SUBMISSION: 9630

Title: A Comparative Study on Reward Models for UI Adaptation with Reinforcement Learning

ESEM 2023 - Registered Reports Track

----------------------- *REVIEW 1* ---------------------

SCORE: 2 (accept)

R1. Comment 1: The registered report is about a study to compare Human-Computer Interaction (HCI) models with HCI models augmented with human feedback as strategies to generate reward models for user interface adaptations. The evaluation considers the impact of the two approaches on user engagement and satisfaction. The proposed research is interesting and can potentially make a real and worthy contribution to Software Engineering and HCI research.

The paper is well-written and easy to read, except for a few minor mistakes. The abstract has the items for a structured abstract as required by the CFP. Also, the research questions and hypotheses are clear. However, some improvements and clarifications are required for acceptance.

R1. Comment 2: This is a bit of a provocation to the authors: is it not expected that the HCI model augmented with human feedback will be better? The Introduction can explore a bit more on this, maybe exploring the importance of knowing how much better/worse it is and analyzing the costs/effort of each approach (as cost/effort is also a relevant variable in decision-making about what approaches to adopt in software development).

R1. Comment 3: Please, in Section III.B, provide the exact list of metrics of user engagement that will be collected. As written, it sounds like the authors provide examples of metrics that can be used. See the end of the first column on Page 4 (and the beginning of the second column).

***ANSWER to R1.Comment 3*:**

R1. Comment 4: In Section III.D, at which points are data collected (e.g., regarding user satisfaction)? Something about what is collected is at the beginning of Section III.E, but the detail of when is missing.

R1. Comment 5: Regarding the threats to validity, would not a long break be more effective in dealing with the maturation effect? The point of allowing participants to rest does not seem to be the problem here.

R1. Comment 6: Are participants in different groups (1 and 2) part of different classes? Can participants from different groups interact during the periods/sessions, creating a threat?

R1. Comment 7: Will the researchers record the thinking aloud from participants? Will this be analyzed at some point for some reason? If so, how?

R1. Comment 8: MINOR ISSUES: There is an incorrect citation in the Introduction ([x]). Also, is the definition of the HCI model really coming from references [10] and [11] (second paragraph of page 2)?

There are a few writing mistakes (e.g., “However, suggest the right adaptation…” instead of “However, suggesting the right…”; adaption instead of adaptation).

---------------------- REVIEW 2 --------------------

SCORE: 2 (accept)

R2. Comment 1. *1. Importance of the research question(s).*

The research report describes a confirmatory study on the user experience and usability of adaptive user interfaces (AUI) for a sporting goods store and a course management system. For both application scenarios, the authors evaluate different AUI versions. One version applies reinforcement learning (RL) models supported by a predictive cognitive model, while the second version uses a cognitive model and human feedback. The report and future study results address an important research topic, and the case is relevant. Especially the use of machine learning methods considered in this context is auspicious. The plans appear to be appropriate but require some improvements. Given that the detailed recommendations below will be addressed, we recommend a publication at the ESEM Conference and Journal with Continuity Acceptance (CA).

2.) The proposed hypotheses’ logic, rationale, and plausibility.

R2. Comment 2. Overall, the proposed research questions and hypotheses present an intriguing investigation into the effectiveness of reward models in adaptive user interfaces (AUI). The research questions are relevant to the current trends in human-computer interaction (HCI) and machine learning. Empirical methods are used to derive relevant knowledge for building adaptive user interfaces. However, several aspects could be refined to enhance the clarity and focus of the research questions and hypotheses.

2.1) Firstly, RQ1 investigates whether AUIs with RL models that use predictive cognition models with human feedback (MCTS-HCI) are more effective than AUIs with RL models that use predictive cognition models without human feedback (MCTS-HCI-HF).

In this regard, the related work section reports two studies of the MCTS-HF that were helpfully applied to train such models (“we believe that human feedback could be used to optimize the reward model”). Given these statements, it seems surprising that no differences between MCTS-HCI and MCTS-HCI+HF are assumed in RQ1 and Hn1. Similarly, RQ2 asks about the difference in user engagement between non-AUIs and MCTS-HCI AUIs. However, the literature has already demonstrated the positive effect of AUIs. Therefore, the authors should reflect on the necessity of the question again and then clarify its relevance to the reader. Has the impact of system adaptivity on user engagement yet to be studied? Even though the decision to formulate all null hypotheses (Hn1 to Hn5) as statements of no effect are joint in hypothesis testing, there might be an opportunity to clarify the specific alternative hypotheses the study seeks to support. The statement “All the hypotheses are two-sided because we did not postulate that any effect would occur as a result of different RL model’s usage.” seems to contradict the purpose of the research, which is to investigate the potential effect of reward models on AUI effectiveness. Thus, this statement could be clarified or reconsidered.

R2. Comment 3. 2.2) Secondly, the article would generally profit from an explicitly described relation between previous work, research questions, and hypothesis. The object of the experiment, the experimental design, and the variables seem to have already been determined, while the research questions were derived afterward. However, the procedure of the Scientific Method is supposed to be the other way around. It is recommended to start with a research question that is specific, measurable, and based on observations or insights from the related work focusing on the relationship between two or more variables. After stating the research problem at the end of a solid motivational part, the reader would have expected more detailed or nearly slightly different scientific literature that would make him/her understand the current state of knowledge and the gaps in knowledge more in detail, followed by a refined statement of the research questions. Then in the final step, I would have expected a hypothesis that would have been derived from the research question explaining observations from the literature by describing the relationship between testable (operationalized) variables clearly and concisely. Making more logical connections between existing works and RQs and hypotheses could be addressed by describing in detail previous works' study variables and the effects of the independent variables on the dependent variables observed. The dependent, independent, and control variables defined in the study plan should be derived from previous work and be justified more substantiated. This applies to all RQs and hypotheses.

R2. Comment 4: 2.3) In addition, Research Question 1 (RQ1) and Research Question 2 (RQ2), according to the hypothesis, are both targeting the effectiveness of AUIs on user engagement. Still, the distinction between the two needs to be clarified. A more evident difference between these two questions would be beneficial, as they explore the same concept but with slightly different facets. This could be achieved by further distinguishing what aspect of 'effectiveness' each question targets. The authors should clarify how 'user engagement' and 'user satisfaction' are defined and differentiated. These terms can be understood in various ways, and it would be necessary to give precise definitions – not only in the related work and variable description but also in the research questions and hypothesis part – to avoid confusion and ambiguity. The hypotheses assume these concepts can be quantified and statistically tested, so it will be necessary to elaborate on how these metrics will also be operationalized in the hypothesis.

R2.Comment 5. *3.) Soundness and feasibility of the methodology and analysis pipeline (including statistical power analysis where appropriate).*
Overall the proposed methodology appears sound and feasible. Still, it requires more clarity and improvement regarding a consistent description and differentiation of independent, dependent, and control variables in all article sections. The methodology and analysis can be improved by addressing the following aspects:
3.1) The design includes a baseline condition for initial performance levels and a detailed analysis plan, this approach to controlling and mitigating threats to validity is highly commendable, but improvements need to be made regarding control of the effect of the variable which the authors describe as "experimental object" variable. First, we recommend renaming this variable with a more explicit label, such as "application domain." However, the label "experimental object" is too general and might also include "RL model" or "AUI." Secondly, the variable "application domain" must be mentioned consistently in the RQs, hypothesis, and related work as described in 2.2). Most importantly, the "application domain" variable must also be varied for the baseline. Otherwise, effects in user engagement and user satisfaction of different AUIs compared with the non-adaptive baseline condition might be attributable to different

application types rather than the adaptivity method. Including three different application domains would lead to a 3x3 factorial repeated measures within-subject design, having nine conditions that all participants conduct the experimental task. However, due to the nine different factor combinations and the associated long test time, the fatigue of the test subjects could influence the test results. To avoid this and still be able to control the variable "application domain," we recommend considering only one "application domain" in one experiment if the conduction of the 3x3 factorial design takes too much time or to include several breaks within the 3x3 experiment.

R2. Comment 6. 3.2) It is not entirely clear to the reader whether the measures of the dependent variable "user engagement" (clicks and time spent on each page, the frequency and number of user activities, recency, and the total number of actions performed during the user interaction) are the same measures used for training the RL models. Properties of an independent variable should not be determined by the measurement points of a dependent variable, as this would compromise independence and make the interpretation of the results for hypothesis testing inadmissible.

R2. Comment 7. 3.3) The experiment uses a combination of objective (user engagement measured by HCI models) and subjective measures (user satisfaction measured by SUS questionnaire). The reward model derived from predictive HCI models exclusively (AUI-HCI) and the reward model derived from predictive HCI models augmented by human feedback (AUI-HCI-HF) are independent variables that should allow for a robust comparison. However, in the designation and interpretation of the variables, the AUI and RL models depend on each other. Therefore, the results will not allow concluding whether the effects can be attributed solely to the RL model or the AUI. The authors should communicate that different AUI types will be investigated, whose factors include "non-AUI," "AUI-RLHCI," and "AUI-RLHCI HF" if point 3.1 will be considered.
In summary, the experimental design, including a balanced within-subject two-treatment (3x3 if 3.1 will be considered) factorial crossover design, effectively addresses the issue of small sample sizes and increases the sensitivity of experiments. Including a baseline period allows initial performance levels to be considered when analyzing the results. The way the authors have set up different sequences and periods to control for order effects is also well-thought-out, given those recommendations from 2) will be addressed.

R2. Comment 8. 3.5) The proposed statistical analysis is well-justified and comprehensive, covering descriptive statistics, Linear Mixed Model (LMM) analysis, effect size measurements, and system usability assessments. Furthermore, acknowledging the different tests based on the final sample size indicates a thoughtful and appropriate approach. However, as part of the analysis, we recommend testing and reporting the assumptions of LMMs that the relationship between the predictors and the response variable is linear and that residuals/errors from individual subjects are independent of one another, as the presence of both fixed effects (parameters that are assumed to be the same across all subjects) and random effects (parameters that are allowed to vary across subjects) need to be present. The random effects can account for within-subject correlations in the data, but the residuals are assumed to be uncorrelated after accounting for these effects. Furthermore, the authors should prove that homoscedasticity and normality of errors are given.

R2. Comment 9. 3.6) A power analysis has yet to be explicitly mentioned. This is important in determining the appropriate sample size to detect an effect of a given size. The authors should consider doing a power analysis.

R2. Comment 10. 3.7) One potential concern might be the operationalization of user engagement using the number of clicks, time spent on each page, frequency of user activities, etc., as these may not accurately reflect engagement. For example, participants might click around out of confusion rather than engagement.

R2. Comment 11. *4. (For confirmatory study) Does the clarity and degree of methodological detail sufficiently replicate the proposed experimental procedures and analysis pipeline?*
Overall, the provided method section is very detailed and includes sufficient information for replication. However, there are a few points where more clarity could be beneficial.

4.1) The description of the independent variables and their operationalization is clear and well-described. In addition, the description of the experimental objects (Sports and Courses) and the meaning of the abbreviations (AUI-HCI and AUI-HCIHF) are easy to understand. However, more detail about the actual implementations of the RL-based UI adaptation strategies and the different experimental tasks the participants will conduct with the system are necessary.

R2. Comment 12. 4.2) Furthermore, the objective of the sampling strategy needs to be clearly defined. Understanding why and how the authors chose participants is essential to evaluating the strength and generalizability of findings. We recommend that the authors provide a clear and concise explanation of the goal behind their sampling approach. A crucial aspect that may also catch readers' attention is the nature of the sample population. The final study results may lack representativeness, a critical component for the generalizability of the results. While achieving a fully representative sample can be challenging, we strongly encourage acknowledging this as a limitation in the final manuscript. Additionally, sampling solely from Masters students in Computer Science might introduce bias. These students may already have an understanding or familiarity with adaptive user interfaces (AUIs), and this prior knowledge could influence their perception and evaluation of AUIs in your study. One way to account for this would be to ask participants about their familiarity with AUIs before or after the experiment, providing a way to control for this potential bias.

R2. Comment 13. 4.3) Considering the stated motivation for the research on AUIs as a solution to context variability, the homogeneity of the sample needs to better align with this goal. If the aim is to examine how AUIs adapt to the context of use (i.e., platform, environment, or user) at run-time, the sample should ideally provide more variability. Although this might not be feasible for the forum or environment, increasing user variability in your sample may lend more strength to your research.

R2. Comment 14. 4.4) The subject recruitment plan is well explained. The consent form and data handling procedures are also defined. However, more detail about the anticipated number of participants or how that number will be determined would be helpful. This is also important for ensuring the statistical power of the analysis.

R2. Comment 15. 4.5) The use of a balanced within-subject two-treatment (three or 3x3 given 2.1) factorial crossover design is clearly explained, as is the need for baseline measurements. The explanation of the sequences and periods is also understandable. However, the final paper must clearly document each period and session's duration.

R2. Comment 16. 4.6) The section on validity is well-considered and suggests appropriate mitigations for potential threats. Nonetheless, providing more details on how threats to statistical conclusion validity, such as low statistical power or violated assumptions of statistical tests, will be addressed could further strengthen this part.

R2. Comment 17. 4.7) However, several aspects are mentioned in the paper where additional clarity could improve the readability and overall understanding of the work. For example, the term "Bandit systems" is introduced without sufficient explanation and doesn't appear to be further referenced in the manuscript. It would be helpful for the readers to provide an explicit definition or context to understand how it relates to the rest of the study. Secondly, the Monte Carlo Tree Search concept is mentioned but needs to be adequately explained. Considering its potential importance to your research, a clear and detailed explanation of this technique would be advantageous. Thirdly, the details of the Human-Computer Interaction (HCI) models utilized could be elaborated more. Understanding these models seems crucial to fully grasp the methodologies and results of your study. Providing a comprehensive description or referencing sources for additional reading on these models would be very beneficial. Finally, the authors should review their manuscript with an eye for potential confusion and provide further explanations or examples for these complex terms and concepts.

R2. Comment 18. *5. Do the authors have pre-specified sufficient outcome-neutral tests to ensure that the results can test the stated hypotheses, including positive controls and quality checks?*
Based on the information provided, the authors have considered designing their experiment to test their hypotheses effectively. Using a non-adaptive user interface as a control condition is a well-considered choice that will allow comparing the reinforcement learning (RL) based interfaces against a baseline condition. This acts as a form of positive control as it is expected to show some level of user engagement and satisfaction, against which the effectiveness of the adaptive interfaces can be compared. The authors have also laid out their plan for checking the validity of their data, including checking for residual normality in their linear mixed model and using established metrics for user engagement and satisfaction (like the System Usability Scale or SUS). The within-subjects crossover design also helps control for participant-related variability. However, the authors should avoid the grouping shown in Fig. 2 b), as this suggests there would be two different groups of participants whose results would be compared. However, the research questions and hypotheses do not describe this between-subject design. While the authors do not explicitly mention any outcome-neutral tests, their use of a within-subjects design and their decision to balance the presentation order of the treatments to each participant will help to control for potential order effects.

R2. Comment 19. Last but not least, the authors could specify interim data quality checks or pilot testing to confirm that their data collection works as intended and that the RL algorithms are correctly implemented and functioning as expected. Details of how outliers or missing data will be handled would also be valuable. Outliers can dramatically skew results, and pre-specifying how they will be identified and addressed can prevent post hoc decisions that might bias the results. The authors might also pre-specify a plan for conducting secondary analyses, such as testing for potential interaction effects between variables or subgroup analyses.

----------------------- REVIEW 3 ---------------------

SCORE: -2 (reject)

R3. Comment 1: This paper aims to perform a confirmatory study to compare two different approaches to generate reward models in the context of UI adaptation using reinforcement learning, derived from predictive HCI models exclusively as well as predictive HCI models augmented with human feedback. While I find the idea interesting and useful I have strong doubts about the defined context (the paper does not define which form of "adaptivity" will be studied) and the operationalization of user satisfaction using SUS. Here my detailed comments:

R3. Comment 2: Introduction

- "However, suggest the right adaptation" should be "However, suggesting the right adaptation".
- The reference is missing in "been proved to be successful [x]."
- Why the square bracket in "[Reinforcement learning...]"? I suggest to use round ones.
- At the end of the Introduction, it is still not clear to me what these HCI models represent/contain. Maybe a toy example would be useful.
- The report does not define which "Adaptivity" will be studied. Are we talking about the catastrophic disappearing adaptive menues of Office 2000 (see e.g. [2], in which the authors discuss that adaptive menues hinder users to develop habits), moving buttons, "Perspectives" as they are used in Eclipse, etc? I think that the chosen form of "adaptability" will be fundamental to the experiment since it will have an impact on user satisfaction.

R3. Comment 3: *Background*

- Fig. 1 is not clear to me. It seems that time is depicted horizontally (t0, t1, tn) and vertically (t0, t1, t2). This is all very confusing. If I understand it correctly, you want to show that there are alternative ways to adapt the UI over time? Then the icon for the initial UI should be depicted on the top node of the tree? And the adapted UIs can be depicted in small next to different nodes? Anyway, this diagram does not follow any standard notation (e.g., UML), therefore a legend is required to understand it. A black arrow seems to have two meanings: horizontally it depicts "sequence", vertically from top to down it depicts "passing of time". Please recreate a clearer version of this figure and clearly depict what it should explain to the reader. Moreover, after reading the text, the picture does not describe at all the four steps described, i.e., also the text needs to be adapted to what is depicted.

Experimental design

R3. Comment 4: - Congratulations on formulating the goal following the GQM guidelines!

- You write "In contrast to task-oriented interactions, UX represents a focus to “experience”, focusing also on hedonic qualities, and positive emotions and affect (e.g., interested, enthusiastic, irritable) that people experience while interacting with software systems." It seems that you see UX as "different" from usabiltiy. Until now I had a view on UX as in [1], that UX ***includes*** usability. Therefore I would be careful of talking about a "shift". Nevertheless, after reading further, I realize that you include "utility" in UX since you operationalize UX using "user engagement and user satisfaction" (page 4) and you define "user satisfaction is a measure of the quality of the user experience and refers to the extent to which a user’s expectations are met".

R3. Comment 5: - I agree with your choice of measuring user engagement but I disagree in using SUS to measure user satisfaction/usability. I find SUS useless since it is overly generic. I think that using SUS would result in a low external validity, since it evaluates usability at a very abstract level. Moreover, I have always felt that participants often do not really understand what to answer (e.g., to the question "I found the various functions in this system were well integrated.") or that some questions have to do with the work context of the participant (e.g., "I think that I would like to use this system frequently."). In the GQM you specify that you want to evaluate these UIs from the point of view of "a group of undergraduate and Master’s students in Computer Science at the Universitat Politecnica de Valencia interacting with user interfaces from the e-commerce and e-learning domains". For me - the question "I think that I would like to use this system frequently." does not make
sense for a course management system or a sports goods store if I am a student. I will not use these systems more than needed even if they have a fantastic UX. Moreover, I must use the course management system, otherwise I cannot pass the exams. What I want to say, I find SUS very imprecise and vague in this context. I would suggest to measure user satisfaction using task-based evaluations (measuring if users are able to accomplish tasks faster with the adaptive UI) or developing your own adaptation of SUS questions that evaluate the perceived complexity of the user interface, the perceived difficulty in accomplishing certain tasks, etc. ***in your particular context of adaptivity*** with questions that take into account the type of participants. This point is connected to my doubts above that I think that the concrete "adaptivity" that will be studied in this paper needs to be defined.

R3. Comment 6: - Fig. 2 a) apparently "shows the entire process of training an RL agent which uses MCTS, with reward models such as HCI models and with human feedback". You use a notation that is not familiar to me and neither has a legend that explains what the colors green, orange, black mean and what dashed and solid arrows represent. Please use BPMN to depict the process.
- Fig. 2 b) is never referenced in the text, the caption explains that it represents the execution plan. Since also this diagram is inconsistent (suddenly two larger arrows, boxes seem to represent systems but also user groups) I suggest to use BPMN to depict the execution plan.

R3. Comment 7: - Be careful: [3] suggested that comparing SUS scores can only occur ordinally. In other words, while SUS scores may indicate that one system is perceived as more usable than another, those comparisons would not be able to conclude how much more usable the system was.

[1] http://www.neospot.se/wordpress/wp-content/uploads/usability-vs-user-experience.jpg

[2] Hornbæk and Frøkjær: "Evaluating User Interfaces with Metaphors of Human Thinking", 2002
[3] Peres, S. C., Pham, T., & Phillips, R. (2013, September). Validation of the System Usability Scale (SUS) SUS in the Wild. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 57, No. 1, pp. 192-196). Sage CA: Los Angeles, CA: SAGE Publications.

----------------------- REVIEW 4 ---------------------
SUBMISSION: 9630
TITLE: A Comparative Study on Reward Models for UI Adaptation with Reinforcement Learning
AUTHORS: Daniel Gaspar Figueiredo, Marta Fernández-Diego, Silvia Abrahao and Emilio Insfrán Pelozo

----------- Overall evaluation -----------
SCORE: 1 (weak accept)
The goal of this study is «to analyze reward models derived from predictive HCI models and predictive HCI models augmented with human feedback in reinforcement learning algorithms with the purpose of assessing their impact to adapt user interfaces with respect to their ability to improve the user experience of software applications from the point-of view of both developers and researchers interested in reward modeling in RL-based methods for user interface adaptation.»

R4. Comment 1: The study may be categorized as mainly being confirmatory. The hypotheses do, however, not include a direction of the effects, which limits the strengths of the hypothesis tests, just the rather uninteresting hypothesis that there is a difference. That there is a difference is mainly a question of the power of the study, as all methods will have different effects given a sufficient number of participants. I guess that the study would not have been conducted if there was no effect from human feedback (similar to the other hypotheses), so why not make the hypotheses directed? To conclude that the hypothesis of no difference is rejected does not give much interesting information.
The dependent variable seems to be the use of predictive HCI models only and predictive HCI models augmented with human feedback, i.e., the same + a bit more information. The only way more information would not contribute positively is, however, when the information is misleading, irrelevant or incorrect. I may misunderstand something here, but is there any reason to believe that the addition of human feedback would not improve the dependent variables (such as the user experience)? If so, the question should perhaps rather be how much this would improve, rather than if there is a difference.

R4. Comment 2: Not being familiar with the concepts and topics of this paper I struggled to find information about how the interfaces were developed and what they will be able to represent when trying to generalize the results to other contexts (external validity). If not there, please add information and discuss the ability to generalize from the interfaces used in the experiments to other contexts.

R4. Comment 3: I also recommend you to add a power analysis to examine to what extent it is likely that you will find significant differences with interesting effect sizes given the number of participants you plan to include. It is common that software engineering experiments of this type have a very low power, typically that there is only 30-40% probability of finding a statistically significant difference even if there is one of medium effect size (using Cohen's d categories).