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.

Unfortunately, I should note that I did not find any applicable literature on defendants interpreting risk assessments, only on judges interpreting them.

Now that we recognize the importance of transparency, we must look into the multiple ways transparency could be implemented into a human AI system.

The paper, **Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?**, explores how best to employ transparency into these AI decision making systems. One potential option is to disclose the weights the algorithm assigns to certain factors it takes in as input. This is called intentional stance explanations. It essentially involves creating a model of a model. **Chopra and White 2011** looks at using trace programs to open up the "black box" of the algorithm without exposing the inner workings that may be proprietary of the algorithm. The trace programs function to provide patterns of very complicated phenomena that are easier to describe at the design level rather than the technical level. While this is one option to the transparency problem, it still raises the question of the most user friendly way to convey the explanation.

Binns et al. 2018 looked into this issue and conducted a study on the best explanation style for interpretable machine learning models. They found that when testing four different explanation styles with their participants, the case-based explanation in justice related decision making by AI systems, led to significantly less understanding than the other three methods: input-influenced based, demographic-based, and sensitivity based. This begs the question: which explanation style or which combination of explanation style is best suited for decision making AI systems? A good point raised in **Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?**, is that the input-influenced style of explanation is perhaps best suited for judicial decision making systems such as recidivism rate predictors because this style more closely resembles typical judicial reasoning and remarks regarding a sentence.

Another proposed solution to the transparency problem with decision making algorithms comes from **Chiao 2019.** This paper does not take on what style of explanation should accompany an AI decision making system, but rather just suggests that an explanation should be given at every layer of the network, yet does not go into further detail. These intermediate explanations could potentially aid the comprehension of the system by a defendant or judge. While the paper proposed this idea, it did not conduct a study using this intermediate explanation idea.

Rudin 2019 looks at transparency in AI decision making systems used in court, but without using explanations and instead creating an inherently transparent system. It explains how not much work has been done in the interpretable ML space because of this unfounded fear of a trade off between accuracy and interpretability. It also argues that when black box decision making models such as COMPAS are explained the explanations tend to be overly complicated and borderline pointless. Rather than setting out a model and testing its interpretability performance, the paper lists out examples of hand-made interpretable logical models with which the AI system could be designed after.

I could not find research or studies evaluating the transparency of interpretable recidivism rate predicting AI systems, so I believe that is where my study could help out and move the status quo further towards my vision. I feel the current issue with recidivism rate predicting systems is that they offer no interpretability. To solve this, I want to use the explanation styles used in the **Binns et al.** study (besides case-based because it was found to be least effective at explaining) and apply them to explaining a recidivism rate predicting human AI system. My study goal is then to answer the following question: When using a recidivism rate predicting AI system, which explanation style out of input influenced, sensitivity based, and demographic based, would best explain the system's recidivism risk assessment for judges and which explanation style would best explain the system's recidivism risk assessment for defendants?

METHOD:

System:

I am proposing an AI system that takes in data regarding a defendant and creates a recidivism risk assessment as well as an explanation as to how the system reached that assessment. The risk assessment will boil down to a number out of 10, with 10 being the most risky to recommit a crime. This system however does not exist yet, so my study will use Wizard of Oz technique to simulate the risk assessments and explanations the system would generate if the system existed.

Why WoZ?

The reason I am using WoZ is because this way I can focus less on developing an accurate and unbiased system, and focus more on its transparency and the issues that follow with just that one concern. Trying to address the bias or fairness would take a whole other paper. Because it is WoZ, I plan on not telling the participants that they are seeing risk assessments and explanations created by me.. I will not tell them that the system is WoZ because that would alter the way in which the participants interact and feel about the proposed system. To best mimic an AI system, I will create a program that has the explanations for each of the three imaginary defendants' decisions hard coded in it, but appears to process the case in real time using artificial intelligence. The person running the session will enter in all the defendants data and then the system will pretend to send it through a neural network and return a recidivism risk assessment and explanation of varying

style. In this way, hopefully the users (either defendants or judges) will believe that it is an actual human AI system.

Explanation Styles:

From my prior research I learned that of the four that **Binns et al.** looked at, the worst explanation style in terms of interpretability was the case-based explanation so I will not include that. Instead, the three forms of explanation will be, from **Binns et. al**:

1. Input Influence:
    a. Presents a list of the input variables with a measurement of their overall influence, positive or negative, on a decision.
2. Sensitivity Explanation:
    a. Presents how much each input variable would have to change in order to change the final decision.
3. Demographic:
    a. Presents aggregate statistics on people in the same demographic categories as the defendant.

Participants:

I will randomly recruit 20 defendants and 20 judges from various backgrounds to prevent contextual bias. Ideally, the defendants and judges will have varying knowledge of law, recidivism risk predicting algorithms, and algorithms in general.

Experiment:

The session leader will brief the group of defendants on the task they are about to complete. The session leader will explain what the imaginary algorithm is and what it does in very basic terms. No technical jargon will be used so that participants that do not have experience with AI, algorithms, or recidivism rate prediction can still grasp the basic concepts.

After the briefing, each participant will be led through the "algorithm's" prediction process. Each defendant participant will receive details regarding each of three imaginary cases that I will devise. The imaginary cases will include a defendant, a collection of details about him/her, and any background information regarding why he is in court. The participant will then watch the session leader enter the case details into the WoZ and the participant will view the risk assessment along with the three explanations (one of each style) one by one. The order the different styles are given will be randomized to prevent one style being more comprehensible because it always comes after the other two styles.

After the risk assessment is viewed and each style of explanation is read, the participant will be asked to rank each style explanation based on how well they feel it explained the algorithm's decision from 1 to 10 with 10 being extremely helpful and 1 being extremely unhelpful. Then, the session leader will introduce a new imaginary case and continue the process until the participant has ranked all the explanation styles. This same process will be done to the group of judges as well, the only difference is more judicial related details will be shared with the judge regarding the imaginary case that will be given to him.

Results:

Once both groups have gone through the experiment session, each ranking will be compiled and those results will be analyzed to find the aggregate rating of explanation styles for both the defendant and judge groups. To do this, I will use the formula:

$$avg.\ rating = \frac{sum\ of\ all\ ranks\ of\ this\ style}{20}$$

for each style of explanation for each group. Then, I can compare each style's average rating within the judge and defendant group. Ultimately, hopefully both the judges and defendant groups will have a clear style that is on average rated highest. These results could potentially be used moving forward if there are efforts to make the WoZ an actual AI system with explanations built into the decision. Also, there is a chance the results can be generalized to other types of decision making algorithms, not just recidivism predicting systems..

Concerns:

Though I do randomize the order of the explanations, I cannot stop the participant from seeing multiple explanation styles and potentially gaining understanding from each explanation that comes before it, and then exaggerating the rating of the final explanation style. Moving forward, one solution to this issue could be to only show each participant or group of participants one style, but this approach would make it harder to measure how effective the style was at explaining the decision relative to other styles because other participants or groups of participants could have some background that affects how effective the explanation styles are at explaining the decision making to the participants or group of participants. By showing each participant all three styles, we control for the participant's contextual background and knowledge.

REFERENCES:

Binns, Reuben, et al. "'It's Reducing a Human Being to a Percentage'; Perceptions of Justice in Algorithmic Decisions." *ArXiv.org*, 31 Jan. 2018, https://arxiv.org/abs/1801.10408.

Chiao, Vincent. "Fairness, Accountability and Transparency: Notes on Algorithmic Decision-Making in Criminal Justice: International Journal of Law in Context." *Cambridge Core*, Cambridge University Press, 20 June 2019, https://www.cambridge.org/core/journals/international-journal-of-law-in-context/article/fairness-accountability-and-transparency-notes-on-algorithmic-decisionmaking-in-criminal-justice/635E1CB265F4F94335D2CAEBDC4D68EE.

Chopra, Samir. “A Legal Theory for Autonomous Artificial Agents.” *University of Michigan Press*, University of Michigan Press, 1 Jan. 1970, https://www.fulcrum.org/concern/monographs/wp988k715.

Green, Ben, and Yiling Chen. “Disparate Interactions.” *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, https://doi.org/10.1145/3287560.3287563.

Practitioner's Guide to Compas Core - Northpointeinc.com. http://www.northpointeinc.com/downloads/compas/Practitioners-Guide-COMPAS-Core-_031915.pdf.

Rudin, Cynthia. “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead.” *Nature News*, Nature Publishing Group, 13 May 2019, https://www.nature.com/articles/s42256-019-0048-x.

Shapiro, David M., and Monet Gonnerman. “State v. Loomis.” *Harvard Law Review*, 10 Mar. 2017, https://harvardlawreview.org/2017/03/state-v-loomis/.

Siegel, Eric. “How to Fight Bias with Predictive Policing.” *Scientific American Blog Network*, Scientific American, 19 Feb. 2018, https://blogs.scientificamerican.com/voices/how-to-fight-bias-with-predictive-policing/.

Team, PixelPlex. “Artificial Intelligence (AI) & Criminal Justice System: How Do They Work Together?” *PixelPlex*, PixelPlex, 2 Dec. 2021, https://pixelplex.io/blog/artificial-intelligence-criminal-justice-system/.

Walmsley, Joel. “Artificial Intelligence and the Value of Transparency.” *AI & SOCIETY*, vol. 36, no. 2, 2020, pp. 585–595., https://doi.org/10.1007/s00146-020-01066-z.

Zerilli, John, et al. “Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard? - Philosophy & Technology.” *SpringerLink*, Springer Netherlands, 5 Sept. 2018, https://link.springer.com/article/10.1007/s13347-018-0330-6#Sec1.