Machine learning bias game

Zhirui Hu[s2071894]

Media Technology, Leiden University
z.hu@umail.leidenuniv.nl

Abstract Machine learning is the science that helps computers uncover data patterns and relationships. It is a powerful tool that studies how computers simulate or implement human learning behaviors to acquire new knowledge or skills and reorganize existing knowledge structures to continuously improve their performance. But as these systems become more complex and powerful, researchers have found that the widespread use of artificial intelligence systems can cause some erroneous decisions. However, it is not the use of these AI algorithms themselves that is an issue, but rather that human biases are incorporated into the resulting models, and into the systems that use these models.
This paper focuses on the concept of how human bias effect on machine learning, analyzes the main reasons for its formation through two practical cases, and the impact of this bias on machine learning algorithms and the impact on practical engineering applications. Then we try to design a game in which we use regional crime rate prediction and the mathematical model of police deployment to combine the biased concepts we studied to show the impact of bias in the game and discuss how to eliminate this bias. Through the above work, we could get a better understanding of the concept of bias, and reflect on the incompleteness and defects of machine learning algorithms, thereby further improving the robustness, efficiency and reliability of machine learning algorithms.

Keywords: ML Bias · Game · Runaway Feedback Loop.

1 Introduction

Factors affecting crime rates are diverse, such as regional geography, education, social welfare, and economic conditions. With the development of the Internet and big data, we have accumulated a large number of relevant statistics. People have been trying to find the rules of these data to help decision-makers make better decisions, such as crime prevention and police deployment.

In recent years, with the development of artificial intelligence technology represented by machine learning, researchers have tried to explain these data through machine learning technology, and processed, labeled, and classified these data through a large number of crime rate data research in various regions, and then make data sets that can be used for modeling. Choosing the right machine learning algorithm, through a large amount of data learning, to achieve a mathematical model that meets the crime characteristics of the region, this

mathematical model can predict the crime rate based on historical data, and provide other effective information. Based on these forecasts, the police can deploy appropriate police forces to achieve optimal police efficiency utilization and to reduce crime rates in all areas most efficiently with a limited number of police officers. This has been a typical artificial intelligence application scenario.

Therefore, after using these data for relatively accurate analysis and research, combined with techniques such as machine learning and statistics and data mining, we can build an accurate model and use it in order to identify the pattern of crime, and finally predict the actual effect of the crime, these predictions are provided to the police for effective prevention.

1.1 What is machine learning bias?

However, human beings have certain judgments, and these judgments are often called priors. This does not mean that it is a necessarily bad thing, it is a survival technique, and in some cases, it can help us avoid dangerous situations. But these priors sometimes have an effect on the models produced by the algorithm, primarily through bias in data gathered for modeling, potentially leading to some terrible consequences.

The main manifestation is that with the development of technology and the development of various machine learning algorithms, people tend to unconsciously prejudice the programs they write or the models they built. This includes poor data selection, inaccurate data classification, missing some key sample data, etc., or letting their own biases penetrate into the programs they write or data they collect. Since machine learning systems are trained based on existing data sets, if the data set itself has some bias, or because the algorithm has certain limitations, the system makes the wrong decisions, which are sometimes contrary to common sense. If we use biased AI and machine learning algorithms to make decisions, this can lead to some serious problems and prediction errors.

1.2 Runaway feedback loops in predict policing

Machine learning algorithms can use big data to predict crimes in certain places. However, the predicted crimes might differ from the real situation. That is because the decision or action could further feedback into the algorithms, therefore, interfere in the next prediction process. Over time, it could be forming a vicious loop. We will find that the results of the learning model are against the attempt if the decision/action data feedback into the algorithm. There will be many police forces in one area, but few police forces in other areas.

We call this phenomenon a feedback loop [5], which means that the algorithm is stuck in a wrong loop and it is difficult to jump out. However, this is a general issue in machine learning, we tend to trust a model with a higher success rate. Therefore, we need to adjust our data and feedback strategies accordingly, for instance in this case correct predictive policing models by randomly dispatching

police forces to other regions, or setting a minimum policing limit for each region, which may offset the machine bias caused by the feedback loop.

We abstract the problem into research. In this process, we have made a simulation game to let people experience what is a runaway feedback loop, in order to draw more attention to this phenomenon.

This paper is structured as follows. In section 2, we look at the detailed problem, the cause of machine bias, and what is a runaway feedback loop. In section 3, we discuss two cases that relate to this problem. We further explain how we built the game, and how it works in section 4, followed by a discussion (section 5) and conclusion (section 6).

2 Literature review

Researchers have already noticed the problem of bias influence on machine learning and trying to find a way to solve this problem. For example, IBM developed a toolkit for detecting, understanding and removing bias in machine learning algorithms, and serve researchers focused on the fairness of machine learning [3]. In this section, we are trying to find out what are the causes of machine learning bias, explain what is a runaway feedback loop, and how it may occur in predictive policing.

2.1 The cause of machine learning bias

According to Pandey [14], there are three different causes for bias in machine learning. First, the use of skewed data. As machine learning depends largely on data, it is possible that something went wrong in our data collection, or data cleaning process. For example, the data that feeds into the machine learning algorithm does not represent the ground truth of certain groups. Secondly, the difference in different groups. For example, if we want to know the ratio of eating rice as a main food might be something like 81% for Asians, and for European people, it might be 12%. This difference will influence largely on the result of the algorithm. Finally, people might report more cases of false positives to avoid punishing based on the output of the model, that is, providing untrue information to get a better outcome from the algorithm.

However, there are ways that we could minimize bias, for example, Srivastava and Rossi [18] propose that third party rating on the decision of the algorithms, then the rating could be further sent back in the algorithms to assess the performance of the algorithms to reduce the bias. As well as the concept of explainable artificial intelligence could also, to some degree, reduce the bias by using a method that human could understand why and how it made such a decision.

2.2 Runaway feedback loop

Automation bias can ultimately lead to decisions that are not based on a comprehensive analysis of all available information, but rather strongly biased by

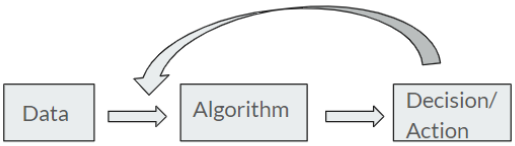


Figure 1: Runaway feedback loop

automatically generated recommendations [15]. We use data and algorithms to build a model, then further use it for predicting, making decisions and other actions. However, these decisions and actions then feed back data and outcomes into the models, forming a loop (Figure 1).

This situation exists in different disciplines in the real world. For example, once a decision is made to patrol a community, crimes found in that community will be sent back to the training apparatus for the next round of decision making, thus, the runaway feedback loop started[5]. Another example is Amazon's 'things you may want to buy' feature. For Amazon, this feature may motivate users to purchase. But this feature is built on top of the products that users have bought and viewed, perhaps after being recommended these products. At this time, the products recommended by the algorithm can only be regarded as one or two categories of the products that the user likes, and cannot represent all. The more users click, the lower the diversity of items that may be recommended, and the algorithm are already stuck in the runaway feedback loop.

2.3 Predictive policing

The definition of predictive policing is to identify possible targets for police intervention through statistical projections and prevent crime according to the RAND Corporation[16]. Lum and Isaac [12] mentioned that predictive policing is predicting future policing, instead of future crime. They have found out that using neighborhood observation data to update the existing model for predicting drug-related policing in Oakland could cause the runaway feedback loop. Ensign et al[5] further confirmed it, and try to find the reason why the runaway feedback loop occurs.

2.4 Other related work

On the basis of racial prejudice, human life has influenced the algorithms of programming and AI systems. There are some articles on gender bias in this aspect, although these studies are currently in the early stage, and most of the content is unpublished[7] [2]. What sparks the debate on this topic is an academic paper on a semantic library called the semantic corpus, which contains human-like bias [13]. There has been a great deal of research on embedded words and their

applications from web searches [11]. However, previous studies did not recognize vocabulary-embedded gender discrimination associations and their different biases that may be introduced in different software systems. The researchers used a benchmark for recording human bias, the Implicit Association Test (IAT). Although this test has been questioned and criticized by some academic circles, criticizing the validity of this research, this test still plays a big role in this research direction.

In order to test the intrinsic relationship between different words, the algorithm can generate the co-occurrence frequency statistics of words, that is, which words and those words have stronger connections [6]. Once the information index is complete, the research team examines a set of target words and filters a large amount of content that tells us about the potential biases that humans might inadvertently possess. The sample words are "programmers, engineers, scientists, nurses, teachers, and librarians", while the two sets of attribute words are male/female and female/female. In the results, the data emphasize bias, such as the preference for flowers over worms (which can be described as harmless bias), but the data identifies subject biases related to gender and ethnicity. A special case is that autonomous intelligence agents associate female names more with words that belong to the family, such as "parents" and "weddings" rather than the names of male characters. On the other hand, male names are more strongly related to words in their careers, such as professional and salary [13].

The project emphasizes that bias in the word embedding is actually closely related to the social concept of gender stereotypes. Stereotypes are described as unconscious and conscious prejudice among a group of people [17]. Many studies have explored stereotypes that contribute to the training of AI data [10].

3 Case Study

To better understand machine learning bias, in this section, we use two case studies to explain more in detail the impact of bias on the algorithm. The first case study, that the machine learning bias could occur in predicting recidivism, is a good example of potential bias in machine learning, as well as in a criminology context. The second case study is an example of runaway feedback loop, and an example of a creative approach towards making bias in machine learning experienceable, and understandable.

3.1 Model bias in recidivism risk assessment: ProPublica analysis of COMPAS

In 2017, ProPublica, a non-profit news organization, launched a project to reverse engineer Northpointe's commercial tools and models for recidivism risk assessment called COMPAS. COMPAS stands for 'alternative sanctions for correctional offender management'. COMPAS is used in the United States to predict the risk of a person to re-offend. Under the guidance of these risk assessments, judges in US courts will impose certain penalties on offenders, such as fines and

imprisonment time. ProPublica wanted to understand whether the COMPAS model contained bias against minority groups[9].

The agency obtained a data set of COMPAS risk assessment scores for a sample of 7,000 people in Florida and analyzed their probability of being re-incarcerated. COMPAS can predict the probability of a convicted criminal re-offending, but when the algorithm makes a mistake in the prediction, the type of error for whites and blacks is quite different. The main manifestation is that the probability of a black criminal re-offending is twice the probability of a white criminal re-offending, but in reality, this is not the case. The rate of recidivism by whites is generally higher, but the software has reached the opposite conclusion.

The United States is a country with the largest number of imprisoned people in the world. A large proportion of the imprisoned people are black people. Through these large amounts of data analysis, some conclusions have been drawn. The research shows that the offenders race, color, and nationality The results of these risk assessments have been largely influenced, black defendants are 45% more likely to be assigned high risk than white defendants, which is an unreasonable result. In 2014, former US Attorney General Eric Holder said that this analysis software for criminals did not perform well in court. At the same time, some research teams have also questioned ProPublica's research. They believe that there are many conflicts between the actual results and ProPublica's predictions.

This risk prediction software has obvious racial and gender discrimination. This is a very obvious erroneous prediction. This kind of erroneous prediction of African American racism is introduced when the sample data set is being constructed, or the learning algorithm is being produced. This is due to the long-standing prejudice against this problem and the introduction of this bias into the artificial intelligence machine learning system.

Since these algorithms do not disclose the details of their research, the specific evaluation details cannot be carried out, but the court has repeatedly evaluated the limitations of the software and started to use the software to make the final decision. Finally, the court concluded that the final score of the COMPAS risk assessment was based on some group data and was biased in the analysis of black crimes. At the same time, with the adjustment of the regional population, the data set used by this software must be regularly updated and tested. At the same time, this software can not be used reliably for criminal justice of a criminal. This software can only be used as an auxiliary tool and cannot give out sentences.

Through this case, some conclusions can be drawn. The evaluation software is biased. The machine bias of these algorithms is artificially introduced, because from the data point of view, the data is not wrong, and the actual evaluation is also correct. The only explanation is that research personnel introduced this kind of bias into the system when writing programs and designing algorithms. However, Northpointe disagrees with ProPublica by pointing out that the AUC scores are similar through both African American and white people [4]. However,

according to Angwin [1], the overall success rate of the algorithm is 60% for both races, but the success rate should not be the only way to assess the algorithm.

3.2 Games for raising awareness about runaway feedback loops: Monster Match

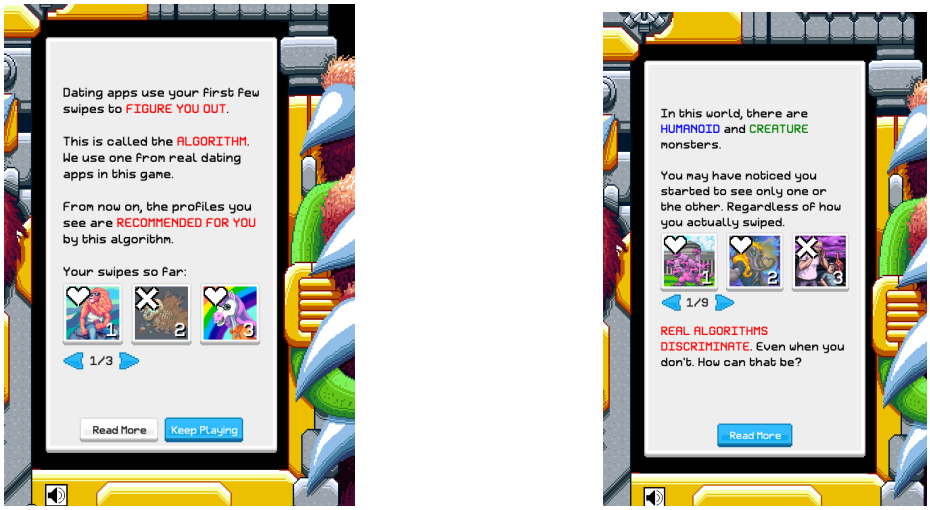
Monster Match [8] is a game for simulating online dating. Through some simple functions, this game shows that the algorithm may be affected and biased. This also reveals the bias in machine learning algorithm(available at `https://monstermatch.hiddenswitch.com/`).

Developer Ben Berman and designer Miguel Perez created a game to reveal the inherent bias in dating application matching algorithms. They have long noticed the strange problem of such an algorithm in machine bias, they are the winners of the Mozilla Creative Media Awards.

The game-play is quite simple. The users first create a profile and using the different features provided such as body, eyes, nose, mouth, etc. to piece together a monster avatar. Click on start and the user will see many other monster avatar photos in the game and these monsters are also profiles created by other players and have the same registration process. When the user sees them, he/she can swipe left or right on the avatar to choose whether he/she likes this monster and whether he/she wants to chat with him. The game will record the user choices and analyze the characteristics of his/her favorite monsters. The longer the user plays the game, the more the game will know the users "monster preferences" and it tries to find a pattern(Figure 2). Later, the system will recommend more of the same monsters, these monsters have many of the same characteristics as the monsters the users have chosen before, showing less and less variety.

Interestingly, many of users choices are unconscious, and these choices are short-lived, only a few seconds or even shorter. This game was trying to teach people how the bias occur in an algorithm. This software uses a "collaborative filtering technology" that can more accurately identify user preferences by analyzing many of the user's choices. This algorithm considers many factors, some of which are not even considered by the user. Therefore, this algorithm may find some potential selection rules on its own, because these biases filter a lot of content. Thereby narrowing the user's choice.

It is worth noting that, based on many choices, this technology can be used to provide users with movies and books they like, as well as people who might like social accounts. At the same time, the recommendation system has become a very important part of the retail, social networking, and entertainment industries. From giving suggestions about songs, recommending books, or fancy clothes to buy, and the recommendation system greatly improves the ability of customers to make choices more easily. Studies have shown that collaborative filtering increases the probability of bias, especially when the algorithm is recommending humans rather than movies or a product, the algorithm can narrow down the range of options based on other people's previous choices, which in turn discriminates against minority races, ethnicity, and sexual orientation.



(a) a (b) b

Figure 2: Monster dating game

The researchers contacted the designers and programmers of the software to try to discuss how to reduce this bias in the algorithm, but the developers refused to comment. The developers of the game have demonstrated that people's understanding of love and dating is quite different from how algorithms match people.

4 Game concept

The purpose of the simulation game is to make the concept machine bias easier to experience and understand by simulating a decision-making process (the game is available at `https://www.openprocessing.org/sketch/742355`). The users should be making decisions about where they want to dispatch the police force based on the predicted crime rate calculated by a machine learning algorithm. We try to solve the problems encountered in reality through this game, gain a better understanding of runaway feedback loops and then discover some methods to upgrade existing machine learning algorithms to avoid bias. At the same time, through this game, we can also attract more attention, to let more people become aware of this problem, and try to find a solution from different angles.

4.1 How does bias manifest itself in the game?

Crime is a common social phenomenon and presents obvious regional characteristics. As time goes by, Area A may become a high crime area, and crime rates in other areas are relatively lower, generally, The machine learning algorithm

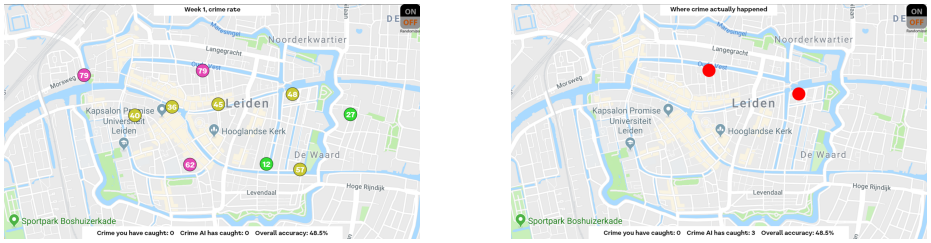
Figure 3: Prototype game

adjusts the arrangement of police forces in the city according to the historical crime rate. At this time, if more police forces are deployed in Area A to prevent crimes from happening, more crimes are observed in A and less in other places, which leads to a bias of the algorithm to Area A and less to other places where crimes may occur.

In the above, two examples are given. As we can see from ProPublica's research that bias is ubiquitous, mainly because all algorithms are created by humans. Which means people's previous information will be added, so bias is inevitable for many things. Therefore, bias is inadvertently passed to the algorithm when designing the model. This leads to these models not being truly objective and specific, and not reasonable enough. It seems this design flaw is impossible to solve.

In the Monster dating game, it further confirms the influence of runaway feedback loop in machine learning. That is, when an algorithm keeps using previous data to predict a future event, it most likely will be trapped in a loop. In the sense of crime prediction, for example, if we keep sending police force to the places which have the highest crime rate that predicts by the machine learning algorithm, and the data then being feedback to the algorithm again, it is possible that certain areas will be marked as high crime rate forever. However, crimes could happen in other places, but it will not be observed by the algorithm, because it focuses on those areas, therefore forming a "tunnel vision", which makes it cannot sense the situation elsewhere.

We have built a prototype (Figure 3) based on the research of Lum and Isaac[12], users could interact with any dots shown on the map. However, for a game that delivers the concept of the runaway feedback loop, the prototype contains too much information that may cause users could not easily operate the game. Thus, we have further simplified the game.



(a) a (b) b

Figure 4: Game play

In this game, we simulate an algorithm to predict the crime rate in the city of Leiden, the Netherlands. However, it could be replaced by any other location, for example a location of choice from a user, as we do not use any real data about the location.

The algorithm highlights the predicted crime location and the percentage of crimes that could occur. As shown in Figure 4. Each round the users has to choose where they want to dispatch the police force based on the crime rate that has been highlighted by the algorithm (Figure 4a). Figure 5 shows the complete game-play without the randomizer on. The red circles represent the users choice. Users, as well as the AI, could choose three places each week/round, however, we could not see the AI's choice. After the selection, the places where the crime happened will be highlighted (Figure 4b). It is worth to mention that the mechanism of the places where crime actually happened is that, the two places are random, but based on the percentage of all predicted places, the higher the crime rate, the easier it is to be selected.

Then the players and the AI will get 1 point for each place they predicted as same as where the crime acutely happened. However, the prediction is just background information and that the users are free to select the three places where they think crime will most likely happen. Meaning they could follow the advice of the predictive policing algorithm but also could decide to send police forces elsewhere.

Under this mode, the users are expected to notice that the algorithm focuses more on the places users have chosen, and start to become biased because it not only increases the predicted criminal rate for those places, also deleting and ignore other places. As the game goes on, there are less and less predicted locations where crime might happen, in the end, there are only three places shown on the map, which means the algorithm stuck in the loop, and becomes less and less useful for making any prediction (Figure 5).

Similarly, the users are also expected to notice that the crime rate has not decreased in the three dispatched locations they have chosen. On the contrary, the crime rate has risen to varying degrees in these places because the influence

Figure 5: Game play without randomizer: The red circles represent the users choice each turn.

of the runaway feedback loop. That is to say the more police forces deployed, the higher crime rate these places will have in the future.

In the above game process, we found that using the machine learning algorithm to predict the crime rate and deploy the corresponding police force, this method has caused a lot of concentration of police forces in particular areas, even saturated, while there is no police force in other areas, thus, the algorithm is trapped in a runaway feedback loop. As the game goes on, these areas with high crime rates are still very high and will be increased in the end game. At the meantime, while in other areas, the crime rate has been ignored by the algorithm completely.

4.2 Countering runaway feedback loops with randomization

It is important to add something to the prediction process to avoid the program getting stuck in such a loop. This problem can be abstracted as a dynamic search for the most advantageous strategy to optimize the objective function, which is in our case the overall number of criminals caught.

In the process of solving each iteration, we can make corresponding adjustments to the police force allocation calculated by the algorithm, which means that for each predicted value each round, a part of the random quantity is added, so that in the game, the function of random is added to the algorithm, when the algorithm ignore and removing the locations, we add two random locations with random value, which may offset the machine bias brought by each iteration.

The specific game-play is that we display an adjustment window in the game, where randomization optimization is performed for each deployment result. Meaning that every time, after users made their choice, the randomizer generates two random dots on the map, to offset the influence of the runaway feedback loop, which is to remove one dot on the map. Therefore, the algorithm could predict for more rounds (Figure 6).

The player should be able to adjust the randomness of the algorithm, thus alleviating the strange phenomenon that the algorithm predicts that crimes continue to rise in the same place. In practice, users could click the randomizer button on the upper right corner, each week/stage, after users choose the place their think where crime will happen, the randomizer generates more places where crime event may occur. At the end of the game, instead of ignoring other places except where users have chosen, the algorithm could continually provide relatively more useful advice for the decision-maker. The difference of with (Figure 7a) and without (Figure 7b) the randomizer is quite clear.

We further compare this modified map with the initial police deployment map. We found that normal machine learning algorithms can achieve the need to deploy a lot of police forces in key areas. However, as time increases, due to the bias, or more specifically, runaway feedback loop, there are a lot of police forces in the key areas, and there are few police forces in other areas, which makes the program fall into an infinite loop. However, by using the optimized algorithm, we could minimize the influence of bias in the machine learning algorithm.

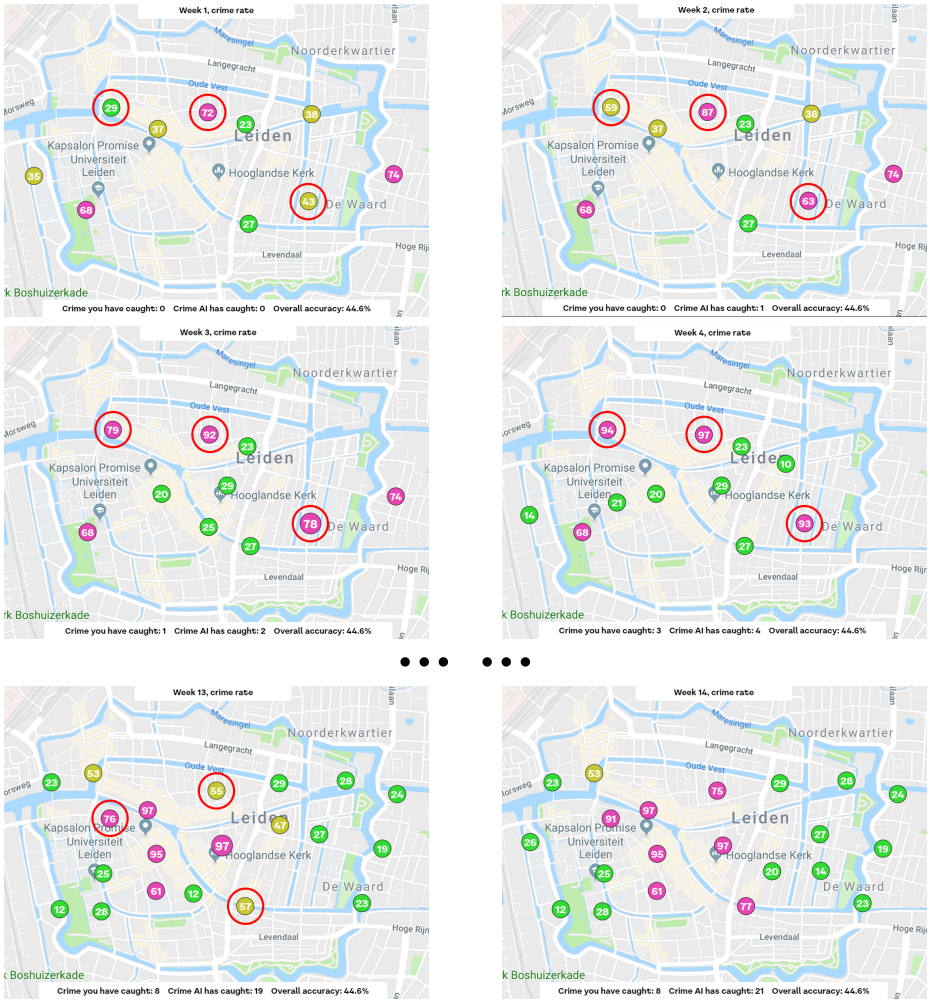
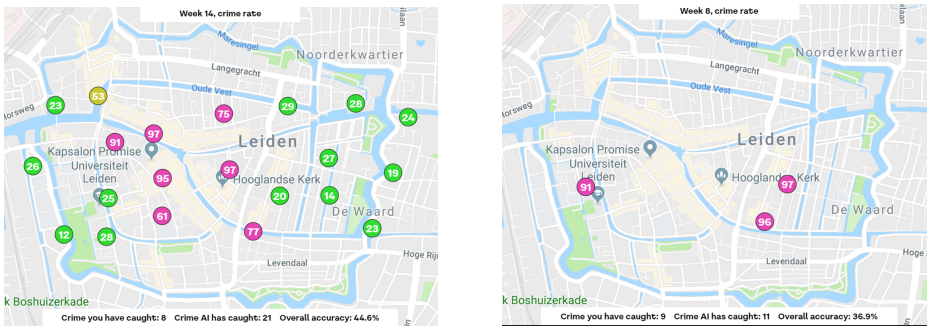


Figure 6: Game play with randomizer: The red circles represent the users choice each turn.



(a) a (b) b

Figure 7: End-game with[a]/without[b] randomizer

5 Discussion

This game focuses on the runaway feedback loop by developing an experience for the users. The goal is to expose the problem of machine learning to users. More specifically, the more rounds the users use the randomizer, the better the algorithm performs, and the AI will suffer less from the runaway feedback loop, so it becomes harder to beat

An interview has been done to make further verification that whether the game achieved the goal or not. We have tested the game on an actual user with a background of communication and now work as an assistant manager at a retail store. The first time the user played the game we build he was using primarily the recommendations that the simulator was giving him. This resulted that the user lost to the AI with 7 points. This because the user kept putting his police forces in the high crime rate areas as predicted by the game. It took the user until the second round of the following game to discover the effect of the randomizer button.

By using the randomizer the game was kept out of the loop and this gave him more control over where to place his police forces and thus he was able to defeat the AI in the game. However, in principle, the randomizer supposed to give better results for the AI algorithm, so it should be harder to beat. But indeed in practice, users get more locations to chose from in this game . In this way, it showed the user the effect of the feedback loop that is created by a bias machine learning. The user thought he was winning the game the first time he was playing it. This because the lower crime rate areas were disappearing and only the high ones were remaining. But the user was losing the game instead. After getting more information about the game, the user was able the understand why the game fell into a loop. This means that instead of putting his police forces in the city more effectively he was focusing on a smaller area, which made his future prediction less accurate and prevented him to effectively place the police in areas where they were needed.

The concept of this game that trying to deliver is the runaway feedback loop and the result of it. Indeed, it is difficult to win the game. However, users could have more chance to win the game when the randomizer was turned on, because the algorithm performs better, therefore less biased.

Future game development could be done such as adding a built-in charts to display the number of criminals caught over the weeks, as well as compare differences in the criminals caught across sessions (and whether the session was randomized or not).

In terms of the user interface, it could be clearer which places have been chosen by the users in the last round, instead of jump into the next round immediately.

This project focuses on finding the possibility of visualizing the runaway feedback loop in a game, due to time limitation, we did not carry out a larger scale user evaluation. The possible questions could be based on the experience of the runaway feedback loop, the awareness of the changes in each round, and between sessions with or without the randomizer. An open question on what do the users think is the best way to improve the game could be used on improvement in game design to better express the concept of the runaway feedback loop.

Moreover, this game was a simulation based on theories of bias in machine learning and runaway feedback loop. However, Lum & Issac [12] used a realistic algorithm based on different types of data collected in Oakland to analyze bias. Further research could be done making the algorithms more realistic or more like real predictive policing algorithms, compare with the simulation game, for example how much crime would the regular AI catch versus the randomized version over large numbers of simulated runs, as well as a more realistic reflection on the bias and certain type of predicting algorithm.

6 Conclusions

This paper mainly focused on the influence of bias in machine learning algorithms, more specifically, the runaway feedback loop, through two preexisting cases and a game that was developed.

In the case of ProPublica, they evaluated commercial tools and found that in the final crime rate prediction process, obvious misjudgment and predicted results are obtained. Although the algorithm has a relatively high success rate for predicting the re-offend for both black and white criminals, However, blacks are 45% more likely to be assigned high risk, which is twice more likely than whites. While white criminals are twice more likely to be assigned as low rick.

There is also Monster Match, a creative dating game. It uses statistical analysis of personal interests to recommend the corresponding date. We found that after a while, the system only recommends a less diverse set of dates to the user, which are not in line with the actual situation and logic.

We have designed a police deployment game. After the deployment of the police force in the city of Leiden, the Netherlands, according to the crime rate of different regions, the machine learning algorithm was used to recommend the

deployment of the police force in the region. After a while, there have been obvious logical errors, and there are many police forces in key areas. There are no police forces in other areas, so we were able to reproduce a runaway feedback loop in the game.

The cases and the game show us how bias affects the final result. Through these cases, we found that the machine learning algorithm is not perfect. The bias is stored in many algorithms, and the main reason for machine learning bias is the runaway feedback loop. However, we noticed that after recognizing this problem, we could minimize the bias in machine learning by adding some randomness into the algorithm, so that it could reduce the chance that the algorithm gets stuck in the loop.

We realize that it is impossible to eliminate bias for now. However, we could use some methods, such as the game, to let more people notice this problem, thus motivating more people to study and solve this problem.

7 Acknowledgements

I would first like to thank my thesis advisor Dr. P.W.H. van der Putten of the Media Technology at Leiden University. The door to Prof. van der Putten's office was always open whenever I ran into a trouble spot or had a question about my research or writing. He steered me in the right the direction whenever he thought I needed it. Without his passionate participation and input, this project could not have been successfully conducted. I would also like to acknowledge Mr. Jichen Wu of Media technology at Leiden University , and I am gratefully indebted to him for his help with putting the game online, and debugging. Finally, I must express my very profound gratitude to my parents and to my grandparents for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you. In memory of my grandfather Yangchun Hu.

References

1. Angwin, J.: Ai now: Uncovering machine bias[video file] (2016), `https://www.youtube.com/watch?v=Ts35luE59d0`
2. Bass, D., Huet, E.: Researchers combat gender and racial bias in artificial intelligence. Bloomberg.com (2017), `https://www.bloomberg.com/news/articles/2017-12-04/researchers-combat-gender-and-racial-bias-in-artificial-intelligence`
3. Bellamy, R.K.E., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K.N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K.R., and, Y.Z.: Ai fairness 360: an extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. IBM (3 Oct 2018)
4. Brennan, T., Dieterich, W., Ehret, B.: Evaluating the predictive validity of the compas risk and needs assessment system. Sage Journals **36**(1), 21–40 (2009). https://doi.org/10.1177/0093854808326545

5. Ensign, D., Friedler, S.A., Neville, S., Scheidegger, C., Venkatasubramanian, S.: Runaway feedback loops in predictive policing. In: Friedler, S.A., Wilson, C. (eds.) Proceedings of the 1st Conference on Fairness, Accountability and Transparency. Proceedings of Machine Learning Research, vol. 81, pp. 160–171. PMLR, New York, NY, USA (23–24 Feb 2018), http://proceedings.mlr.press/v81/ensign18a.html
6. Gelman, A., Fagan, J., Kiss, A.: An analysis of the new york city police departments stop-and-frisk policy in the context of claims of racial bias. Journal of the American Statistical Association **102**(479), 813–823 (2007). https://doi.org/10.1198/016214506000001040
7. Gorner, J.: Chicago police use heat list as strategy to prevent violence. Chicago Tribune (2017), https://www.chicagotribune.com/news/ct-xpm-2013-08-21-ct-met-heat-list-20130821-story.html
8. Kraus, R.: A dating app for literal monsters exposes the bias in our swipes. Mashable (2019), https://in.mashable.com/tech/3599/a-dating-app-for-literal-monsters-exposes-the-bias-in-our-swipes
9. Larson, J., Mattu, S., Kirchner, L., Angwin, J.: How we analyzed the compas recidivism algorithm. ProPublica (2016), https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
10. Lerman, A.E., Weaver, V.: Staying out of sight? concentrated policing and local political action. European Journal of Social Psychology **156**(1), 202–219 (2013). https://doi.org/10.1177/0002716213503085
11. Levitt, S.D.: The relationship between crime reporting and police: Implications for the use of uniform crime reports. Journal of Quantitative Criminology **14**(1), 61–81 (1998). https://doi.org/10.1023/A:1023096425367
12. Lum, K., Isaac, W.: To predict and serve? Significance pp. 14–18 (2016)
13. Mohler, G.O., Short, M.B., Malinowski, S., Johnson, M., Tita, G.E., Bertozzi, A.L., Brantingham, P.J.: Randomized controlled field trials of predictive policing. Journal of the American Statistical Association **110**(512), 1399–1411 (2015). https://doi.org/10.1080/01621459.2015.1077710
14. Pandey, P.: Is your machine learning model biased? how to measure your models fairness and decide on the best fairness metrics. Towards Data Science (2019), https://towardsdatascience.com/is-your-machine-learning-model-biased-94f9ee176b67
15. Parasuraman, R., Manzey, D.: Complacency and bias in human use of automation: An attentional integration. Human Factors: The Journal Of The Human Factors And Ergonomics Society **52**(3), 381–410 (2010). https://doi.org/doi:10.1177/0018720810376055
16. Perry, W.L., McInnis, B., Price, C.C., Smith, S., Hollywood, J.S.: Predictive policing: The role of crime forecasting in law enforcement operations. RAND Corporation (2013). https://doi.org/10.7249-RR233
17. Sewell, A.A., Jefferson, K.A., Lee, H.: Living under surveillance: gender, psychological distress, and stop-question-and-frisk policing in new york city. Social Science and Medicine **159**, 1–13 (2016). https://doi.org/10.1016/j.socscimed.2016.04.024
18. Srivastava, B., Rossi, F.: Towards composable bias rating of ai services. IBM T. J. Watson Research Center (2018)