Human-AI Collaboration with Misaligned Preferences

Jiaxin Song [*1], Parnian Shahkar [†2], Kate Donahue [‡1,3], and Bhaskar Ray Chaudhury [§1]

[1]University of Illinois, Urbana-Champaign
[2]University of California, Irvine
[3]Massachusetts Institute of Technology

Abstract

In many real-life settings, algorithms play the role of assistants, while humans ultimately make the final decision. Often, algorithms specifically act as curators, narrowing down a wide range of options into a smaller subset that the human picks between: consider content recommendation or chatbot responses to questions with multiple valid answers. This type of role is one of the most common in human-algorithm systems: e.g., in content recommendation, search, and some types of categorical prediction. For example, when a user requests directions, Google maps returns a small set of routes, and the user typically picks her final route from those routes presented. Crucially, humans may not know their own preferences perfectly either, but instead may only have access to a *noisy* sampling over preferences. Algorithms can assist humans by curating a smaller subset of items, but must also face the challenge of *misalignment*: humans may have different preferences from each other (and from the algorithm), and the algorithm may not know the exact preferences of the human they are facing at any point in time.

In this paper, we model and theoretically study such a setting. There are two actors: an algorithm and a human in the human-algorithm collaboration model. A set of items $M = \{x_1, \ldots, x_m\}$ representing different outcomes (e.g., labels in prediction tasks, generative model output). The specific mode of interaction we assume is where the algorithm picks their top k from some noisy ranking over items, and the human picks their favorite among that set, according to their own noisy ranking over items.

Our first finding is that there are settings where misalignment can be *helpful*. We consider the setting where the human may work with multiple algorithms with different ground-truth. We show instances where the human benefit by collaborating with a misaligned algorithm. Surprisingly, we show that humans gain more utility from a misaligned algorithm (which makes different mistakes) than from an aligned algorithm. The key insight is that appropriate misalignment on low-valued items can increase the likelihood that humans select high-valued items. Next, we build the algorithm by studying two welfare objectives: *utilitarian welfare* and *uplift*. Social welfare measures the expected utility of the human, while uplift requires each type of human to benefit from the collaboration. While these two objectives do not always align with each other (i.e., there exist instances where the utilitarian welfare maximizing algorithm does not achieve uplift and vice versa), we design different approaches for achieving these two objectives. For utilitarian welfare maximization, we show that it is NP-hard and design an MIP that solves it effectively in practice. Meanwhile, although an uplift algorithm may not always exist in general, we provide some special scenarios where a naturally designed algorithm can always achieve uplift. We conclude by discussing implications for designers of algorithmic tools and policymakers.

Acknowledgements

The work is supported by an NSF CAREER grant CCF No. 2441580 and an MIT METEOR fellowship.

[*]Email: jiaxins8@illinois.edu
[†]Email: shakarp@uci.edu
[‡]Email: kpd@illinois.edu
[§]Email: braycha@illinois.edu