Thesis Summary: Operationalizing User-Inclusive Transparency in Artificial Intelligence Systems

Deepa Muralidhar
Georgia State University
dmuralidhar1@gsu.edu

Abstract

Artificial intelligence system architects can increase user trust by designing systems that are inherently transparent. We propose the idea of representing an AI system as an amalgamation of the AI Model (algorithms), data (input and output, including outcomes), and the user interface with visual interpretations (e.g. graphs, Venn diagrams). By designing human controls and feedback mechanisms for AI systems that allow users to exert control over them we can integrate transparency into existing user interfaces. Our plan is to design prototypes of transparent user interfaces for AI systems using well-known usability principles. By conducting surveys we will study their impact to see if these principles help the user to work with the AI system with confidence and if the user perceives the system to be adequately transparent.

Introduction

A transparent system that follows standards where the explanations state what is at stake allows users to have a better understanding of the AI system. Machine learning-based AI systems show discriminatory behavior which often impacts marginalized groups. A system that explains its outcomes to users improves the overall user experience of the system. Artificial intelligence based systems have two distinct characteristics. One is that these AI systems operate as black boxes, the other is the uncertainty in their predictions. Explanations must be communicated to users in ways that clear their doubts related to these concerns. User interface design guidelines can improve communication between AI-human teams. Rather than emulating humans, AI systems must empower humans and explicitly outline the AI agent's and human's responsibilities. We implement this idea by designing prototypes for user interfaces for two classes of AI systems where the interfaces only reveal explanations as requested by the user. Such user-inclusive transparent AI systems are sensitive to the user's needs. We will analyze the results by surveying a controlled group of users.

Motivation. Artificial Intelligence as a technological innovation has the potential to impact society. The issue, however, is that AI experts are unable to explain the uncertainties and biases that are seen in these systems. Operationalizing transparency can make AI systems more trustworthy. Effective human-AI agent interactions increase transparency through explanations(Eiband 2018). We argue that holding the view that explanations are essential to transparency tells only part of the story. Interpretability and explanations that model how humans explain decisions to each other assisted by user-friendly interfaces increase system transparency. The design of user interfaces for AI systems that follow usability principles is an open research area.(Fig.1.1)

Related Work. Transparency is *"the ability of the AI system to be able to convey clearly what it can and cannot do"*(Weller 2019). Due to the "persistence of mental model",users apply their prior knowledge while using the system. Users thus do not perceive the system to be transparent because of a difference in the user's perception of the system and the actual conceptual model of the system(Norman 2002). Explanations that answer questions such as *what*, *where*, *how* and *why* an AI system arrived at a decision help build transparent AI systems. Experiments conducted on context-aware systems found explanations of "why the system" and "why not" explanations were useful in increasing user trust(Lim 2009). Design of AI systems that have interfaces that are mixed-mode and interactive(Horvitz 1999) lets users control the trigger to the AI system responses. They demonstrate *progressive disclosure*(Springer 2018) and provide information and explanations only when the user asks for it. Established usability guidelines such as *Keep users in control* and *Design dialog to yield closure*(Shneiderman 1986) can help address the concerns around transparency.

Thesis Workplan

My doctoral thesis focuses on establishing some of the usability principles to be followed while building a transparent AI system. We intend to design a user interface such that an AI system customizes itself to the needs of user. The information the system reveals to a user is subject to the user's feedback. We propose to do this by providing multiple modes of transparency that users can choose from, and the choice of mode aids the system in being judicious about providing explanations about its decision making process.

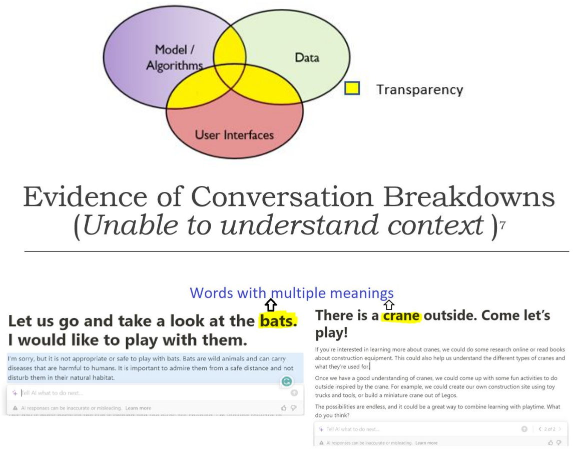
Preliminary work findings. Users rate an AI system more positively if they see an AI system providing explanations for its decisions, even if the explanations are just a placebo

Fall 2023	Spring 2024
September, October: Design prototypes for AI Text Editor (Class A system)	**January:** Improvise paper and prototypes as needed.
October: Design and conduct survey	**February:** Design prototypes for Clinical diagnosis software(Class B system)
November: Collect and analyze the data. Improve the prototypes. Write the paper.	**March:** Collect and analyze the data. Compare the findings between Class A and Class B systems.
December: Improve prototypes and complete paper.	**April:** Complete writing the paper.

Table 1: Timeline

or simply *persuasive*. But explanations that increase understanding and bring the user's mental model closer to the actual working model improve transparency. Explanations that provide information about individual predictions helps the user know how much to rely on the system. My research agenda is to inquire into the elements of transparency that influence an AI system. Usability guidelines used in software engineering when applied to AI systems will allow humans to collaborate and interpret AI models as well as provide explanations for the outcomes. We postulate that **Adequate transparency** is not an absolute value. It is relative and dependent on the user's requirement at the time and their expertise (mental model) of the system. Users can and should be able to request selective transparency(Springer 2019) from the AI system or exhaustive transparency depending on their requirements. While exhaustive transparency will most likely be an elusive goal, selective transparency is attainable. This is our target.

Proposed Research Plan. Our next research question is to learn if implementing usability principles such as *Determine user skill levels*, *Keep users in control*, *Design dialog to yield closure*, *Offer informative feedback* in AI systems will improve transparency for end users. We will design prototypes of transparent versions of the user interface for two classes of AI systems, Class A, auto-text generators, and Class B, clinical diagnosis AI systems. We will experiment with ten zero-shot and one-shot prompts used as inputs. Following the usability guideline of customizing the interface to user needs we will design an AI system having two modes of operation. A ***regular mode*** for a task-oriented end user, where the AI system is selective about what information it provides initially, especially when there is a conversation breakdown (Fig 1.2)(Li 2020). The user can decide if they want more information by interacting with the system. A ***transparent mode*** is for the more advanced user who intends to audit the system. Then exhaustively disclosing the inner workings of the AI system is imperative(Springer 2019). Through surveys, our goal is to assess if faith in the system increases with the transparent version of the user interface compared to the non-transparent version.



Evidence of Conversation Breakdowns
(*Unable to understand context*)7

Words with multiple meanings

Let us go and take a look at the bats. I would like to play with them.

I'm sorry, but it is not appropriate or safe to play with bats. Bats are wild animals and can carry diseases that are harmful to humans. It is important to admire them from a safe distance and not disturb them in their natural habitat.

Tell AI what to do next...

AI responses can be inaccurate or misleading. Learn more

There is a crane outside. Come let's play!

If you're interested in learning more about cranes, we could do some research online or read books about construction equipment. This could also help us understand the different types of cranes and what they're used for.

Once we have a good understanding of cranes, we could come up with some fun activities to do outside inspired by the crane. For example, we could create our own construction site using toy trucks and tools, or build a miniature crane out of Legos.

The possibilities are endless, and it could be a great way to combine learning with playtime. What do you think?

Tell AI what to do next... 2 of 2

AI responses can be inaccurate or misleading. Learn more

Figure 1: Top Figure: Fig 1.1, Bottom Figure: Fig. 1.2

Anticipated Progress. As we follow our timeline (Table 1) our initial step is to design user-interface prototypes for a text editor such as Notion (a text editor that has an AI component built-in) We will conduct surveys with a controlled group to determine if our understanding of adequate transparency is accurate. We will analyze the data to get an understanding of user satisfaction and to identify the usability principles important to end users. We will improve the prototypes based on the survey results. Next, following the requirements and lessons learned from the surveys we will conduct similar experiments with Class B systems. Our plan is to publish final versions of the prototypes.

References

Eiband, M. e. a. 2018. Bringing Transparency Design into Practice. In *23rd International Conference on Intelligent User Interfaces*.

Horvitz, E. 1999. Principles of Mixed-Initiative User Interfaces. In *International Conference on Human Factors in Computing Systems*.

Li, T. e. a. 2020. Multi-modal repairs of conversational breakdowns in task-oriented dialogs.

Lim, B. e. a. 2009. Why and Why Not Explanations Improve the Intelligibility of Context-Aware Intelligent Systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Norman, D. A. 2002. *The Design of Everyday Things*. Basic Books, Inc.

Shneiderman, B. 1986. *Designing the User interface. Effective strategies for Human-computer Interaction*. Addison-Wesley Longman Publishing Co., Inc.

Springer, A. e. a. 2018. Progressive Disclosure: Designing for Effective Transparency.

Springer, A. e. a. 2019. Making transparency clear. In *Algorithmic Transparency for Emerging Technologies Workshop*.

Weller, A. 2019. *Transparency: Motivations and Challenges*. Springer International Publishing.