Daniel Rusk

INFO 4940

Interpretable Human AI Risk Assessment System

VISION STATEMENT:

I am proposing an Human AI system that transparently communicates to both the judge and the defendant when predicting recidivism rates of a defendant. The system will be used in pre-trial and sentencing decisions to assess the risk the defendant has of reoffending. The direct users of this system will be judges and defendants. The data the system will learn from is thousands of surveys of people in jail recording over 130 variables regarding the person and whether or not they recommited a crime in the past two years. The system will return an estimated recidivism risk of the given defendant. This estimate will be given to the judges and defendant(s) involved in the case. The system will communicate effectively with the judge and defendant how it came to its conclusion. The system will use different techniques of communication for both the judge and defendant because they have different contextual backgrounds and in this way, comprehension will be optimized. The broader humans impacted by the system are the general public. The general public will be impacted because they will be able to understand the way the AI system is predicting recidivism rates. Also, the way in which the general public accepts the use of the system could influence the development of the system. For instance, if the general public does not believe that the system is transparent enough, that could inspire an initiative to further the transparency of the communication the system makes.

My proposed system will be transparent and the judges and defendants will all be able to view the inner workings and see what the system does with inputted data from a case. In this way, I believe my Human AI recidivism rate predicting system will be able to communicate transparently. Later in the paper I will clarify the details of the proposed system.

RELATED WORK:

There is an existing system that predicts recidivism rates called COMPAS and there are many papers describing the many issues it has, including lack of transparency. **Walmsey 2020** explains how one of the large issues with COMPAS currently is that since it is privately owned, its deeper workings and algorithm is not publicly disclosed. This means that defendants do not know what considerations the system is making when making a decision. This is a problem because this is not helpful to the judge that may want more insight on how the system made its calculations and this is not helpful to the defendant who desires solid reasoning regarding his sentencing. However, before we get into how to make a system like this transparent we must first understand what the system is and the data on which it feeds..

Rudin, Wang, Coker 2019 explains how COMPAS generates two scores, general and violent, both representing the risk of the defendant re-committing a general or violent crime respectively. These raw scores are out of 10, with 10 being the highest risk. To generate these scores the algorithm collects 137 variables from a questionnaire the defendant answers and from those computes a variety of subscales using a method that is not disclosed to the public. Once the algorithm has the subscales it then linearly combines them all for a final, raw risk score.

According to the **Practitioner's Guide to COMPAS Core,** the data that was used to create the weights of each feature that is inputted into the COMPAS system came from a study that the COMPAS owners, Northpointe, conducted. They sampled over 30,000 of the 137 question surveys conducted at prisons, parole, jail, and probation sites across America and used the results to assign weights to each feature depending on how strongly associated a given factor was with recidivism rate.

According to **Practitioner's Guide to COMPAS Core**, the system was created in 1998 and using various unnamed recidivism studies, the algorithm was adjusted over time. This is a problem. Northpointe does not disclose what kind of data is being collected or being analyzed to then adjust the algorithm. This data or study findings could be biased or poorly collected but the public would have no clue.

Now that we understand slightly more about COMPAS we can look into how it is interpreted and what interpretation skills do judges typically have. The ACM paper **Disparate Interactions: An Algorithm-in-the-Loop,** explains how there are two types of errors judges can make when they interpret automated results. The judge can either fail to see when the system makes an error or fail to incorporate outside information that the system does not consider. It is important to realize that even a very accurate system could lead to incorrect sentencings. However, a more transparent system could perhaps decrease the chances of these two types of errors. A transparent system could show the judge exactly what features it is looking at and how those features are ultimately affecting the risk assessment. This could then decrease the chance of the judge making one of the two types of errors mentioned above.