Thesis Summary: Tackling Algorithmic Bias using Human-In-The-Loop AI

There have been rising concerns with regards to fairness, accountability and transparency in automated decision-making algorithms. We propose a human centered AI approach where humans engage with the algorithm to detect and mitigate the bias. Unlike algorithms, humans are capable of distinguishing between desirable and undesirable biases even for unseen situations. Humans can incorporate this knowledge into the system by interacting with a visual interface backed by interpretable models. The interactive visual interface boosts transparency by visualizing the underlying state of the system, the debiasing process and different metrics of bias, distortion, etc. This leads to increased trust into the system because humans are part of it and can guide the process when they feel it is necessary (ref.). With human(s) supervising the entire process, this approach also ensures accountability.
In the following, I will describe two of my current projects following this general premise:

D-BIAS: A Human-in-the-loop Methodology for Assessment and Mitigation of Social biases in Algorithmic Decision Making (ADM)

ADM is becoming omnipresent in a wide spectrum of applications, such as hiring, admissions, social care, law enforcement, and others. Initially conceived as a mechanism to eliminate human bias from the decision making process, there is an increasing recognition that ADM is also not without bias, mostly due to biased data. Bias in the data relates to societal constructs, and algorithmic techniques cannot be expected to understand these complicated relationships.
In this research, we propose a human-in-the-loop approach that leverages human understanding to manipulate data and debias the dataset. The visual tool we propose, called D-BIAS, learns a causal network from the underlying data to model the flow of bias. Each node represents a data attribute and each edge represents a causal relationship. The decision maker (DM) infuses domain knowledge into the system by identifying sensitive attribute(s) and interacting with the causal network. The DM can remove inappropriate edges or add edges based on his/her domain knowledge and institutional goals. On debasing, all edges connected with sensitive attribute(s) are removed and the underlying data is modified to reflect changes in the causal network. D-BIAS also recommends addition/deletion of edges to the DM based on semantic relations between data attributes. Various interactive visualizations and charts are available to show how different debiasing techniques affect bias, accuracy, and different metrics of fairness. We tested our approach on both synthetic and real-world datasets such as the German credit dataset, adult income dataset, etc. and the results were very promising.

WEBVis: A Human-in-the-loop Auditing Tool for Exploration and Mitigation of Social Biases encoded in Word Embeddings

Social biases such as race, gender, etc. encoded in word embeddings plague many applications of NLP (e.g. sentiment analysis, machine translation) and cause representation harm. Representation harm occur when systems reinforce the subordination of some groups along the lines of identity. Sustained exposure to representation harm can lead to the alienation of certain factions of society and even social unrest. To address these problems, we have devised what we call WEBVis, an interactive visual tool for exploring different kinds of biases like race, age, etc. across different word embedding models followed by real-time mitigation. Given a word embedding, WEBVis returns its debiased version which can be safely used for any downstream application. The novel interactive visualization design expedites the bias identification process and makes the debiasing process transparent and interpretable. WEBVis employs a novel post-modeling debiasing method in which a word's representation in higher dimensional space is modified such that bias is mitigated while preserving its semantic meaning as far as possible. We have successfully tested WEBVis for popular word embedding models like word2vec, Glove, FastText and multiple languages like English, French and Hindi.

There are many additional areas I wish to explore using this human-in-the loop AI approach, such as take on social biases in recommender systems.