Pacific Symposium on Biocomputing 25:731-735(2020)

AI Ethics and Values in Biomedicine – Technical Challenges and Solutions

Dragutin Petkovic
Computer Science, San Francisco State University, 1600 Holloway Ave
San Francisco CA 94132
Petkovic@sfsu.edu

Lester Kobzik
Environmental Health, T.H.Chan Harvard School of Public Health
Boston, MA 02115
lkobzik@hsph.harvard.edu

Reza Ghanadan
Google AI, 1600 Amphitheater Pkw
Mountain View, CA 94043
rezaghanadan@google.com

There is increasing recognition of the need to ensure better consideration of ethics and values in artificial intelligence (AI) applications for biomedicine. This 3-hour workshop will provide an overview and discussion on the technical challenges and potential solutions that can enable ethical and value-based AI algorithms and tools in biomedicine. Five expert speakers will present their work and engage the audience in discussion with the aim of defining the best ways to move forward.

Keywords: Artificial Intelligence; AI ethics; AI values, AI bias, explainability

1. Introduction

We are witnessing the emergence of an “AI economy and society”. AI technologies are increasingly impacting many aspects of modern life, including biomedicine and healthcare. However, AI systems may produce errors, can exhibit overt or subtle bias, may be sensitive to noise in the data, and often lack transparency and explainability. These shortcomings raise many ethical and policy concerns that impede wider adoption of this potentially very beneficial technology. These broad concerns about AI are often grouped under the rubric “AI Ethics and Values” and these issues are especially important in the biomedical area. The technical community, the media, as well as political and legal stakeholders have recognized the problem and have begun to seek solutions. Recent examples of such efforts include regulatory actions such as the EU GDPR privacy and data protection laws (May 2018), and the recent California Assembly endorsement of the 23 Asilomar AI Principles (see below).

In the highly regarded Asilomar AI Principles established by the Future of Life Institute (https://futureoflife.org/ai-principles/), the section on AI Ethics and Values includes a number of powerful, well-reasoned and noble principles which guide our focus: Safety; Failure Transparency; Judicial Transparency; Responsibility; Value Alignment; Human Values; Personal Privacy; Liberty and Privacy; Shared Benefit; Shared Prosperity; Human Control; Non-subversion. However, developing technical solutions and practices to help implement and verify these AI Ethics and Values principles has proven to be a substantial challenge. We believe that it is imperative for the scientific and research community to focus on this problem. The main goal of our workshop is to address technical obstacles and approaches in addressing those issues

2. Workshop Goals and Organization

The proposed workshop will focus on *scientific and technical issues* that are central to ensuring better AI Ethics and Values in health and biomedical fields. The workshop will aim to address, among others, the following questions *all in the context of biomedicine*:

- What are the key measurable and auditable components and issues comprising AI ethics and values?
- What state of the art algorithms and solutions are available today that address AI ethics and value issues in a principled and measurable way? Do we need to develop radically new AI algorithms or simply enhance existing ones?
- Can we develop compliance algorithms, tools and processes to verify for adherence AI ethics and values principles
- How can we make AI solutions easier to understand and explain (both at model and sample level), and easier to adopt by experts and non-experts alike?
- How to better leverage currently underutilized "human-in-the-loop" and "user centric methods" in order to design easier to use and control AI systems for experts and non-expects alike
- What specific practices can be recommended to ensure design, development, evaluation, deployment and verification of ethical and value based AI systems?

After an introduction by the workshop Chair, speakers will share their expertise and views on the current state of the art and their vision for answers to the questions posed above. Speakers will then form a panel and engage the audience in discussion, moderated by workshop organizers. At the end, the workshop Chair will summarize the main points of the presentations and subsequent discussion.

3. Panelists' abstracts

The workshop includes five accomplished speakers/panelists whose research addresses the goals of the workshop and whose contributions and expertise as a group cover the technical questions related to the workshop goals. They are: Prof. Su-In Lee (Allen School of Computer Science, U. of Washington); Dr. Claudia Perlich, (Senior Data Scientist, Two Sigma; Stern NYU); Prof. Chris Re, (Computer Science, Stanford University); Prof. Sameer Singh (Information and Computer Science, UC Irvine); Prof. Jessica Tenenbaum (Biostatistics and Bioinformatics, Duke University). Below we list panelists' abstracts reflecting their initial thoughts and ideas to be discussed at the workshop.

Opening the Black Box of Machine Learning Models

Prof. Su-In Lee, Allen School of Computer Science & Engineering, University of Washington

Modern machine learning (ML) models can accurately predict patient progress, an individual's phenotype, or molecular events such as transcription factor binding. However, they do not explain why selected features make sense or why a particular prediction was made. For example, a model may predict that a patient will get chronic kidney disease, which can lead to kidney failure. The lack of explanations about which features drove the prediction – e.g., high systolic blood pressure, high BMI, or others – hinders medical professionals in making diagnoses and decisions on appropriate clinical actions. The black box nature of AI and ML techniques is one of the primary causes of various ethical problems of AI-based technology. The talk will describe current efforts to develop interpretable ML techniques including the SHAP (SHapley Additive exPlanations) for varied biological and medical applications and then present recent efforts on how to capture when a ML model fails for safe deployment of an ML system, and how to move from explanations to actions.

Social Biases: Propagation and Creation through Predictive Models

Dr. Claudia Perlich, Senior Data Scientist, Two Sigma; Stern NYU

Our world is increasing shaped by the predictions of models that learn from data. Whether it is the ads that we see, who to become friends with on Facebook, your chances of repaying a loan or the likelihood of developing cancer – more and more of our environment is shaped by predictive models. While often beneficial and de facto better at informing our decisions than say human experts, there is an increasing concern that while being better at making predictions, they may be equally limited when it comes to discrimination simply because the data the models were build on data that itself reflected our all too human biases. If there was no (or very few) female data scientist who got hired/invited for an interview in the database, the recommender system would not ever recommend a female for that position. So much has been argued that the training data needs to be assessed for biases and if needed somehow 'be-biased' to ensure that models are truly fair and reflect our human values (rather than just the statistics of the data). While this is a important concern, this work looks at a much less easily diagnosed second order effect: even if the first order of the data is unbiased (say 50% of the invited applicants for data science jobs in the data are in fact female), we can show that predictive models can easily create biases in their prediction and as a result, the candidates predicted to be most likely to be invited can deviate strongly from the desired 50% and skew to one or the other subpopulation. This effect is not related to the choice of algorithm but primarily to the relative signal to noise ration in one sub-population over the other. We finally discuss the limitations of trying to detect such model-induced biases.

Stories from the Sewer: A Call to Improve How We Shape Training Sets
Prof. Chris Re, Computer Science, Stanford University

A training set is the primary means that one describes the goal of a supervised machine learning system. In the last few years, we have built tools in the Snorkel project (http://snorkel.org) to help improve the training set construction process with both statistical theory and accompanying software. The methods first developed in Snorkel are part of widely used products from Apple, Google, and others and in scientific and clinical efforts. This talk will describe subtle--and not so subtle—ways that the training set creation process can lead to a variety of undesirable outcomes including over optimism in model performance. Initial technical progress will be presented, but the hope is that these case studies will spur discussion toward more responsible construction of training data in medical sciences and machine learning more broadly.

Detecting Bugs in Machine Learning Models Using Data Perturbations
Prof. Sameer Singh, Information and Computer Science, UC Irvine

Machine learning is at the forefront of many recent advances in science and technology, enabled in part by the sophisticated models and algorithms that have been recently introduced. However, as a consequence of this complexity, machine learning essentially acts as a black-box as far as users are concerned, making it incredibly difficult to detect bugs in their behavior. For example, determining when a machine learning model is “good enough” is challenging since held-out accuracy metrics significantly overestimate real-world performance. In this talk, I will describe automated techniques to detect bugs that can occur when a model is deployed. We will use semantic and natural perturbations of the individual instances to probe the model behavior and describe the detected bugs using high-level summaries over the whole dataset. The talk will include applications of these ideas on a number of NLP tasks, such as reading comprehension, visual QA, and knowledge graph completion.

I'm not biased, I'm a computer. Trust me (at your own peril)

Prof. Jessica Tenenbaum, Biostatistics and Bioinformatics, Duke University

In a recent "The Future of Everything" podcast by Russ Altman, Dr. Altman's guest had studied police bias in traffic stops by gathering 200 million traffic stop records across all 50 states. In rare cases, race data was not available. One explanation provided was "we don't record the races of the people that we stop because we're not racist." This serves to illustrate that people with the best of intentions may not always do what's best. Likewise, researchers attempting to harness the power of artificial intelligence often have the best of intentions. By using an impartial computer and not a fallible and bias-prone human being, they may hope to eliminate bias and instead to make predictions or classifications based on "just the facts." Unfortunately, it is increasingly clear that artificial intelligence is not without its own "biases." For humans the reasons are complex and multi-faceted: cultural influence, irrationality, emotions, and faulty memory, among other reasons. One of the most important, and insidious, causes of bias in AI is that of skewed training data.

The "learning health system" paradigm, in which data collected through the course of clinical care may be used for research, has generated tremendous opportunity to leverage large clinical datasets from large medical centers. In theory these datasets are representative of the population of the medical center's region. However it is important to remember that patients do not visit these medical centers at random. For example, in profiling the racial distribution of patients at Duke Medical Center with a diagnosis of schizophrenia, the proportion of African Americans, and black men in particular, was significantly higher than one would expect based only on prevalence in the population. The enrichment may have been related to insurance status, socioeconomic status, geographical convenience, social support networks, or more likely some combination of these and other factors. It would be all too easy for someone with basic coding skills to apply machine learning packages for any number of predictive analytics. It is critically important in cases like this that the AI algorithm be explainable and transparent in order to ascertain what features it is using, and why. It is also critical that students of AI be given training on the limitations of the technology, and how to account for bias in the data. This talk will feature some examples of AI gone wrong, and why it is crucial to be able to see behind the curtains and understand why an AI algorithm behaves as it does, before we turn over key areas of clinical decision making to "intelligent" systems that are only as "perfect" as their creators.

(1)