Position paper to the NordiCHI 2024 workshop on AI meets PD

Two challenges to consider when designing with AI.

Guri B. Verne, Department of Informatics, University of Oslo

Working with AI as a material for doing participatory design will give some special challenges when designing technology where AI will be (part of) the results. A fundamental property with AI is that it cannot be expected to give deterministic results, understood as the same input will not necessarily give the same results next time. In the public debate, AI is presented in very technooptimistic ways, almost without any consideration of the challenges it involves. This may lead users participating in design processes to be overly optimist about what AI can do.

Results from AI is based on stochastic calculations where unexpected or nondeterministic assessments can be the result. Such outcomes – or properties of the algorithms and the data – is difficult to assess, or evaluate, in advance. Knowledge and experience with working with the results from AI and machine learning will help to do that. Bratteteig and Verne (2018) argue that it will be difficult for both users and designers to evaluate design suggestions in a PD design process when they cannot evaluate the expected future results from the design.

In this position paper, I will discuss two challenges when designing with AI in a PD process. The first challenge is the quality of the recommendations or decisions which will be the results from the AI, and the second is how to work with AI in the design process to aim for quality over time in the results.

Based on Schön´s (1995) conceptualization of design as iterations of see-move-see activities, Bratteteig and Verne (2018) suggest that AI can be used differently in these three "activities" of a design process. These are creating design ideas, selecting and concretizing ideas, and seeing/evaluating a design choice. In the following I will focus on the activity of evaluating a design choice as part of a PD process. Knowledge about challenges with the quality of results from AI will prepare awareness about the risk of weak or even incorrect results when selecting or evaluating a design choice. Then I will discuss how mundane experiences with machine learning and AI in everyday practice can be included in mutual learning, a central element of PD to support users in making more informed design decisions.

- **Evaluating the results from AI**

Results from AI, or to be more precise, an AI-infused system, depends on data and the "training" of the algorithms. Large amounts of training data will be necessary to train and adjust the model. This training can be supervised by a human or unsupervised, but that is not the point here where I will focus on the implications from this dependence on the data on the design work. Designing an AI-infused system will include selecting and preparing the training data that will make the basis for the results from the system. Research has shown that the results from AI can be unexpected or insufficient, based on the data used for training the AI in unexpected ways (Besse et al., 2019; Ribeiro et al., 2016; Zech et al., 2018). Illustrating examples that circulates are the husky that is classified as a wolf when there is snow in the background (Besse et al., 2019; Ribeiro et al., 2016) and the x-ray diagnosing image recognition that recognised the metal token that identified each participating hospital and gave a response based on the general treatment success score of that hospital. Four cooperating hospitals had received very good results from using image recognition to detect

pneumonia in chest radiograph, but when a new hospital joined after the training period was over, their results were weaker. It turned out that the AI image recognition did not recognise the metal token of the new hospital, and since it had not been trained with images from the new hospital, it did not have data for their success rate and without any notice gave random guesses as results (Zech et al., 2018).

Also, rare and unexpected results will occur. Antun et al. (2020) showed that small changes in an image where no tumours are detected, gave the opposite interpretation from image recognition AI when the image was turned 90 degrees. Antun et al. (2020) use the term "instability" to discuss this phenomenon, which pose a challenge for both designers and users in evaluating future results when working with AI. Both designers and users participating in a design process will be challenged when they as part of a PD process evaluate results from AI that may be unstable in unforeseen ways.

The Norwegian tax and welfare chatbots use machine learning to match the user's request to the answer. All answers are correct as they are formulated by experience advisors, but the matching between the request and the response is based on machine learning. The user will need background knowledge in the tax or welfare domain to assess the relevance of the chatbot´s reply for their life situation (Simonsen et al., 2020; Verne et al., 2022). This stands in contrast with the chatbot being presented to the public as a first contact point with the welfare administration.

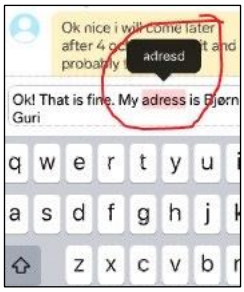
- Mutual learning for developing awareness to unexpected results

The data and algorithms that constitute the basis for the results from AI will change over time, and the results may become more precise - or possibly less precise if there are instabilities in the data (Antun et al. , 2020). Evaluating the quality of the results can take more time than the time allowed for the design process. In addition, applications with AI needs to be configured before use to improve usefulness (Zając et al., 2024),

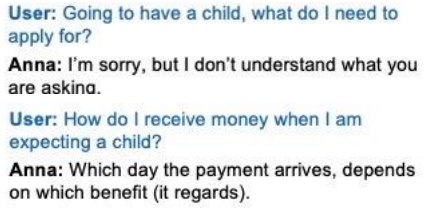
A way for preparing participants in a PD project to be aware of the risk for unexpected results from an AI can be to facilitate gaining experience with results from mundane AI and machine learning from everyday applications. For example, experience with annoying Autocorrect suggestions that give erroneous interpretation of the word that is written can give an indication of how the algorithms of larger scale AI can give weak or even wrong results.

Mutual learning is an important element of a PD process (J. Simonsen & Robertson, 2013). The point is that non-designers will learn enough about the technology involved and design thinking to be able to make informed design choices about this technology. PD is sometimes misunderstood to mean that designers act as voiceless facilitators that design whatever the users want. The element of mutual learning in PD is that the designers also have a voice so that their competence help bring forth better design decisions. In this case, knowledge about how AI works and what to expect about the quality of the results will be important when designing with AI. One approach to building an understanding of the risk of instable results can be to inspire reflection from experiences with mundane examples of AI and machine learning. In the following I give some examples for inspiration and reflection.

Example 1: Autocorrect often gives wrong, irrelevant, or just strange spelling suggestions. Reflecting on how these come about can be elements of mutual learning in a design process with AI. What if a medical image recognition system makes "similar" strange recommendations? In the figure to the right, which is a screenshot from a chat, autocorrect suggests the word "adresd", which is neither a Norwegian nor an English word, for the word "address" which was written.



Example 2: Experiences with irrelevant or wrong responses from chatbots can be food for thought about how the responses are produced. In the example to the right, the welfare chatbot Anna does not give an answer to the user's request, and the chat ended. The example is from Verne et al. (2022). NB: Wrong answers from the chatbot will be difficult to detect without previous knowledge of the domain in question!



Examlpe 3: My car, a VW ID3 1st has a voice input option to ask for heating, making phone calls and other non-driving activities. The car is German made and it almost never recognises my voice. We have stopped using it but sometimes it abruptly and annoyingly interrupts our conversations in the car, asking "Hva sa du?" (What did you say?) in a foreign accent. This car also gives weaker navigation suggestions than what I can get from my phone, I think it is because the VW company use their own map data instead of buying google's map data. Sometimes I compare them and find that google map has registered locations which my car's map does not have.

The research examples given above can also be illustrating and inspire reflections on the quality of results from AI. Below is Figure 2 from Besse et al. (2018) with their explanation.

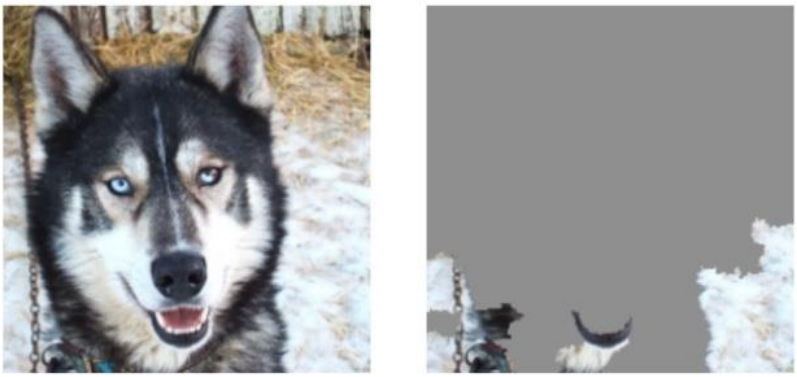


Figure 2. A husky (on the left) is confused with a wolf, because the pixels (on the right) characterizing wolves are those of the snowy background. This artifact is due to a learning base that was insufficiently representative.

- **Summing up and discussion**

In this position paper, I discuss how we can work with AI in a PD process taking into account that AI can give instable results that change over time and that the data involved in training is crucial for the results. In particular, the activity of evaluating a design choice can be challenging. Reflecting on illustrating examples of wrong or unsatisfactory AI results can be

part of mutual learning in PD design processes. Mundane examples from everyday life can give valuable experiences. Critical reflections on personal experiences with AI can be an antidote to unrealistic and overly technooptimistic expectations to its results, and give a different understanding than what is conveyed through industry, which has strong interests in hyping AI.

Designers can grow their own familiarity with insufficient or instable results from AI to strengthen their knowledge base for mutual learning. PD will be important for reflection on technological choice and mutual learning of AI with users, as a basis for more democratic processes for design and development of powerful and extremely expensive technology.

In addition to the challenges of working with results that may be instable and change over time, there are more problematic aspects of designing with AI. Further research on the enormous energy demand for answering relatively small and simple requests, will be valuable when considering designing with AI.

References

Antun, V., Renna, F., Poon, C., Adcock, B., & Hansen, A. C. (2020). On instabilities of deep learning in image reconstruction—Does AI come at a cost? *Proceedings of the National Academy of Sciences*, *117*(48), 30088–30095. https://doi.org/10.1073/pnas.1907377117

Besse, P., Castets-Renard, C., Garivier, A., & Loubes, J.-M. (2019). Can Everyday AI be Ethical? Machine Learning Algorithm Fairness. *Statistiques et Société*, *6*(3). https://doi.org/DOI: 10.13140/RG.2.2.22973.31207

Bratteteig, T., & Verne, G. (2018, August 20). *Does AI make PD obsolete? Exploring challenges from Artificial Intelligence to Participatory Design*. Participatory Design Conference (PDC) 2018.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *'Why Should I Trust You?': Explaining the Predictions of Any Classifier* (arXiv:1602.04938). arXiv. http://arxiv.org/abs/1602.04938

Schön, D. A. (1995). *The reflective practitioner: How professionals think in action*. Aldershot: Arena.

Simonsen, J., & Robertson, T. (2013). *Routledge International Handbook of Participatory Design*. Routledge.

Simonsen, L., Steinstø, T., Verne, G., & Bratteteig, T. (2020). "I'm Disabled and Married to a Foreign Single Mother". Public Service Chatbot's Advice on Citizens' Complex Lives. *Electronic Participation: 12th IFIP WG 8.5 International Conference, ePart 2020, Linköping, Sweden, August 31 – September 2, 2020, Proceedings*, 133–146. https://doi.org/10.1007/978-3-030-58141-1_11

Verne, G., Steinstø, T., Simonsen, L., & Bratteteig, T. (2022). How Can I Help You? A chatbot's answers to citizens' information needs. *Scandinavian Journal of Information Systems*, *34*(2). https://aisel.aisnet.org/sjis/vol34/iss2/7

Zając, H. D., Ribeiro, J. M. N., Ingala, S., Gentile, S., Wanjohi, R., Gitau, S. N., Carlsen, J. F., Nielsen, M. B., & Andersen, T. O. (2024). 'It depends': Configuring AI to Improve Clinical Usefulness Across Contexts. *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, 874–889. https://doi.org/10.1145/3643834.3660707

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, *15*(11), e1002683. https://doi.org/10.1371/journal.pmed.1002683