Designing AI Systems for Mental Model Development

Michael Hoefer[1]

[1]University of St. Thomas, 2115 Summit Ave, Saint Paul, MN, 55105, USA

Abstract

Users of interactive systems form mental models of how those systems work, and they operate those models during system interaction. AI-augmented systems exhibit characteristics that make forming accurate mental models with high predictive power more difficult. In particular, AI-augmented systems that utilize large language models (LLMs) exhibit more uncertainty, natural language interaction, and dynamism than traditional (non-LLM based) interactive systems. In this paper I discuss each of these characteristics and how they impact the formation of mental models. Drawing from a user study of an AI-augmented calendar, I highlight how participants' interaction styles with the system appeared to be influenced by their mental models. I close with design considerations that may help AI-augmented systems better support the development of user mental models.

Keywords

Human-AI interaction, Human-AI collaboration, Human-AI teaming, mental models, cognition, interactive system design, large language models, HCI theory

1. Introduction and Background

The concept of "mental models" has long been a topic of inquiry for HCI researchers. Traditionally, designers would have complete knowledge of the underlying system, and they would seek to present a "system image" to the user. This image represents the specific underlying mechanisms that compose the system, in a way that supports the user learning the system. This learning involves construction of a "mental model" of the target system. The user will then interact with the system in a way that corresponds with their understanding of the system, or in other words, they will operate their mental model of the system in order to make decisions in how they interact with the system. Norman refers to this as the "predictive power" of a mental model [1].

The introduction of large language models (LLMs) as a component of interactive systems poses a new challenge for designers. Previously, the "conceptual model" was known by the designer, and designing an interactive system was focused on the user and helping the user create their mental model. The training details of many LLMs, including the sources of the training data, are proprietary. Even open-source LLMs are necessarily highly complex "black box" systems which makes model interpretation extremely challenging [2]. Thus, LLMs add an element of unavoidable uncertainty into the operation of the system, which makes constructing mental models of AI-augmented systems more challenging and requires specific consideration in the design process. In the rest of this paper, I will discuss ways in which AI-augmented systems present new challenges for the formation of mental models of users. I will present a case study of an AI-augmented calendar, and highlight how users' mental models impacted their interaction with the system and provide design recommendations to support mental model development.

2. Related Work

While a full literature review is beyond the scope of this paper, I will briefly note that there is research discussing mental models in the context of AI. One approach of interest is that of expressing *explainable* AI in terms of mental models. Merry et al. suggest that much existing work on explainable AI focuses on

Joint Proceedings of the ACM IUI Workshops 2025, March 24-27, 2025, Cagliari, Italy
michael.hoefer@stthomas.edu (M. Hoefer)
https://michaelhoefer.com/ (M. Hoefer)
0000-0002-9407-8145 (M. Hoefer)

computational methods that do not take into account the context of use of the system ("understandable *by whom*" [3]). Explainable AI is AI where the *users* of the AI system form, maintain, and use accurate mental models of the system, in the context of the system's use. This approach to explainable AI corresponds with the research ethos of the HCI community, which focuses on understanding the user and the user's needs in the context of use (*in situ*).

There has also been discussion about treating AI-augmented systems as "agents," and drawing from literature on mental model development and maintenance in team situations. In this case, the AI is seen as another member of the team. The human-AI teams may form shared mental models (SMMs) that represent the collective understanding of the task at hand [4]. The adaptation of the system, in response to a particular user or users, could also be thought of as the AI itself having a "mental model" of the user. While this might be over-anthropomorphizing, the opacity of the inner workings of LLMs does perhaps mirror the illusive nature of mental models that humans hold of systems.

Readers interested in a more thorough literature review can read the work of Andrews et al, who explore various concepts about mental models and how they relate to human-AI teams [4].

3. Characteristics of AI-Augmented Systems Relevant to Mental Model Formation

What makes AI-augmented systems "different" enough to warrant special consideration with regards to mental model development? Traditionally, designers would work with a *conceptual model* of a target system, and seek to design an interface that naturally communicates this model to the user, such that they form an accurate mental model. According to Norman, this conceptual model should be an "accurate, consistent, and complete" representation of the target system [1]. When part of the system involves stochastic output from an LLM, creating a "complete" conceptual model becomes much more difficult or even impossible. If we do not have an complete conceptual model of LLMs, how can we design interfaces that help users form complete mental models of the systems they are using?

I suggest the following traits of AI-augmented systems motivate special considerations for design to support the development of mental models of users.

- Uncertainty
- Natural Language Interaction
- Dynamism

3.1. Uncertainty in LLM Output

One primary way in which users form mental models of systems is by interacting with the system. Users identify patterns and form mental models based on repeated interactions. In an AI-augmented system, repeated interactions of a similar nature may result in different outputs. This presents a new challenge in mental model formation, as users may be less able to predict the output of the LLM or the behavior of the AI-augmented system.

3.2. Natural Language Interaction

AI-augmented systems may be more likely to utilize natural language interaction. Users may be more likely to ascribe anthropomorphic qualities to systems that utilize natural language interaction, thinking of the system as more human-like due to natural language being a social cue [5]. This may lead to a more complex mental model, or a mental model that contains more human-like qualities than the system actually has. Beyond the effects these mental models may have on task performance, according to Lombard et al., when humans over-anthropomorphize systems, they could even be considered to be "victims of deception, unconscious responses, and the manipulation of presence" [5].

3.3. Dynamism

AI-augmented systems can change their behavior over time for multiple reasons. LLM base models are being continuously improved, and if base models underlying AI-augmented systems are also updated, then the user may be interacting with a system with entirely different "brains" but the same old "skin." There is a challenge in helping the user adjust their mental models of systems that have an increasing capability over time. Moran presents research that highlights how users of intelligent assistants (such a Siri or Alexa) internalized those system's poor capabilities early on, and then later, after the system had improved, did not "push the interaction limits" of those systems [6]. Presumably, these users had "written off" the capabilities of these assistants. Perhaps one reason is that the system did not support the user in updating their mental models to include the system's newly added capability.

4. Case Study: An AI-Augmented Calendar

The following case study is about an AI-augmented voice interface to a calendar, and is intended to show how mental models play a role in human-AI interaction.

4.1. System Design and Evaluation

TellTime is a calendar interface intended to help users collect data about how they spend their time, using spoken natural language rather than (or in addition to) manual calendar modifications [7]. The intended audience of the system ranges from individuals engaged in self-tracking and time management, to researchers who wish to conduct large scale time-use studies, such as the American Time Use Survey (ATUS) [8] or Multinational Time Use Study [9]. Self-reports of time use are historically captured manually using paper diaries, phone interviews, or electronic systems requiring manual data entry. These methods tend to place a significant cognitive burden on the users (or those surveying them).

The TellTime system was designed in order to reduce the burden of gathering time-use data, both for individuals and researchers. TellTime supports hybrid human-AI interaction, enabling both spoken natural language and manual interaction. The manual interaction is similar to that of commercially available calendar systems (click and drag to create an event, for example, or click to modify an event's details). The voice interface also enables modifications to be made to the time-use record, either via additional spoken commands or via manual interaction. The users therefore have a choice between interacting with the system either via voice or manually, making it a good test system to study human-AI interaction.

An evaluation study was conducted with 18 participants, in the form of a qualitative randomized controlled trial investigating a fully manual interface (no AI), a fully voice-AI interface (AI only), and the hybrid interface. Participants shared their experience after each session, and completed a closing interview. In addition, participants were asked to "think-aloud" their thoughts as they interacted with each version of the system, which was useful to help get a sense of the mental models participants may have had [10, 1]. The fully voice-AI interface was useful for provoking interaction styles in "frustrating" situations, as the participants could not make manual modifications to the calendar events in that version.

4.2. Influence of Mental Models on Interaction

Participants exhibited a variety of interaction styles with the system that appeared to correlate with their mental models of the system. Some participants communicated with very simple, short statements such as "9 am ate breakfast" and then "10 am got ready for the day." Others were very comfortable speaking in long narratives, which included events happening out of order, side comments, and revisions to events mentioned earlier in the narrative. The system was generally quite capable at handling these narratives and parsing them into activities (see Figure 1 for one example).

One of the overly cautious participants said "if I was talking to a person about my day, I would speak in that way where I kinda just like talk about all the things I did that day as they came in my brain. But

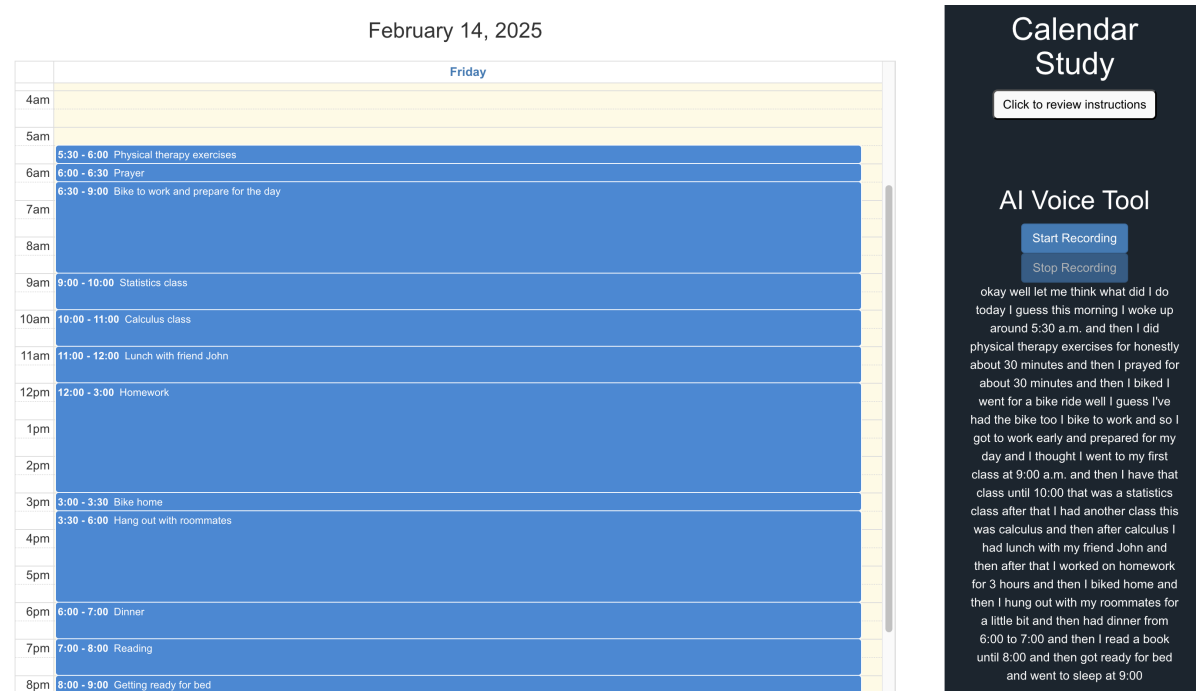

Figure 1: A screenshot of the AI-powered calendar interface, TellTime. Users can record a spoken narrative of their daily activities, which are parsed into structured events using an LLM, and shown on the calendar.

for some reason when I'm like working with a system like this...I have to stick to like 'I did A and then B, and then C, and then d"' (P1). P1 also expressed how their previous interactions with Siri led them to feel "rushed" in their interaction, saying "I'm so used to like the virtual assistants like Siri, or something where like, if you stop for 5 seconds, it's like, 'Huh? I'm listening' and kind of like nudges you. So I have to like force myself to not like rush, and like, actually think through [how I spent my time]"

Another participant adopted mental models of the system's capability based on previous interaction with Zoom's transcript feature, leading them to the conclusion, "I just assumed that I needed to...I needed to *speak like a computer* in order for it to be able to know, to put it in there [the calendar]" (P4).

The participants were recruited from a convenience sample that included some computer science students, and some participants recruited from social media who had less experience with technology. Perhaps ironically, it was the computer science students who were the most cautious with the system. Participants with less technology experience seemed more trusting and willing to experiment with the system's features, treating the experience more similar to "speaking to a friend."

After interacting with the system further, some participants did appear to adapt their mental model and begin to use a more conversational, narrative based interaction style in their voice commands. Some participants specifically "played around" with the system, giving "bogus" commands to test the limits (such as shifting the start time of all events by one hour) and understand the capability of the system.

Participants in the TellTime evaluation study also expressed anthropomorphic beliefs in the system. For example, P12 said "I felt like I was talking to someone telling them about my events of the day." P8 started an edit by greeting the AI, saying "oh hey, let's make a change." P9 explicitly said, "I was kind of thinking as the AI as just kind of like a friendly, helpful little guy in the computer if I'm being honest."

5. Design Considerations for Supporting Mental Model Development

This section contains design considerations that arose after studying the 18 participants interact with the AI-augmented calendar system.

5.1. Understand the Sources of Default Mental Models

Participants' previous experiences with other interactive systems "muddied" the mental model they formed of the TellTime system. This suggests one heuristic: *users will inform their mental model of a new system based on their experience with systems that had similar interaction modalities.* In this case, the transcription feature of both Siri and Zoom led participants to be more cautious and "speak like a computer" when they used the TellTime system.

A recommendation for designers is to consider common systems that users may have previous experience with, and in particular, those who share interaction modalities (in this case, speech recognition). Designers can create affordances or training opportunities that provide interaction experiences that directly contradict their previous experience. In the case of TellTime, a video could have been provided that shows a user speaking a complex narrative, and the system correctly parsing the events. In addition, the system may be able to intelligently identify when the user is "talking like a computer" to it, and provide feedback indicating that they can "talk normally."

5.2. Challenge Previous Assumptions of Non-Augmented Tasks

Another source of initial mental model formation is from the task itself. When participants are asked to complete a task, their conceptualization of the steps necessary to complete the task will be based on their previous (successful) completions of the task. In the case of creating events on a calendar, P1 interacted with the AI system in the same way they would complete the task without the AI system, saying "If I was to, like, schedule my day ahead of time, how would I do it? And so I was thinking...let me, like, make this [single] event." As a human, they would reconstruct their calendar event-by-event, and so they asked the AI to do the same.

When augmenting a traditionally manual task with AI, it may be important to *consider the users' prior mental models of the* ***task itself***, and challenge these mental models which may be less useful in light of the more capable AI-augmented system.

5.3. Provide Intelligent Playgrounds or Sandboxes

While some participants naturally played around with the model to test it's capabilities, most did not. Designers could create "playgrounds" or "sandboxes" where users can safely experiment with different ways of interacting with the system. This can act as a kind of "training ground."

While simply providing a worry-free environment could help, it may be even better to scaffold the play by providing examples that showcase the limits of the system's capabilities. For example, a sandbox for the TellTime system could contain a complex narrative as a prompt, and allow the user to run the narrative through the system to see the output. The user could then make modifications to the narrative to see how the output changes.

6. Conclusion

The incorporation of "black box" LLMs into interactive systems makes modeling the system, both for the designer, and the user, more difficult. Specifically, AI-augmented systems tend to include more uncertainty, natural language interaction, and dynamism that increase the difficulty of forming accurate mental models with high-predictive power. I discuss how some of these issues arose in a user-study of TellTime, an AI-Augmented calendar system for gathering time-use data, and present design considerations to support mental model development.

Declaration on Generative AI

The author has not employed any Generative AI tools.

References

[1] D. A. Norman, Some observations on mental models. Mental Models, Lawrence Erlbaum: (1983) 99–129.

[2] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, M. Du, Explainability for large language models: A survey, ACM Trans. Intell. Syst. Technol. 15 (2024). URL: https://doi.org/10.1145/3639372. doi:10.1145/3639372.

[3] M. Merry, P. Riddle, J. Warren, A mental models approach for defining explainable artificial intelligence, BMC Medical Informatics and Decision Making 21 (2021) 344. URL: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-021-01703-7. doi:10.1186/s12911-021-01703-7.

[4] D. S. Robert W. Andrews, J. Mason Lilly, K. M. Feigh, The role of shared mental models in human-ai teams: a theoretical review, Theoretical Issues in Ergonomics Science 24 (2023) 129–175. URL: https://doi.org/10.1080/1463922X.2022.2061080. doi:10.1080/1463922X.2022.2061080.

[5] M. Lombard, K. Xu, Social Responses to Media Technologies in the 21st Century: The Media are Social Actors Paradigm, Human-Machine Communication 2 (2021) 29–55. URL: https://stars.library.ucf.edu/hmc/vol2/iss1/2/. doi:10.30658/hmc.2.2.

[6] K. Moran, Mental models of ai assistants, https://www.nngroup.com/articles/mental-model-ai-assistants/, 2023. Nielsen Norman Group. Accessed: 2025-01-15.

[7] M. Hoefer, M. Gong, R. Rychecky, S. Voida, TellTime: An AI-Augmented Calendar with a Voice Interface for Collecting Time-Use Data, in: 30th International Conference on Intelligent User Interfaces (IUI '25), ACM, Cagliari, Italy, 2025. doi:10.1145/3708359.3712116.

[8] B. of Labor Statistics, American Time Use Survey User's Guide, 2023.

[9] J. Gershuny, M. Vega-Rapun, J. Lamote, Multinational time use study, 2020.

[10] C. Lewis, Using the "Thinking-Aloud" Method in Cognitive Interface Design, Technical Report RC 9265, IBM T. J. Watson Research Center, Yorktown Heights, NY, 1982.