Statement of Interest: Center for Human-Compatible AI

Duncan C McElfresh | https://duncanmcelfresh.github.io/

There are two reasons I am excited to work with the Center for Human-Compatible AI: First, I believe that aligning AI with human interests through stakeholder feedback is essential for responsible AI design. Second, with CHAI experts I am eager to study how AI influences human decision making, and to build AI systems that reflect—and adapt to—human interests.

My research focuses on AI applications for resource allocation and decision support, including kidney exchange and blood donation. This involves a mix of theory, computational simulations, and human subjects research. Some of my research takes a "top-down" approach to AI design; for example, addressing uncertainty in kidney exchanges using mathematical optimization techniques [2, 4, 6] or matching blood donors and recipients using an online platform [7]. However, a top-down approach is not always appropriate: AI researchers are often far-removed from the application domain, and it is impossible plan for every scenario an AI system might encounter. In many applications, such as organ allocation, it is more appropriate to take a"bottom-up" approach driven by stakeholder input. To this end I use human-centered studies to understand how AI influences decision making—my research has touched on indecision [1], perceptions of fairness [5], and the influence of AI suggestions on consequential decisions [3]. **As a CHAI postdoc I would like to address the questions raised by this research.**

One promising context for this research is natural language processing (NLP). While NLP is already commonplace in advertising, customer relations, and internet search, it remains a challenge to align these methods with complex human values in a variety of contexts. Failure to align NLP models with human values has led to memorable public relations disasters[1] and creates a risk of bias and long-term harm.[2] For this research I am especially interested in working with Andrew Critch and his collaborators, building on their ICLR'21 paper "Aligning AI With Shared Human Values" and the ETHICS dataset.

Another important step is to develop a *framework* of AI-influenced decision making. First, humans are inconsistent and indecisive [1]; AI systems need to account for this human quirkiness. Furthermore, AI systems are often designed to *influence* human behavior; indeed this is the purpose of decision support systems and recommender systems. If AI systems are trained on human behavior, and human behavior is in turn influenced by AI suggestions [3], it is difficult to determine the objective quality of AI-assisted human decisions. How can we build AI systems that are useful while avoiding self-reinforcing human-AI feedback loops?

In my view, an effective theoretical framework should draw inspiration from computational social choice and preference modeling (on the technical side), and moral philosophy and behavioral economics (on the human side); to account for the dynamism and uncertainty of human values and behavior, it is natural to build on reinforcement learning concepts as well. To this end, I am interested in learning from reinforcement learning experts, including Anca Dragan and Pieter Abbeel, to explore these ideas further.

Thank you for your consideration,
Duncan C McElfresh

[1] https://en.wikipedia.org/wiki/Tay_(bot)

[2] Bender et al., On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *FAccT'21*

REFERENCES

1. McElfresh, D. C., Chan, L., Dole, K., Sinott-Armstrong, W., Conitzer, V., Borg, J. S. & Dickerson, J. P. Indecision Modeling. *AAAI* (2021).
2. Bidkhori, H., Dickerson, J. P., McElfresh, D. C. & Ren, K. Kidney exchange with inhomogeneous edge existence uncertainty. *UAI* (2020).
3. Chan, L., Doyle, K., McElfresh, D. C., Conitzer, V., Dickerson, J. P., Schaich Borg, J. & Sinnott-Armstrong, W. Artificial Artificial Intelligence: Measuring Influence of AI'Assessments' on Moral Decision-Making. *AIES* (2020).
4. McElfresh, D. C., Curry, M., Sandholm, T. & Dickerson, J. P. Improving Policy-Constrained Kidney Exchange via Pre-Screening. *NeurIPS* (2020).
5. Saha, D., Schumann, C., McElfresh, D. C., Dickerson, J. P., Mazurek, M. L. & Tschantz, M. C. Measuring Non-Expert Comprehension of Machine Learning Fairness Metrics. *ICML* (2020).
6. McElfresh, D. C., Bidkhori, H. & Dickerson, J. P. Scalable Robust Kidney Exchange. *AAAI* (2019).
7. McElfresh, D. C., Kroer, C., Pupyrev, S., Sodomka, E., Sankararaman, K. A., Chauvin, Z., Dexter, N. & Dickerson, J. P. Matching Algorithms for Blood Donation. *Under review at PNAS. An earlier version appeared at EC 2020.*