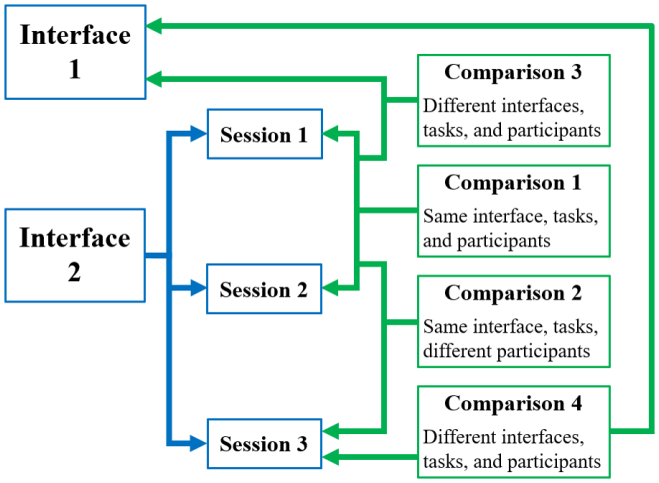


Fig. 3. Comparisons between experiments with different interfaces

control multiple UAVs in a supervisory control scenario. It includes both UAV navigational and imagery analysis tasks where operators focus on the map area when navigating UAVs to different targets, and they shift their attention to individual vehicle payload cameras that look down when a UAV reaches a target [32]. Data from this interface included a single experiment with 30 people [31].

Three experimental sessions were conducted using Interface 2, the Security-Aware RESCHU (RESCHU-SA) platform (Figure 2). Derived from RESCHU, RESCHU-SA adds a primary task of detecting possible GPS spoofing [33]. In RESCHU-SA, once operators receive a system hacking notification, they attempt to match a UAV camera view against the GPS-reported position on the map to determine potential hacking events [32]. The experiment with Interface 2 included 36 participants experiencing two experimental sessions (Sessions 1 and 2 in Figure 3). Session 3 repeated the same experiment but with 45 different participants. There was no significant difference in participants' overall performance across these three sessions based on statistical analyses with a significance level of $\alpha = 0.05$.

As shown in Figure 3, four comparisons between strategy models were conducted to quantitatively measure the potential differences between the four experimental sessions. In Comparison 1, strategy models from the two experimental sessions from Interface 2 were compared because the interface and tasks were the same and the participants were the same. Thus, the expectation is that the divergence measure would be the least. For Comparison 2, we compared the same interface and same task, but with different groups of people, and our expectation was that this comparison would yield an increased divergence measure. Comparisons 3 and 4 were expected to have the highest divergence distances because the participants, interfaces and tasks were all different. However, because the two different interfaces included different observations, we needed to formulate an observation reduction approach to be able to calculate the divergence measure.

TABLE I

HMM OBSERVATIONS FROM BOTH EXPERIMENT PLATFORMS

| Observations in both RESCHU and RESCHU-SA platform | | |
|---|---|---|
| 1 Add waypoint | 2 Move waypoint | 3 Delete waypoint |
| 4 Move endpoint | 5 Switch target | 6 Engage task |
| 7 Monitor UAV | | |
| Hacking detection observations only in the RESCHU-SA platform | | |
| 8 Perceive hacking | 9 Detection decision | 10 Adjust zoom level |

### B. Model Development Process

HMM strategy models were developed based on operators' actions (i.e., observations) collected during experiment sessions as listed in Table I. Given that the original RESCHU platform (Interface 1) only includes two major tasks of navigation and image analysis for targets, the first experiment dataset only contains observations $1-7$ in Table I. The task of UAV hacking detection was added in the RESCHU-SA platform, resulting in ten observations, $1-10$ in Table I.

HMM models were trained using the unsupervised multi-sequence Baum-Welch algorithm [22], which is a common expectation-maximization (EM) algorithm, and evaluated by the Bayesian Information Criterion (BIC) [35], which balances the model generalizability and complexity. To increase the confidence of the training results, over 100 randomly generated initializations were used for each specific model structure with a certain number of hidden states, and the resulted model with the highest likelihood was selected.

## IV. COMPARISONS WITHIN THE RESCHU-SA INTERFACE

To quantitatively determine the probabilistic difference between two models, we first focused on comparing the two experimental sessions conducted on the RESCHU-SA interface with the same participants (Comparison 1), and then with different participants (Comparison 2). Comparison 1 should result in the least difference between the strategy models when compared with Comparison 2. The 10 RESCHU-SA platform observations, shown in Table I, were used for all comparisons in this section.

Given the fact that divergence metrics can be different depending on the subjective interpretation of the BIC curve and selection of hidden states, divergence value meshes were plotted that depict the divergence values for model comparison across the number of all possible hidden states. Understanding that the RESCHU-SA interface provides three primary tasks, we consider that the minimum hidden state number in model comparisons is 3. Also, given that hidden states represent abstract cognitive groupings, the maximum number of hidden states should not be greater than the number of observation types, which is 10. Thus, such calculations can depict not only a central measure but also the variation, which would indicate the stability of a particular selection of model structures.

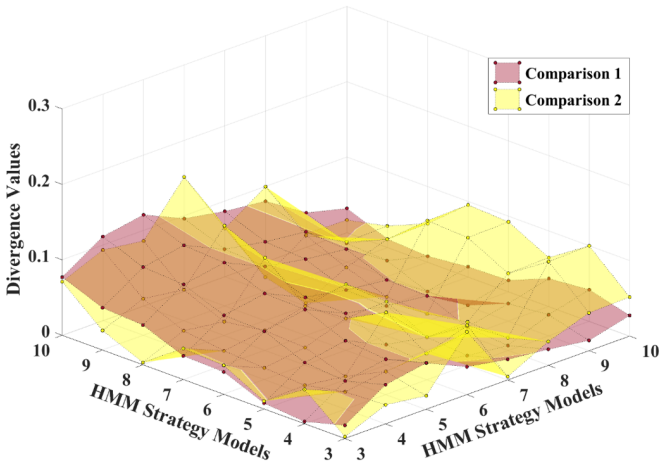


Fig. 4. Divergence value meshes for Comparisons 1 and 2

Figure 4 plots the mesh for Comparison 1 and 2, and both their divergence averages are 0.045 with (SD (standard deviation) = 0.029) and (SD = 0.033) respectively. A non-parametric Mann-Whitney test with a significance level of $\alpha = 0.05$ showed no significant difference between these two divergence value distributions $(p = 0.768)$. Given that both experiments had over 35 participants, the high similarity between Comparisons 1 and 2 indicates that different participant groups introduce limited variability in participants' overall strategies between experimental sessions with the same interface. These results align with the experimental data in that there was no significant difference in participants' overall performance. This data also establishes that in terms of human-in-the-loop performance in an HSC system, what it means to be similar can be roughly measured at 0.045.

As depicted in Figure 4, there are two remaining comparisons of interest which aim to assess the probabilistic difference between two different interfaces with different tasks. However, this comparison is not straightforward since the RESCHU interface only has 7 observations and RESCHU-SA has 10. The next section explores an observation reduction approach that was used to enable such a comparison.

## V. THE OBSERVATION REDUCTION APPROACH

In order to compare HMM models using the divergence measure in Equation (1), there need to be equivalent numbers of observations. To address this, we proposed an observation reduction approach, which modifies the selection criteria of observation types for training data and reduces observations for the HMM model with a higher number of observations. So, in the previous example, the experiment with 10 observations needs to be reduced to 7. Recall the notation in the previous sections, $\lambda_1$ has $M_1$ types of observations and $\lambda_2$ has $M_2$ observation types. Assume $M_1 < M_2$ so that the first model has fewer observations than the second model. To align the observation types, data points in the training dataset of $\lambda_2$, $O_{\lambda_2}$, are re-screened to match the $M_1$ observation types as the dataset of $\lambda_1$, $O_{\lambda_1}$.

TABLE II
OBSERVATION REDUCTION CRITERIA FOR THE SENSITIVITY TEST

| Index | 7 Obs | 8 Obs | 9 Obs | 10 Obs |
|---|---|---|---|---|
| 1 | Add waypoint | Same as 7 Obs | Same as 7 Obs | Same as 7 Obs |
| 2 | Move waypoint | | | |
| 3 | Delete waypoint | | | |
| 4 | Move endpoint | | | |
| 5 | Switch target | | | |
| 6 | Engage task | | | |
| 7 | Hacking detection | Monitor UAV | Monitor UAV | Monitor UAV |
| 8 | - | Hacking detection | Hacking detection | Perceive hacking |
| 9 | - | - | Adjust zoom level | Detection decision |
| 10 | - | - | - | Adjust zoom level |

### A. The Viterbi Propagation with Observation Reduction

With this observation reduction approach, the number of observations in $\lambda_1$ and $\lambda_2$ are aligned to $M_1$ that the dataset $O_{\lambda_2}$ is reformulated to $M_1$ types of observations as $O_{\lambda_1}$. Thus, the emission matrices of $\lambda_1$ and $\lambda_2$ share the same number of columns, $M_1$. Applying $\lambda_1$ and $\lambda_2$ to a one-dimensional sequence $O_{seq1} = (o_1, o_2, \cdots, o_t)$, the Viterbi propagation in Equation (1) can be updated for both models.

$$V_{\lambda_1:t,s_t} = \max_{s_t \in S_1}(b_{\lambda_1:s_t \to o_t} \cdot a_{\lambda_1:s_{t-1} \to s_t} \cdot V_{\lambda_1:t-1,s_{t-1}}) \quad (2)$$

$$V_{\lambda_2:t,s_t} = \max_{s_t \in S_2}(b_{\lambda_2:s_t \to o_t} \cdot a_{\lambda_2:s_{t-1} \to s_t} \cdot V_{\lambda_2:t-1,s_{t-1}}) \quad (3)$$

Given that both models contain $M_1$ types of observation, they share same expectations of emission probabilities.

In an HMM model, a higher number of hidden states could lead to a lower expectation of state transition probabilities and a highly differentiated model structure with a higher model complexity. However, high hidden state number could cause high corresponding emission probabilities. Thus, regardless of the number of hidden states, different model structures would only bring limited influence to the product of $a_{s_{t-1} \to s_t} \cdot V_{t-1,s_{t-1}}$ if models share the same number and type of observation.

Therefore, the difference in model likelihood values between both HMM models in Equation (2) and (3) would only be affected by the underlying patterns in the dataset, rather by the Viterbi algorithm propagation process. In this case, the divergence measure value, which is directly calculated from $\log(P(O_{all}|\lambda_1))$ and $\log(P(O_{all}|\lambda_2))$, will reflect the quantitative measure of the similarity level between the two models more precisely.

### B. The Observation Reduction Sensitivity Test

To understand the reliability of such an observation reduction approach, a sensitivity test was conducted to assess how collapsing observations impacted the overall divergence

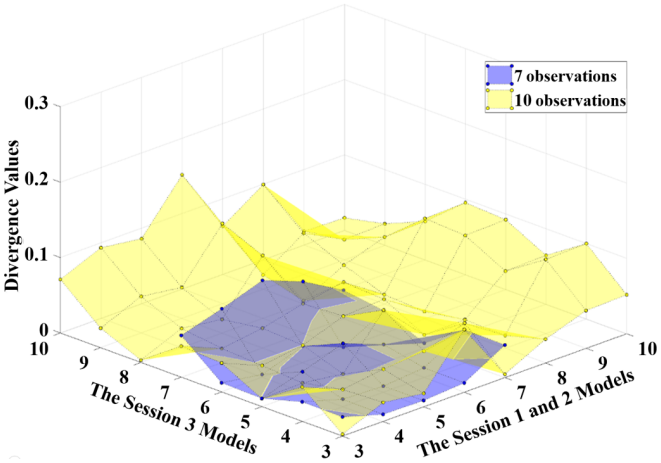


Fig. 5. Divergence value meshes between Sessions 1, 2, and 3 based on different observation reduction criteria and all possible hidden state numbers

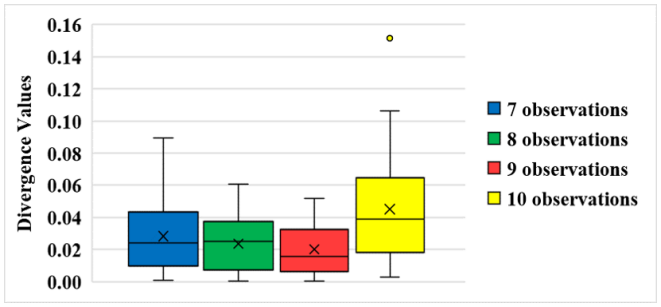


Fig. 6. Divergence value boxplots of model comparisons between Sessions 1, 2, and 3 based on different observation reduction criteria

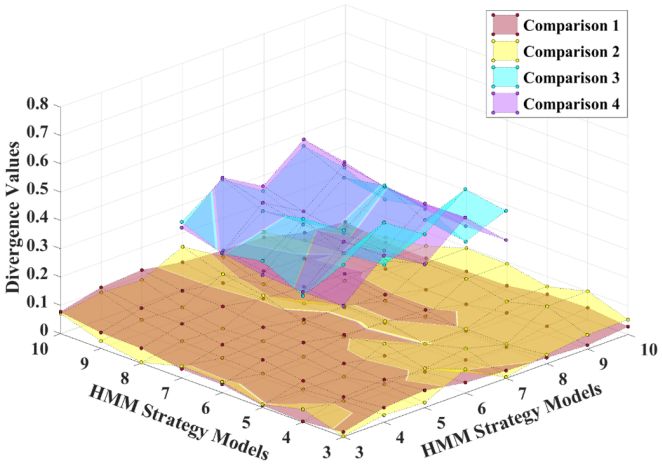


Fig. 7. Divergence value meshes of operator strategy model comparisons across all experimental sessions

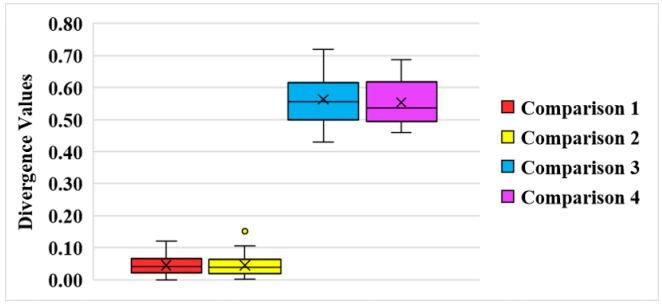


Fig. 8. Divergence value boxplots of operator strategy model comparisons across all experimental sessions

distance metric. Considering that collapsing observations may cause loss of information, it is necessary to ensure collapsing certain observations would not significantly change the divergence metrics.

Datasets from the RESCHU-SA interface, including Session 1, 2, and 3, were used with the 10 types of observations as shown in Table I. Understanding that Session 1 and 2 had same participants and Session 3 had another group of participants, further comparisons between different interfaces require the combined dataset of Session 1 and 2 and the Session 3 dataset. Thus, the sensitivity test was conducted on datasets from these three sessions. Given that the hacking detection task was uniquely embedded in the RESCHU-SA platform, we combined hacking detection-related observations based on different levels of abstraction. In order to understand the impact of collapsing the data from 10 to 7 observations, the revised data selection criteria for 7, 8, and 9 observations are shown in Table II.

With the revised observation-reduced models from Table II, HMM strategy models with all possible numbers of hidden states were retrained on the realigned datasets using the same methods discussed in the model development section. Then, visualizations of divergence measures between the HMM models were created, composed of divergence values from all possible combinations of model comparisons based on the different number of hidden states.

As shown in Figure 5, two divergence meshes represent the 7 and 10 observation model comparisons from the two RESCHU-SA sessions with corresponding observation reduction criteria. The 8 and 9 states were omitted for clarity but shared the same space. Quantitatively, the ranges of the average divergence values of these four meshes are all within $0.02-0.05$ (Figure 6). Mann-Whitney tests with a family-wise significance level of $\sim 0.008$ $(0.05/6)$ were conducted on the divergence values between two meshes. The statistical results show that the distribution of divergence values of both 8 and 9 observations meshes are significantly lower than the divergence distribution in the 10 observations mesh (both $p < 0.001$). However, the value distribution in the 7 observations mesh is not significantly different from the 10 observations mesh ($p = 0.017 > 0.008$).

Thus, although the observation reduction criteria of 8 and 9 observations may change the underlying patterns in datasets and affect model comparisons with the original dataset of 10 observations, the reduction criteria of 7 observations only introduce limited influence to the divergence measures. In this case, all hacking detection related observations could be collapsed to a single observation. In other words, the observation selection criterion for RESCHU-SA experiment models can follow the 7-observation rule instead of the original 10 observations without significant loss of information.

TABLE III
REVISED HMM OBSERVATIONS FROM BOTH EXPERIMENT PLATFORMS

| | | |
|---|---|---|
| 1 Add waypoint | 2 Move waypoint | 3 Delete waypoint |
| 4 Move endpoint | 5 Switch target | 6 Engage task |
| RESCHU | 7 Monitor UAV | |
| RESCHU-SA | 7 Hacking detection | |

## VI. COMPARISONS BETWEEN THE RESCHU AND RESCHU-SA INTERFACES

Based on the observation reduction sensitivity test from the previous sections, the model comparison with the original 10 observations was not significantly different from the modified 7 observations. Thus, we can compare HMM strategy models between RESCHU-SA using 7-observation reduction criterion and RESCHU with the original 7 observations as shown in Table III.

Comparison 3 focuses on the difference between RESCHU and RESCHU-SA with one group of people and Comparison 4 also examined the difference between RESCHU and RESCHU-SA, but with a different group of people. The expectation is that such divergence distance metrics should be similar between them but very different from Comparisons 1 and 2.

The divergence value meshes shown in Figure 7 illustrate the differences in strategy models developed between RESCHU and RESCHU-SA interfaces across the 4 comparisons, and their means and standard deviations are plotted in Figure 8. For Comparisons 3 and 4, both divergence meshes range between $0.40-0.75$, which indicates a relatively large difference. The average for Comparison 3 is 0.563 ($SD = 0.078$), while Comparison 4 average is 0.553 ($SD = 0.067$). These two meshes interleave, and a Mann-Whitney test shows no significant difference ($p = 0.691$) between the two, which agrees with the expectation that they would be similar. Also, such a high similarity level between Comparison 3 and 4 supports the fact that different participant groups across the same interface would only bring limited variance to the general strategies.

One important comparison to be made in Figure 7 is between the clustering of Comparisons 1 and 2 and Comparisons 3 and 4. It is clear that the divergence distance metric not only accurately captures differences between two interfaces, but this difference is relatively consistent across Comparisons 3 and 4. Understanding that the RESCHU-SA interface contains an additional primary task of UAV hacking detection, which is not provided in the original RESCHU platform, participants using RESCHU-SA had quantitatively different strategies and behavioral patterns comparing to those using RESCHU.

Another element of Figure 7 worth noting is that the meshes of Comparisons 1 and 2 are relatively flat. This flatness indicates model stability in capturing operators' general strategies with three or more hidden states when two scenarios share the same interface and primary tasks. Meanwhile, the meshes of Comparison 3 and 4 are relatively uneven across comparisons with all possible hidden states. So, comparisons of HMM strategy models with different interfaces may be less stable, which could be a result of the non-equivalent observation manipulations. Further studies will investigate how such manipulations affect the quantitative measures.

Figure 8 illustrates a range of what divergence metrics can likely capture. In this analysis, the most similar comparison was between the same people using the same interface for the same tasks, Comparison 1. Comparisons 3 and 4, which had almost identical mean divergence metrics, looked at different interfaces and tasks with different people. The overall difference in these means is about $0.500-0.520$. Understanding that different groups of operators only introduce limited variance to the divergence metrics, such an overall mean difference is a quantitative similarity metric of adding an additional primary task in an HSC scenario. However, it remains to be seen whether this relative difference holds between other interfaces with different tasks and how it can be leveraged in various applications, which is also a future research direction.

## VII. DISCUSSION AND CONCLUSION

Given that operator strategy models could help researchers investigate operators' strategies with varying HRI system designs, quantitative comparisons between models from different supervisory control scenarios may provide a more objective basis for assessment. In this effort, we developed and compared strategy models from four human-in-the-loop experimental sessions with varying degrees of difference, including different participant groups, experimental tasks and interfaces. To allow for model comparisons with non-equivalent observations using the HMM divergence measure, we proposed an observation reduction approach and justified it with a set of sensitivity tests.

The model comparison results show that the divergence distance approach can quantitatively capture differences in HMM strategy models, including when people use different interfaces. This ability to detect such a difference is especially important given the fact that the RESCHU-SA was derived from the RESCHU platform.

One caveat in this effort is that this approach was not able to capture a probabilistic difference in models when different groups of people used the same interface, so there is likely a lower limit of a just noticeable difference. What remains untested in this effort, and is the subject of future work, is whether this method can detect changes in strategies when there are different tasks in the same interface or modified interfaces.

In addition to this future work, more effort is needed to further investigate issues surrounding the observation reduction approach. Information loss occurs when observations are collapsed, which could be important in signaling important model differences. So, further studies are needed to investigate such an approach for non-equivalent model datasets and other HSC applications.

Lastly, observation reduction can be a subjective process so knowing which observations to reduce and determining when a reduction loses critical information is also an area of future work. While such models are intended to primarily diagnose whether changes in a system lead to detectable changes in operator strategies, another area that deserves further investigation is whether the utilization of a divergence metric could be used to predict future performance based on the strategy changes, which has clear safety implications for many HSC systems.

## ACKNOWLEDGMENT

We gratefully acknowledge the assistance of Andy Wang, who worked as a summer intern in the Duke Humans and Autonomy Lab (HAL), in implementing the RESCHU-SA platform, and the assistance of Matthew Seong, a UNC graduate, in conducting experiments. The Office of Naval Research partially sponsored this work.

## REFERENCES

[1] T. B. Sheridan, *Telerobotics, Automation, and Human Supervisory Control*. Cambridge, MA: MIT Press, 1992.

[2] C. D. Wickens, S. E. Gordon, and Y. Liu, *An Introduction to Human Factors Engineering*. New York, NY: Longman, 1998.

[3] C. D. Wickens, J. G. Hollands, S. Banbury, and R. Parasuraman, *Engineering Psychology and Human Performance*. Psychology Press, 2015.

[4] M. L. Cummings, S. Bruni, and P. J. Mitchell, "Human supervisory control challenges in network-centric operations," *Reviews of Human Factors and Ergonomics*, vol. 6, no. 1, pp. 34–78, 2010.

[5] J. Y. Chen and M. J. Barnes, "Human–agent teaming for multirobot control: A review of human factors issues," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 1, pp. 13–29, 2014.

[6] G. Carrigan, D. Long, M. L. Cummings, and J. Duffner, "Human factors analysis of predator B crash," in *Association for Unmanned Vehicle Systems International (AUVSI) 2008: Unmanned Systems North America*, San Diego, CA, June 2008.

[7] Z. Wang, A. Peer, and M. Buss, "An HMM approach to realistic haptic human-robot interaction," in *World Haptics 2009-Third Joint EuroHaptics conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*. IEEE, 2009, pp. 374–379.

[8] D. Kulic and E. A. Croft, "Affective state estimation for human-robot interaction," *IEEE Transactions on Robotics*, vol. 23, no. 5, pp. 991–1000, 2007.

[9] R. Kelley, A. Tavakkoli, C. King, M. Nicolescu, M. Nicolescu, and G. Bebis, "Understanding human intentions via hidden Markov models in autonomous mobile robots," in *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction*, 2008, pp. 367–374.

[10] D. Kragic, P. Marayong, M. Li, A. M. Okamura, and G. D. Hager, "Human-machine collaborative systems for microsurgical applications," *The International Journal of Robotics Research*, vol. 24, no. 9, pp. 731–741, 2005.

[11] T. Petković, D. Puljiz, I. Marković, and B. Hein, "Human intention estimation based on hidden Markov model motion validation for safe flexible robotized warehouses," *Robotics and Computer-Integrated Manufacturing*, vol. 57, pp. 182–196, 2019.

[12] R. Fu, H. Wang, and W. Zhao, "Dynamic driver fatigue detection using hidden Markov model in real driving condition," *Expert Systems with Applications*, vol. 63, pp. 397–411, 2016.

[13] B.-H. Juang and L. R. Rabiner, "A probabilistic distance measure for hidden Markov models," *AT&T Technical Journal*, vol. 64, no. 2, pp. 391–408, Feb. 1985.

[14] J. Silva and S. Narayanan, "Average divergence distance as a statistical discrimination measure for hidden Markov models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 890–906, 2006.

[15] H. Zhu and M. L. Cummings, "The stability of human supervisory control operator behavioral models using hidden Markov models," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 6971–6978.

[16] H. A. Yanco, A. Norton, W. Ober, D. Shane, A. Skinner, and J. Vice, "Analysis of human-robot interaction at the darpa robotics challenge trials," *Journal of Field Robotics*, vol. 32, no. 3, pp. 420–444, 2015.

[17] B. Sadrfaridpour, H. Saeidi, J. Burke, K. Madathil, and Y. Wang, "Modeling and control of trust in human-robot collaborative manufacturing," in *Robust Intelligence and Trust in Autonomous Systems*. Springer, 2016, pp. 115–141.

[18] J. N. Marewski and K. Mehlhorn, "Using the ACT-R architecture to specify 39 quantitative process models of decision making," *Judgment and Decision making*, vol. 6, pp. 439–519, 2011.

[19] S. Prezenski, A. Brechmann, S. Wolff, and N. Russwinkel, "A cognitive modeling approach to strategy formation in dynamic decision making," *Frontiers in Psychology*, vol. 8, p. 1335, August 2017.

[20] C. Gonzalez, V. Dutt, A. F. Healy, M. D. Young, and L. E. Bourne Jr, "Comparison of instance and strategy models in ACT-R," in *Proceedings of the 9th International Conference on Cognitive Modeling (ICCM 2009), A. Howes, D. Peebles and R. Cooper (Eds.)*, Manchester, UK, 2009.

[21] R. Geng and J. Tian, "Improving web navigation usability by comparing actual and anticipated usage," *IEEE Transactions on Human-Machine Systems*, vol. 45, no. 1, pp. 84–94, 2014.

[22] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *The Annals of Mathematical Statistics*, vol. 37, no. 6, pp. 1554–1563, 1966.

[23] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[24] Y. Boussemart, M. L. Cummings, J. L. Fargeas, and N. Roy, "Supervised vs. unsupervised learning for operator state modeling in unmanned vehicle settings," *Journal of Aerospace Computing, Information, and Communication*, vol. 8, no. 3, pp. 71–85, 2011.

[25] V. Rodríguez-Fernández, A. Gonzalez-Pardo, and D. Camacho, "Finding behavioral patterns of UAV operators using multichannel hidden Markov models," in *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2016, pp. 1–8.

[26] L. R. Rabiner and B.-H. Juang, "An introduction to hidden Markov models," *IEEE Acoust., Speech, Signal Processing Mag.*, vol. 3, no. 1, pp. 4–16, Jan. 1986.

[27] S. M. E. Sahraeian and B.-J. Yoon, "A novel low-complexity HMM similarity measure," *IEEE Signal Processing Letters*, vol. 18, no. 2, pp. 87–90, 2011.

[28] Y. Qi, J. W. Paisley, and L. Carin, "Music analysis using hidden Markov mixture models," *IEEE Transactions on Signal Processing*, vol. 55, no. 11, pp. 5209–5224, 2007.

[29] C. Bahlmann and H. Burkhardt, "Measuring HMM similarity with the Bayes probability of error and its application to online handwriting recognition," in *Proceedings of Sixth International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2001, pp. 406–411.

[30] N. Nguyen-Duc-Thanh, S. Lee, and D. Kim, "Two-stage hidden Markov model in gesture recognition for human robot interaction," *International Journal of Advanced Robotic Systems*, vol. 9, no. 2, p. 39, 2012.

[31] M. Cummings, L. Huang, H. Zhu, D. Finkelstein, and R. Wei, "The impact of increasing autonomy on training requirements in a UAV supervisory control task," *Journal of Cognitive Engineering and Decision Making*, vol. 13, no. 4, pp. 295–309, 2019.

[32] H. Zhu, M. Elfar, M. Pajic, Z. Wang, and M. L. Cummings, "Human augmentation of UAV cyber-attack detection," in *International Conference on Augmented Cognition*. Springer, 2018, pp. 154–167.

[33] H. Zhu, M. L. Cummings, M. Elfar, Z. Wang, and M. Pajic, "Operator strategy model development in UAV hacking detection," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 6, pp. 540–549, 2019.

[34] B. Donmez, C. Nehme, and M. L. Cummings, "Modeling workload impact in multiple unmanned vehicle supervisory control," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 40, no. 6, pp. 1180–1190, Nov. 2010.

[35] G. Schwarz *et al.*, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.