Two or Three Things You Need to Know about AI Design

Myounghoon Jeon
Industrial & Systems Engineering/ Computer Science
Virginia Tech

AI system design requires different approaches from traditional product design because it is dynamically influenced by the trained data, user interaction, and environment. This visionary statement includes design constructs and methods that Human Factors designers and AI designers need to consider when designing AI systems.

Failures of Automation

Mode Confusion In automated systems, **function allocation** (who's responsible for what?) dynamically changes. For example, one of the most difficult design problems in automated vehicles is to let drivers notice what level they belong to. Display is needed when the transition is happening. But drivers also need to be reminded of it when there is no transition. Readers might experience mode confusion while using software, "Am I in the editing mode or performance mode?" In automated vehicles, the problem is much more serious because there are currently six different levels of automation and the outcomes of confusion can be fatal.

Bias & Discrimination Human Factors design's role in AI design would be contributing a **humanist perspective** that considers the social, political, ethical, cultural, and environmental factors of implementing AI into daily human-system interactions (Lau, Hildebrandt, & Jeon, 2020). Hamidi, Scheuerman, and Branham (2018) showed that how identifying an individual's gender by automatic gender recognition can intrude privacy and cause potential harms that can result from being incorrectly gendered, or misgendered by interviewing transgender participants.

Theories and Constructs

Mode confusion is related to **mental model**, which refers to an explanation of one's thought process about how the system works. Users can intuitively use the system only when designers' conceptual model reflects users' mental model. In automated vehicles, it is problematic to make users fully understand and memorize how six different modes are distinctively working and what the users are supposed to do in each mode. Of course, "Now, you are at level 3 automation" type display is useless. Another critical construct in AI design is **trust**. Parasuraman and Riley (1997) classified three relevant terms about automation, depending on perceived/true reliability and complexity of the system. If users are doubtful about frequent false alarms and do not use the system at all, it is "**disuse**". If users overtrust the system (complacency), it is "**misuse**". In this case, users' **situation awareness** and skills will be degraded. Because AI systems evolve based on the interaction with users, **situation awareness** is getting more and more important. Situation awareness emphasizes that users need to not only understand what is going on, but also predict what will be happening in near future. The second example above would be "**abuse**", meaning applications of automation without consideration of human side consequences. In terms of the interaction strategy to enhance trust, **anthropomorphism** is frequently adopted by designers. Research showed that people more prefer and trust agents like themselves (Eyssel et al., 2012). On the flip side of the coin, users might overtrust and expect more intelligence than the AI system actually has (**false mental model**). Also, some studies showed that people are frightened by human-like robot (Lohse et al., 2007) or its' facial expressions, due to the **Uncanny Valley** (Seyama, & Nagayama, 2007).

Methods and Approaches

In terms of design methods, there would be different phases Human Factors can intervene with in AI design. First is the human-centered design method for **generation of training data**. Considerations include sociotechnical impacts their AI system can bring about, such as algorithmic **fairness, bias, discrimination, trust, & transparency**. We can highlight the role of **humans as teachers** and their interaction with data with a key factor in building machine learning-based systems (Lindvall, Molin, & Lowgren, 2018). Also, **user corrections** of the machine learning predictions can be used to generate additional training data (i.e., **human in the loop**). At the same time, we need to emphasize that machine learning's reliance on data collected from human annotation and transcription makes it susceptible to the same biases that plague human cognition. Human Factors can also provide methods of exploring the consequences of design choices when creating AI systems (**Explainable AI**). For example, **visualization** can help users understand how features and their tweaking affect the prediction of the outcomes (= **mental model**) (e.g., Krause, Perer, & Ng, 2016). Then, how can we evaluate the system? One plausible method would be using **feedback models** describing how people react when receiving machine-aided decisions and how much the outcomes of the decisions made are desirable (Zhang, Khalili, & Liu, 2019). A recent study shows that **interpretability tools** are misused by data scientists and there is a need to improve their **mental models** of interpretability tools (Kauer et al., 2020).

Attitude or Mindset

In AI systems, key features evolve through interaction with users and the environment. Designers need to consider how information flows through these systems, how data can make operations more efficient and user experiences more meaningful, and how feedback creates opportunities for learning. However, designers will no longer craft the interaction per se between a user and a system. They will need to design the **meta-systems** which will design the system's interaction (Martelaro & Ju, 2018).

References

Eyssel, F., De Ruiter, L., Kuchenbrandt, D., Bobinger, S., & Hegel, F. (2012, March). ‘If you sound like me, you must be more human’: On the interplay of robot and user features on human-robot acceptance and anthropomorphism. In Proceedings of the *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 125-126). IEEE.

Hamidi, F., Scheuerman, M. K., & Branham, S. M. (2018, April). Gender recognition or gender reductionism? The social implications of embedded gender recognition systems. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1-13).

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020, April). Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-14).

Krause, J., Perer, A., & Ng, K. (2016, May). Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5686-5697).

Lau, N., Hildebrandt, M., & Jeon, M. (2020). Ergonomics in AI: Designing and interacting with machine learning and AI. *Ergonomics in Design, 28*(3), 3-3.

Lindvall, M., Molin, J., & Löwgren, J. (2018). From machine learning to machine teaching: the importance of UX. *Interactions*, *25*(6), 52-57.

Lohse, M., Hegel, F., Swadzba, A., Rohlfing, K., Wachsmuth, S., & Wrede, B. (2007, February). What can I do for you? Appearance and application of robots. *In Proceedings of AISB* (Vol. 7, pp. 121-126).

Martelaro, N., & Ju, W. (2018). Cybernetics and the design of the user experience of AI systems. *Interactions*, *25*(6), 38-41.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230-253.

Seyama, J. I., & Nagayama, R. S. (2007). The uncanny valley: Effect of realism on the impression of artificial human faces. *Presence: Teleoperators and Virtual Environments, 16*(4), 337-351.

Zhang, X., Khalili, M. M., & Liu, M. (2019). Long-Term Impacts of Fair Machine Learning. *Ergonomics in Design,* 1064804619884160.