Research Statement

FOROUGH POURSABZI-SANGDEH

Overview

Machine learning (ML) systems increasingly affect every facet of our lives including recommendations on what movies to watch, what music to listen to, what news to see, helping doctors make medical diagnoses, and helping judges set bail amounts. While these systems have been adopted in many scenarios to automatize decision making, they still require a large amount of human involvement and effort; from data used to train models to decisions made with the help of ML systems, humans are at the heart of machine learning.

Because of the central role of humans in machine learning, it is important to study humans and models jointly. The key challenges arise from the need for knowledge in machine learning as well as Human–Computer Interaction (HCI), behavioral psychology, and social sciences. My research lies at the intersection of these fields to empirically study how humans interact with ML systems. I use my knowledge and experience in designing, training, debugging, and evaluating ML models and employ principles and methodologies from HCI and psychology in controlled human-subject experiments to study how people behave when they interact with ML systems. My long-term research goal is to leverage these insights and create systems that foster an effective and responsible collaboration between humans and ML systems.

Below, I describe my past contributions in studying the interaction between humans and machine learning systems. I then elaborate on how these experiments, findings, and insights can be leveraged to further my research agenda in future work.

Understanding and Exploring Datasets

Human-centric evaluation of exploratory tools

In many scientific fields, especially the social sciences, ML models are nearly exclusively used for understanding, explaining, or exploring data. Take, for example, topic models that are commonly used to generate thematic structure from large corpora in the form of *topics*—which are a set of semantically coherent words—and help people understand the content. But how does one assess the effectiveness of these models in helping people understand or explore large corpora? In two lines of work, we take a human-centric approach to understand how people interpret individual topics and use them to complete a real-world exploratory task—namely, understanding of and answering questions about science policy.

We designed a human-subject experiment to systematically study the effect of topic visualizations on human comprehension of a topic's concept [1]. We ran a crowdsourced experiment on Amazon's Mechanical Turk in two phases. In the first phase, participants were shown a visualization of a topic and provided a textual label to describe the topic. In the second phase, a new set of participants assessed the quality of these labels alongside automatically-generated labels. Better labels would have implied that people had a more accurate understanding of the topic.

While we found no meaningful difference in the quality of labels generated by participants, those who saw more complex visualizations (i.e., a network graph) spent more time to understand (and label) topics compared to those who saw simpler visualizations (i.e., a word list or a word cloud). As expected, labels which were generated by humans were superior to automatically-generate labels. In a nutshell, people were better than an algorithm in understanding and summarizing topics but the specific type of visualization they saw did not make any meaningful difference in their comprehension.

In another line of work, we took a task-based evaluation approach to assess the effectiveness of topic models in helping people explore and make sense of a dataset (see Chapter 5 of [2]). We focused on understanding science policy as a use case since it requires understanding the large dataset of research reports and affects important decisions such as the amount of funding that gets allocated to a specific field or the types of research that get transferred into technological innovations.

We hypothesized that providing an overview of the dataset through topics would help people understand the collection of research documents better and faster. In an online experiment (on Upwork), we asked participants a set of questions related to the funded proposals by the National Science Foundation.[1] Participants were given a system with or without topic models to explore the proposals and submit answers to the questions. We compared the quality of their answers. A higher quality answer would have implied that people had a more accurate understanding of the dataset.

Surprisingly, we found no meaningful difference in the quality of answers provided by participants. However, participants who were given a topic model rated the task less mentally demanding and the interface more helpful. We also found evidence of people being quicker and topics being inspiring for further exploration of the dataset.

These findings suggest that scientists should harbor a healthy skepticism of their intuitions and rely on empirical evaluations with humans in context to verify the effectiveness of their methods.

Labeling Data and Training Models

Design and human-centric evaluation of tools that help people collect data and train models

Supervised machine learning models make predictions based on a training set which includes a set of data points assigned with metadata—often in the form of a *label*. Creating the training set requires a lot of manual human effort: annotators need to wade through an often large dataset, select some data points from the dataset, get an understanding of these data points, and apply appropriate labels to the data points. On top of these challenges, the scenarios where a pre-defined label set is not at hand are even more complicated. Take, for example, the case of classifying news articles based on their topic (e.g., sports, art, technology) or academic computer science papers based on their area (e.g., computer systems, machine learning, HCI). In these cases, human annotators need to induce a label set first and only then they can apply labels to data points.

We designed a human-subject experiment to understand how annotators induce and apply labels to a large document collection when they are given various ML tools [3]. We hypothesized that a lack of knowledge on the dataset *overview* makes inducing a label set challenging and therefore, providing an overview of the documents using *topic models* will be beneficial. We also hypothesized that *selecting* which documents to label is time consuming and therefore, suggesting documents to focus on using *active learning* methods will be helpful. Through an online, one-hour, randomized experiment on Upwork followed by a large-scale crowdsourced experiment on Figure Eight, we found that topic overviews and active learning selections led to a higher quality label set and training data, which in turn, was used to train a more accurate model. Furthermore, we found evidence that under strict time constraints, overviews in the form of topics could potentially be overwhelming and therefore not as helpful.

These findings provide insights on how people craft label sets and apply them to data. They also can serve as guidelines for designing systems that help annotators label data under different time constraints. Our system is now publicly available and has been adopted by Snagajob (`https://www.snagajob.com/`).[2]

[1] The questions were inspired by Questions for the Record (`https://www.congress.gov/congressional-record`).

[2] `https://github.com/Snagajob/alto-boot`

Making Decisions

Human-centric study of decision aids in the context of interpretability approaches

ML systems are often used as decision aids. That is, predictions help in making decisions where humans used to decide alone. To convince decision makers that a system is trustworthy, reliable, justifiable, ethical, or fair in practice, many have argued for the need for research on *interpretable* models and methods.[3] Despite the progress in this area, there is still no consensus on how to define or measure interpretability. I take the perspective that interpretability cannot be directly manipulated or measured. Rather, interpretability is a latent property that is influenced by several manipulable factors (such as the number of features, the transparency of the model, and the user interface) and these impact several measurable outcomes (such as people's tendency to follow the model's predictions and people's ability to simulate or debug the model's predictions). Different users in different scenarios may have different needs. As such, interpretability is a sociotechnical concept and should be defined, measured, and evaluated in context with relevant people.

We designed a human-subject experiment to study how two factors—the number of features and whether a model is transparent or black box—affect people's abilities to simulate the model's predictions, follow the model's predictions when it is beneficial to do so and deviate from the model's predictions otherwise [4]. We found that a transparent model with a small number of features was easiest for people to simulate. However, contrary to expectation, simple, transparent models did not improve the degree to which people followed the model's predictions when it was beneficial to do so. Even more surprisingly, we found that transparency hampered people's abilities to detect model's sizable mistakes, seemingly due to information overload. This was in line with our findings in previous experiments that too much information (in the form of topic overviews, for example) can be overwhelming under strict time constraints.

These findings shed light on how the presentation of models can affect people's decisions, provided potential guidelines for designing ML systems as decision aids, and most importantly, underscored the necessity of evaluating interpretability methodologies with humans rather than relying on intuition to ensure that they achieve their intended effects.

Future Work

The central role of people in designing, implementing, evaluating, and deploying ML systems comes with great opportunities for studying human behavior and designing for a more effective collaboration between humans and models. Below I describe several themes of promising future directions. These directions are inherently interdisciplinary and the challenges come from the need for an in-depth knowledge in ML, HCI, psychology, and social sciences. I am excited about continuing my existing collaborations and building new ones across different fields to do new interdisciplinary research and further my research agenda.

Studying different people in different scenarios

When it comes to human-subject experiments, I find it useful to think through multiple dimensions and reflect on them in the process of designing, analyzing, and reporting. No experiment is perfect. Critical thinking about these aspects helps in identifying limitations of the experiment and has the added benefit of providing natural directions for future work.

[3]The terms "interpretable", "intelligible", and "transparent" are often used interchangeably. In this document, I have selected to use "interpretable" and "interpretability" because I think they better capture a wider variety of efforts in this research area.

First, **participants** and their level of expertise play an important role. Relatedly, qualitative experiments with relevant stakeholders in context are necessary for getting a better understanding of the nuances of how people interact with and use ML systems. Second, the particular task and **domain** of the experiment affects results. While empirical findings from an experiment in a specific domain are undoubtedly valuable, expanding out to a variety of domains can provide insights on generalizability. Third, it is important to think about the experiment **setting**, how well it mimics a real-world scenario, where and how it fails to do so, and what this means for the conclusions drawn from the experiment. Fourth, the details of a **user interface** or a certain **visualization** can have a huge effect on human behavior. This leads to various research opportunities related to the user experience when interacting with ML systems to achieve specific goals.

Fairness and ethics

Issues around fairness and ethics have played an important role in motivating the progress in interpretable machine learning. There are several reasons why interpretability is desirable from this perspective. First, interpretability can provide evidence of compliance with the General Data Protection Regulation's requirements to provide "meaningful information about the logic involved" in models. Second, ML systems that are easier to understand can potentially be easier to audit, ensuring that they are working on ethical grounds. Third, it is often assumed that humans can help identify and mitigate potentially existing unfairness if they are given a chance to oversee these systems and interpretability can potentially facilitate this.

Each of these motivational points calls for further research and experimentation to make real impact. For example, getting a better understanding of regulations is necessary to assess the effectiveness of a model or an explanation method in providing the required information to people. A thorough understanding can also motivate designing new methods that meet regulatory requirements. Furthermore, it is crucial to study how various tools for interpretability affect various notions of (un)fairness. Given that falsifying common intuition has been an emerging theme in the experiments I have run, one should empirically study whether people can identify issues around unfairness through their interaction with systems.

As a first step in this direction (in an ongoing project), we study the role of humans in the context of Facial Recognition (FR) [5]. There has been growing concern around the deployment of FR systems in high-stakes scenarios (e.g., authorizing to enter a building or detecting people who belong in a "watch list"). As a result, many have argued for involving humans to oversee the decisions made by FR systems with the hope of more fair decisions. We seek to study how humans interact and make decisions with FR systems, and how these decisions impact different populations. These experiments will have implications for designing, auditing, deploying, and regulating FR systems. Similar human-subject experiments with other systems in other scenarios and domains are important future directions.

More human engagement

More and more often researchers and practitioners claim that complex models that are hard to understand hamper trust and adoption. To address this issue, they turn to simple models or explanations for complex models. I take the perspective that simplicity or explainability is not sufficient to ensure that a model or a system is trustworthy. We should identify people's needs in using these systems and design for an ongoing and efficient engagement and interaction. This perspective is in line with what we discuss in the context of text analysis [6, 7] as well as health-wearable technology [8] and what Kluttz et al. [9] phrase as "contestability".

Interpretability tools provide a way to peak through models, but often are a "take it or leave it" proposition; there is not a trivial mechanism to improve or personalize a model. Research on how to interact with models or explanations, incorporate feedback in them, and improve them is promising for promoting

reliability and adoption. This is a large interdisciplinary research agenda. It requires research from the ML perspective to design mechanisms that allow incorporating human feedback in model training. Furthermore, methodologies from HCI and psychology are necessary to empirically assess the effectiveness of such mechanisms. And finally, interdisciplinary research at the intersection would let us design, build, and empirically evaluate systems that facilitate an effective and responsible collaboration between humans and models.

References

1. Alison Smith, Tak Yeon Lee, Forough Poursabzi-Sangdeh, Leah Findlater, Jordan Boyd-Graber, and Niklas Elmqvist. Evaluating visual representations for topic understanding and their effects on manually generated labels. *Transactions of the Association for Computational Linguistics*, 2016.
2. Forough Poursabzi-Sangdeh. Design and empirical evaluation of interactive and interpretable machine learning. 2018.
3. Forough Poursabzi-Sangdeh, Jordan Boyd-Graber, Leah Findlater, and Kevin Seppi. ALTO: Active learning with topic overviews for speeding label induction and document labeling. In *Association for Computational Linguistics*, 2016.
4. Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810*, 2019.
5. Forough Poursabzi-Sangdeh, Samira Samadi, Jennifer Wortman Vaughan, and Hanna Wallach. A human in the loop is not enough: The need for human-subject experiments in facial recognition. In *Human-Centered Approach to Fair and Responsible AI Workshop at CHI*, 2020.
6. Jason Chuang, John D Wilkerson, Rebecca Weiss, Dustin Tingley, Brandon M Stewart, Margaret E Roberts, Forough Poursabzi-Sangdeh, Justin Grimmer, Leah Findlater, Jordan Boyd-Graber, et al. Computer-assisted content analysis: Topic models for exploring multiple subjective interpretations. In *Human-Propelled Machine Learning at Advances in Neural Information Processing Systems*, 2014.
7. Alison Smith, Tak Yeon Lee, Forough Poursabzi-Sangdeh, Jordan Boyd-Graber, Niklas Elmqvist, Kevin Seppi, and Leah Findlater. Human-centered and interactive: Expanding the impact of topic models. In *Human Centered Machine Learning Workshop at CHI*, 2016.
8. Bran Knowles, Alison Smith, Forough Poursabzi-Sangdeh, Di Lu, and Halimat Alabi. Attending to the problem of uncertainty in current and future health wearables. *Communications of the ACM*, 61(12):62–67, 2018.
9. Daniel Kluttz, Nitin Kohli, and Deirdre K Mulligan. Contestability and professionals: From explanations to engagement with algorithmic systems. *Available at SSRN 3311894*, 2018.