Human Factors and Simulation, Vol. 83, 2023, 150–157
https://doi.org/10.54941/ahfe1003572

Using Cognitive Models to Develop Digital Twin Synthetic Known User Personas

Audrey Reinert[1], Baptiste Prebot[2], Summer Rebensky[1], Don Morrison[2], Valarie Yerdon[1], Maria Chaparro Osman[1], Daniel Nguyen[1], and Cleotilde Gonzalez[2]

[1]Aptima, Inc., 12 Gill Street, Suite 1400, Woburn, MA 01801, USA
[2]Dynamic Decision Making Laboratory, Carnegie Mellon University, Pittsburgh, PA, USA

ABSTRACT

A recurring challenge in user testing is that it can be difficult to recruit a large enough sample population to reflect diversity of interaction behaviors within a given user persona. Although solutions to this challenge exist in the form of proxy users (i.e., individuals with abilities and skill sets similar to those of a desired group), the data collected from proxy users may not truly be reflective of actual users. One way to address this problem is to use digital twins of user personas to represent the range of decisions that could be made by a persona. This paper presents a potential use of cognitive models of user personas from a single complete record of a persona to test a web-based decision support system.

Keywords: Human factors, Human-systems integration, Systems engineering, Machine learning, Data science

INTRODUCTION

Human factors and user experience (UX) textbooks are filled with instances of design errors that range from mildly annoying (e.g., too many menus cluttering the screen) to potentially fatal confusion. The value of engaging in user testing is that repeated testing allows product development and design teams to identify and correct usability and design issues before the product is distributed. There are a variety of user testing methods a design team can leverage to identify these issues, ranging from card sorting (Ntouvaleti & Katsanos, 2022) to moderated usability testing (Larsen et al., 2021).

Depending on the design context, it can be difficult to recruit a large enough sample population to reflect the diversity of behaviors within a given user persona (Nunes et al., 2010). This challenge is relevant when assessing designs used by individuals with highly specialized skill sets (e.g., intelligence analysts) because these individuals have limited testing availability and or skillsets that are not readily found in proxy users. Although they can help identify mechanical shortcomings such as placement of menus, these proxy users cannot be relied on to consistently reproduce idiosyncratic behaviors of the primary users.

In the intelligence analysis domain, analysts are tasked with evaluating a large volume of data, including financial records or news reports, and making decisions about the data to formulate conclusions. As the magnitude of data to be analyzed increases, so does the cognitive load strain on the analyst. This increased workload can reduce performance, so we look to technology to support the analyst in some of the decision making and interpretation of relevant information. Recommendation systems are designed to assist analysts with real-time document suggestions according to a live relevance score while monitoring user interactions to model user intention and interest to determine which documents are most relevant. When there are a limited number of users to provide data to initially test with, these adaptive recommendation systems cannot adequately respond to the diversity of users.

Although data-driven personas, which use readily available data to uncover behavioral trends a researcher may not have captured through other means, can help to overcome this limitation (Salminen et al., 2021; Zhang et al., 2016), these personas are subject to sampling and availability bias. However, advances in the development of digital and cognitive twins present a potential solution to this problem. The synthetic data as a digital twin reflecting a user persona type's behavior is the applied practice proposed here. Digital twins have been used as digital representations of intended or actual real-world physical systems that serve as digital instances for testing purposes. Instance-based learning theory (IBLT) is a general postulation of the processes and mathematical mechanisms globally applicable to dynamic decision-making tasks (Gonzalez et al., 2003). We believe that cognitive models, like those built with IBL, could help create synthetic data that optimally represent actual users.

In this study, three participants were tasked to interact with an intelligence analysis recommendation system, ALFRED the BUTLER (ATB), while adopting the behavior of different user personas. Participants were asked to interact with ATB and provide feedback based on the pertinence of the recommendation to their intelligence search (Principal Information Requirement [PIR]). The interaction data were then used as a ground truth to evaluate the accuracy of a cognitive model that predicts user behaviors and its ability to reproduce specific persona decision sequences.

In this paper, we present the ATB system and the method used to establish and operationalize three user personas, followed by an overview of IBLT and the computational methods used to create the IBL cognitive model of the user personas. We first present the results of the cognitive model predictions and close with a discussion of future applications, one of which is using this type of cognitive modeling to help create synthetic datasets of persona behaviors to assist in trust calibration.

ALFRED THE BUTLER AND PERSONA CREATION

Alfred the Butler

Intelligence analysts are responsible for filtering a large volume of text-based information in a short period of time. ALFRED the BUTLER (ATB) is a recommendation system designed to assist intelligence analysts by enabling them to filter the articles they are presented with. As with other recommendation systems, an ATB user can give an article a thumbs up to indicate that they want more articles like that or a thumbs downs to indicate that they want to see fewer articles like that.

ATB users search for relevant information on a specific topic in a database of articles by entering natural language sentence-based searches such as, *What is the projected growth rate of personal IoT devices in the U.S.?* ATB computes a relevance score for every article in the database and presents the user with classified suggestions presented as cards (see Fig. 1) that can be opened to consult the entire content of the article. Cards are presented 8 at a time, ordered by relevance.

Persona Development and Operationalization of Behaviors

The team developed a set of three user personas to help guide their interactions with the ATB system: (1) Disuser, (2) Early Terminator, and (3) Feature Abuser. These personas were designed to represent different types of human-automation user interactions as outlined by Parasuraman & Riley (1997). The differentiation between persona types was based on behavioral measures using features and attributes of an online search assistant. The research team operationalized the definitions of use, misuse, disuse, and abuse, to fit the current context. In the context of this project, these terms have the following operational definitions:

- *Use*: Using the feature/tool as intended with the intended frequency
- *Misuse*: Overreliance on a recommendation system
- *Disuse*: Not using a particular feature/recommendation
- *Abuse*: Using automation and features for non-intended purposes

Each persona had a different goal and termination condition when using ATB, as follows:

- *Disuser*: A user who performs a certain action (either thumbing up or thumbing down) less frequently than expected, providing ATB with insufficient information regarding what to show the user.

Figure 1: Close up of card (article recommended by ATB).

- *Early Terminator:* A user with a specific goal who stops a search once the goal is met, e.g., stops thumbing up or down articles after thumbing up 50 articles.
- *Feature Abuser:* A user who uses a feature for a non-intended purpose, such as using the thumbs up action as a replacement for the save action.

Persona Interaction Data Gathering

Interactions with ATB recommendations included deciding to upvote or downvote a card, indicating the subjective relevance of the card, or saving it for future evaluation. In the context of this experiment, users needed to provide a single interaction. Initially, participants were attributed one specific persona and provided with a description of their intended behavior. Participants were then put into the situation of an Intelligence analyst looking for information concerning the following Principal Information Requirement (PIR):

- *What is the projected growth rate of personal IoT devices in the U.S.?*
- *Are there large-scale techniques to protect IoT devices within the U.S. from global adversaries?*

We recruited three participants from Aptima to act as one of the user personas listed above. These three user personas were selected because they provided the greatest breadth and depth of different interaction behaviors that could be used for model generation. Each participant was asked to upvote, downvote or ignore 280 different recommendations from ATB (35 sets of 8 articles).

COGNITIVE MODEL OF USER PERSONAS

Instance-Based Learning Theory (IBLT)

The general cognitive decision process proposed in the IBLT (Gonzalez et al., 2003; Nguyen et al., 2022) suggests that decisions are represented in the form of instances involving three parts: state, action, and utility. In general, state is a representation of the features of a decision situation, action is a decision an agent makes in such a state, and utility is an expectation the agent generates from experience or an outcome the agent observes because of such action. The theory generally assumes that instances accumulate over time, and past instances are recalled based on their similarity to a current decision situation. The expected utility of each decision alternative is generated as a function of the utilities in past similar instances and the probability of retrieving those instances from memory. A choice is made for the option that has the highest expected utility. The corresponding mathematically concrete algorithms have been presented in multiple papers. We refer the readers to a recent publication (Nguyen et al., 2022), which explains the mathematical formulations used in this theory.

IBL Model of Personas

IBL models are particular computational implementations of the IBLT process applied to specific tasks. In this case, we defined an IBL model that acts as a user analyst, receiving input from ATB and processing the cards obtained in sequential order. Although our IBL model was general for all three personas tested, it updated its predictions of the actions to take in the task (up, down, ignore), according to the actions taken by the human testers. The specific instance representation and process for the human tracing method is explained below.

The instances adopted a structure as shown in Fig. 2. Instance attributes are generally contextual features that the user would consider when making the decision, such as article title, content, keywords, source, and date of publication. Usually, instances are constructed to resemble the information that would be presented to the user, such as the source or date of the article recommended. Here, however, some attributes directly reflect the similarity one can evaluate between some features, like the similarity between the PIR and the article title, and the similarity between the content of the article and the PIR. These are supposed to be judgments made by the user as part of the decision-making process.

For the first two attributes, the card (title and content) and PIR were broken into words (ignoring noise words such as “the” or “and”). Then, the fraction of words contained in the PIR that also appear in the name, and respectively in the content of the card, was calculated to give a sense of their similarity.

When the model was run, the values of the eight attributes shown in Fig. 2 were compared to one another by partial matching. All but the source are numeric attributes, and were compared linearly over the range of possible values (i.e., 0 to 1 for most, though the confidence and date have a larger upper limit). The source was compared such that if identical it had similarity 1 and otherwise 0. The model assigned a utility of 1 to a correct prediction, and 0 to an incorrect prediction. It was also primed with three prepopulated instances, each with a utility of 1, and each with other attributes set to be maximally matching.

Simulation Methods and Metrics

Each of the three persona was tested in a different simulation. We ran 100 IBL runs, over the first 280 recommendations. For each persona, the model examined each recommendation (or card) in turn, in the order in which they were

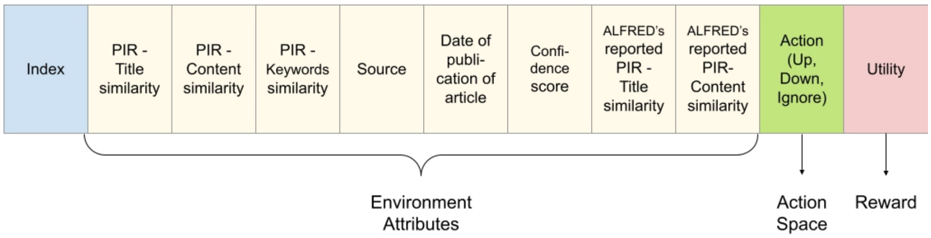

Figure 2: Instance structure of the IBL model.

reacted to by the user. All IBL models were run with default decay d = 0.5, noise $\sigma = 0.25$, and a mismatch penalty of 0.75.

For each persona, a cumulated accuracy score was computed that represented the average proportion of correct predictions of the model, after the 280 recommendations, and over the 100 runs.

RESULTS

User Persona Testing

Table 1 shows the following results. First, the Feature Abuser provided almost double the number of thumbs up responses than the other two users. The Disuser rarely made no choice and instead used the thumbs down option almost exclusively. From the perspective of the ATB system, the Disuser and Feature Abuser provide imbalanced information for ATB to learn from. This can lead ATB to over-exclude articles or fail to include articles from the search because of a lack of counterbalancing signals. Abstracting these results to different behaviors suggests similar consequences. For example, a user who rarely provided a thumbs down would render the system unable to clearly determine what the user did not want to see. This led to the conclusion that a notable disparity in the types of signals will naturally lead to the creation of an "unbalanced ALFRED."

IBL Model

For each persona, the mean final cumulated prediction accuracy of the model is reported in Table 2 and Fig. 3.

Mainly, we observed a lower accuracy at predicting the Early Terminator decisions (0.56), whereas the model was able to predict the persona behavior with 0.68 accuracy for the Feature Abuser and 0.79 accuracy for the Disuser.

Table 1. Overview of the interactions made by the different users.

	Total Interactions	
	Thumbs Up	**Thumbs Down**
Disuser	47	226
Feature Abuser	155	10
Early Terminator	51	33

Table 2. Average final cumulated prediction accuracy from the IBL model per persona type over 100 runs.

	Cumulated Accuracy	
	Mean	**SD**
Disuser	0.79	0.01
Feature Abuser	0.68	0.01
Early Terminator	0.56	0.14

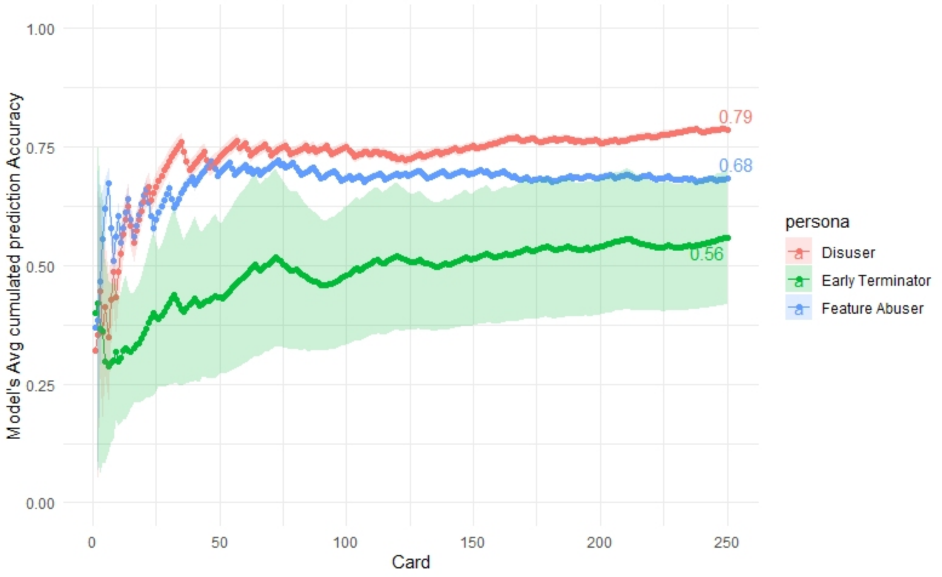

Figure 3: Cumulated prediction accuracy of the IBL model for each of the persona: Disuser in red, early terminator in green, feature abuser in blue.

DISCUSSION

A user experiment, in which three participants posed as personas using ALFRED the BUTLER, allowed us to gather three datasets of user decision behaviors. Based on these three datasets and using an IBL model, we created 100 new datasets per persona of decision data that could be used to train other learning algorithms. The results of this study demonstrated that a synthetic user persona can be generated using sample data and that the synthetic persona behave in a similar manner to the real user.

Although prediction accuracy for the Early Terminator was quite low for a prediction system (0.56), it was still much higher than a random choice (0.33). A possible explanation for why the prediction accuracy for Early Terminators was lower than that for the Feature Abuser or Disuser is that the Early Terminator, by definition, has a fixed goal but the number of interactions needed to reach that goal is different than it is for the Feature Abuser or Disuser. That is, an Early Terminator could interact with 100–200 cards before stopping whereas the Feature Abuser and Disuser each review all cards. This results in datasets for generating the Early Terminator Model that contain imbalances that are not present in the other models.

There are two important implications that stem from this work. First, we have shown that a single dataset of interactions can be used to generate synthetic twins of that persona that are statistically similar to the original set. This allows researchers to create a diverse range of responses that describe a user from one record. Second, the practical implication is the ability to reduce data-collection costs through the deployment of synthetic personas. Because different analysts operate within different time constraints and resource requirements, recommendation systems such as ALFRED the BUTLER will need to adapt to users in order to gain and keep their trust. In practice, this can be achieved by ATB inferring what type of user or persona is interacting with it and adjusting recommendations accordingly to dynamically calibrate trust. Cognitive models and synthetic personas could well be of help to generate enough diverse interaction data to train these recommendation systems in this endeavor.

ACKNOWLEDGMENT

The authors would like to thank Mr. Gabe Ganberg and Dr. Don Nelson for their technical contributions to this paper.

REFERENCES

Gonzalez, C., Lerch, J. F., Lebiere, C.: Instance-Based Learning in Dynamic Decision Making. Cog. Sci., 27(4), 591–635 (2003).

Larsen, L. B., Øvad, T., Nielsen, L., Larusdottir, M.: Remote User Testing. IFIP Conference on Human-Computer Interaction, 579–583, volume 12936 (2021).

Nguyen, T. N., Phan, D. N., Gonzalez, C.: (2022). SpeedyIBL: A Comprehensive, Precise, and Fast Implementation of Instance-Based Learning Theory. Behav. Res. Meth., 1–24 (2022).

Ntouvaleti, M., Katsanos, C.:. Validity of the Open Card Sorting Method for Producing Website Information Structures. CHI Conference on Human Factors in Computing Systems Extended Abstracts, 1–7, New Orleans, LA, USA (2022).

Nunes, F., Silva, P. A., Abrantes, F.: Human-Computer Interaction and the Older Adult: An Example Using User Research and Personas. Proceedings of the 3rd International Conference on PErvasive Technologies Related to Assistive Environments, 1–8, Samos, Greece (2010).

Parasuraman, R., Riley, V.: Humans and Automation: Use, Misuse, Disuse, Abuse. Hum. Factors, 39(2), 230–253 (1997).

Salminen, J., Jung, S.-G., Jansen, B. J.: Are Data-Driven Personas Considered Harmful?: Diversifying User Understandings with more than Algorithms. Persona Stud., 7(1), 48–63 (2021).

Zhang, X., Brown, H.-F., Shankar, A.: Data-Driven Personas: Constructing Archetypal Users with Clickstreams and User Telemetry. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 5350–5359, San Jose, California, USA (2016).