EasyChair Preprint

№ 6170

Competition and Conflict Between Frames in Using Machine Learning

Hebah Bubakr and Chris Baber

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

July 27, 2021

Conflict Between Frames When Using Machine Learning

Hebah BUBAKR[a] and Christopher BABER[b]
[a] The Department of Electronic, Electrical and Systems Engineering, University of Birmingham, HXB815@student.dham.ac.uk, on leave from King Faisal University, Saudi Arabia, Habubakr@kfu.edu.sa
[b] University of Birmingham

ABSTRACT

Artificial intelligence and Machine learning (AI/ML) systems promise to improve organisational decision making by avoiding bias because the machine ought to remain unaffected by moods, prejudices or personal opinions when interpreting the data. However, this promise rests on the fact that these tools are independent of the biases of their developers. The purpose of this paper is to investigate what bias means to developers of AI / ML systems, and how they interpret bias through the result of the system. Our concern is with the relationship between developer - algorithm - data - output. In this paper, we applied the Data/Frame Model (DFM) to understand what decisions are made by developers of AI/ML. We propose that developers work with three distinct frames. First, they need to define a suitable dataset that will answer specific questions and also be amenable to analysis. We term this the 'dataset frame' and it includes factors such as size, representativeness and coverage, type of questions that could be addressed using these data. Second, having selected datasets, participants then explored different algorithms to test the selected datasets. We terms this the 'Algorithm Frame'. Third, once the algorithm produces answers, then these are reviewed. We terms this the 'Interpretation Frame' which includes both judgement on the performance of the algorithm (so overlaps with the 'algorithm frame'), plus judgement on the interpretation of the output to the original questions, and also judgement of the implications of this interpretation. Our conclusions suggest that developers of AI / ML might take a narrow perspective on 'bias' (as a statistical problem rather than a social or ethical problem). This is not because they are unaware of wider, ethical concerns but because the requirements relating to the management of data and the implementation of algorithms might narrow their focus to technical challenges. Consequently, biased outcomes can be produced unconsciously because developers are simply not attending to these broader concerns. This suggests that the 'interpretation frame' ought to be elaborated to encompass the implications arising from possible interpretations of the algorithms' output.

KEYWORDS

Judgment and Decision Making, Bias, Data frame model, Artificial intelligence, Machine learning

INTRODUCTION

Recent studies have focussed on explainable AI (Borgo et al., 2018; Anjomshoae et al., 2019; Baber et al., 2020, 2021) in which people interpret the output of the AI. The focus of these studies has been on developing our understanding of the ways in which people make use of the output of AI / ML systems. There has, we believe, been less attention given to the ways in which the systems are developed. Thus, our concern in this paper is with the relationship between developer - algorithm - data – output during the development of AI/ML. In order to explore this, we applied the Data/Frame Model (DFM) to understand what decisions are made by develolpers of AI/ML.

SENSE-MAKING AND DATA-FRAME MODEL

For Klein et al. (2006 a) "Sense-making is a motivated, continuous effort to understand connections (which can be among people, places, and events) in order to anticipate their trajectories and act effectively" [p. 71]. An important point to emphasise is that frames change as we add, change, delete or modify data. As shown in Figure 1, a frame allows a person not only to decide if data are sufficient, but to also project into the future what data will be needed. In this case, data can provide anchors on which to construct a frame which can then lead to search for further data. Klein et al. (2007) pointed out that the environment, individual characteristics, and available information will additionally influence the frame that is chosen. Once a frame has been selected, we seek data that relate to the frame. Any data that does not fit the frame is likely to be ignored or redefined. If the data are not suitable, then we might switch to a different frame.

Each element in the sense-making process forms relationships with others, using anchors, cues and any relevant data found in that environment. Thus, "when people try to make sense of events, they begin with some perspective, viewpoint, or framework" (Klein et al (2006 b). In terms of this perspective, there may be situations in which the output produced by AI might not be acceptable socially, at least, to the people who are affected by that decision. This might be due to failure of the algorithm, but more likely, it arises from the selection of data that impose the 'frame' upon which the algorithm operates. This means that the datasets might simply reflect biases and injustices that are prevalent in the society that produced these data (O'Neill, 2016). The issue that we explore in this paper is how do developers of AI / ML deal with the data and what sort of frames do they apply in their development and application of algorithms?

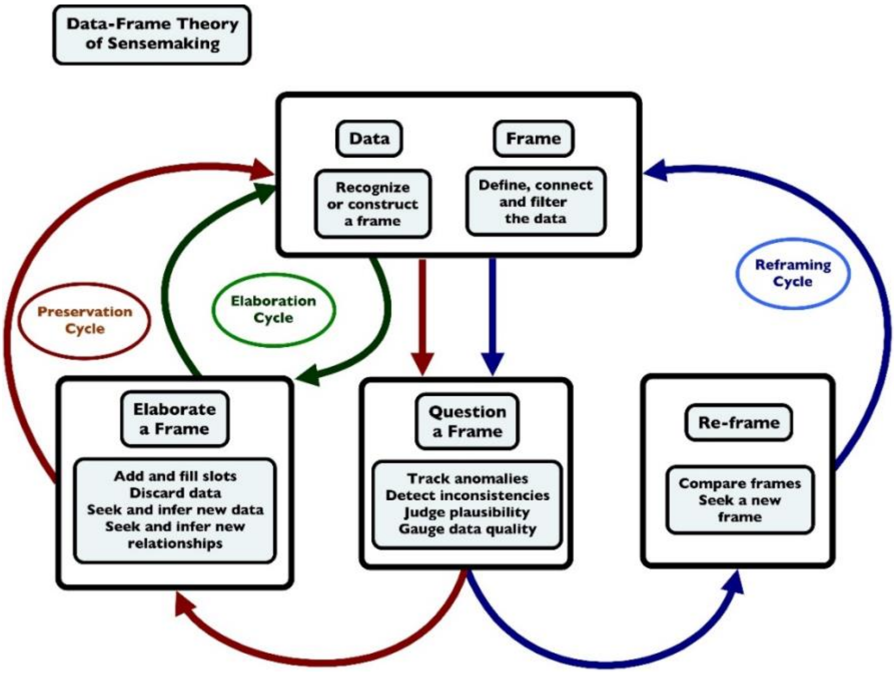

Figure 1: Data-Frame Model (after Klein et al., 2007)

METHOD

To understand how AI / ML developers work with their 'frames' we need to understand the process of creating an AI system. Interviews were conducted with Computer Science undergraduates who completed a project implementing Machine Learning on medical datasets. In order to encourage them to focus on the potential issues relating to 'bias' we devised a project which we called 'the bias machine'. The goal of the project was to deliberately produce biased output from datasets in a which could allow 'bias' to be dialled up or down, i.e., from no bias to high-levels of bias which would disadvantage particular groups of people. Using open access databases, participants selected two datasets to 1) make predictions about future smoking habits of different user groups based on racial profiles, and 2) classify severe visual impairment based on age, gender and race. From these datasets, participants would adjust datasets to inflict bias and implement the algorithms to produce biased results. After the AI / ML development was complete, participants gave a presentation on their work. Following this, we conducted Critical Decision Method (CDM) interviews with the participants, focusing on three high-level questions:

- What frame do developers apply when they select a dataset?
- What frame do developers apply when they select an algorithm?
- What frame do developers apply when they are interpreting the results?

Participants

This study involved Undergraduate students from the School of Computer Science, University of Birmingham. We accept that this might be seen as a limitation in the study but argue in our defence, that many implementations of 'routine' AI / ML will be performed by junior employees, i.e., recent graduates, in companies, and that their knowledge of methods could be similar to those of our student participants. The study employed in-depth qualitative investigation, involving interviews with three participants. The design of the study was approved by the University Ethics Review Board. Prospective participants were approached based on their experience in computer systems and software engineering, and the three who agreed to participate did so voluntarily.

The project

Rather than simply ask people to analyse datasets, we set them the challenge of manipulating the data to produce different levels of 'bias' in the output. In this way, they were working on a project that aimed to produce a 'bias machine' that could dial up or down 'bias' (from no bias to high levels of bias). Since the participants chose to address the health care system, working with medical datasets, 'bias' could be defined in terms of 'social' factors (age, gender, race, social class, level of healthcare etc.) or 'medical' factors (e.g., disease prevalence, predictive validity etc.). The activity began with a search for suitable datasets; the definition of suitability included factors such as size, representativeness and coverage of data, type of questions that could be addressed using these data etc. Most of the medical datasets have missing or incomplete data on patients, or certain patient groups or medical conditions could be over- or under-represented. Often this will require the analysts to modify the dataset, e.g., through normalisation, through consistency of labelling, through ensuring a balanced distribution of factors in the data. These processes could introduce distortions to the original dataset. Furthermore, processes of attribute sampling (in which elements in the dataset are selected on the basis of their relevance or expected contribution to the research question) can also introduce bias into the choice of data. In this project, two datasets were selected on the basis of record sampling, i.e., the project would use the complete dataset and there was little need for cleaning of the data (e.g., in terms of dealing with missing values):

- The Tobacco dataset to make predictions about future smoking behaviors relative to age and race.
- The vision and eye health dataset showed patients who were severely visually impaired, or blind based on a person's age, gender and race.

Working with datasets

The datasets were collected from a US website that provides access to a huge number of large datasets from different disciplines. Since the project focused on the medical field, the participants selected two databases to test bias using different AI / ML algorithms. The selected datasets were chosen for their size and variation. At first, participants kept 100% of the data so that they could demonstrate how selective filtering by the algorithms could force bias. Unfortunately, datasets were still subject to problems in sampling, labelling and representativeness. For example, certain race, gender or age groups were disproportionately represented which reflect the demographics of the source of the data but could have an impact on the quality of the output from the algorithms. However, decreasing the quantity of data for a particular group (perhaps to provide 'balance' in the dataset) by 6% was enough to produce a marked change in the results.

Working with algorithms

Having selected datasets, participants then explored different algorithms. Three different algorithms were applied (Neural Network, K-NN Classifier and Linear Regression). Linear Regression uses data points to identify a trend, and then makes a prediction by extrapolating this trend; K-NN Classifier predicts classes based on the percentage of given parameters; Neural Networks classify data based on two-thirds of the dataset for training, and one-third for testing. Participants applied the selected algorithms to the datasets, modifying parameters in the algorithms or datasets to introduce different outcomes. The aim was to identify parameters that would reliably produce 'biased' results (and to confirm that these parameters could also be specified to minimise bias). Meaning, some fields were removed and that affected the data, for instance, year end was removed (from tobacco dataset), so data collected at the end of a year (overlapping into the following year) would be treated the same as data collected at the start of the year. Moreover, Location would likely have a large effect on data - poorer areas with worse healthcare would have quite different results compared to richer areas. The sample size had also been ignored or in one dataset was seen as an 'equal' spread but it was not representative of the population distribution of the USA (where the dataset was collected from) so could be biased to a particular race. Therefore, the results of this exercise formed the basis of group presentations of the design and development activity. The choice of algorithm is considered in the analysis and results section.

Interviews

Three participants were interviewed on an individual basis to explore their beliefs and understanding of bias in the study. These interviews were conducted using the Critical Decision Method, with the probes shown in Table 1. Each interview took approximately 45-60 minutes.

Table 1: CDM questions

Probe type	**Probe content (the interview question)**
Cues:	Think about the last time you worked on machine learning (ML) project? What was about? Can you please talk in general about your project on Bias in Machine learning? What was the most difficult thing about the project? What were the feature that you were looking for in a dataset? What made a good dataset? What were the feature of the algorithms you apply? what made a good algorithm?
Information:	Why you think that the feature you defined first, were the best feature you would use? Why do you think this dataset is the best choice for your model? Can you explain the dataset? Do you think the data can be improved? How where the data collected? What are the reasons for selecting these specific datasets? Who would benefit from this?
Analogues & Experience:	What do you think is the main purpose of machine learning? Did you have previous experience with other algorithms? What about that previous experience did it seemed relevant to this project? Were you reminded of any previous experience using this method? What specific training or experience was necessary or helpful in making this project?
Standard Operating Procedures:	What is the process that you use to make these algorithms? (what data sets, where from, how validated, how sampled) Can you draw a flowchart or timeline for doing this? what decisions happen at each point… What is the process that you use to make the final decision? Can you draw a flowchart or timeline for doing this?
Goals:	Is there different output you can or want to make? Do you think the model can present more biased results? How? What do you think is the main purpose of presenting bias? Did the model present the specific goals you planed?
Assessment:	How do you know that you produced a biased result? What would make a good bias result? What would make a good unbiased result?
Mental model:	How do you train the models? How do you tune the algorithms? What model fitting do you do?
Decision making & Options:	Which algorithms do you think would be most useful for this application? Which ones did you use? Why did you select these? Which other (data) method do you think would be most useful for this project? Why did you select these methods? What other courses of action were considered or available to you?

Analysis and Results

Interviewing the participants in this study was conducted after their team had presented the results of their project (implementing bias Machine Learning algorithms on medical datasets). We started the interviews by asking the CDM questions (table 1) in order, encouraging the participant to express themselves as much as possible. However, in this section, we are highlighting only the most salient questions and answers.

Why do the project?

We asked the participants "What do you think is the main purpose of presenting bias?", this was to understand their motivation and how they viewed bias. P1 said, "The main purpose was essentially an educational tool to show how easy it is to have bias algorithms and more like how hard it is to have unbiased algorithms to the point of almost impossible." P2 said the project was aimed at showing that not every algorithm is suitable for any dataset, people can misuse them. You will only get a bias output if your inputs or process are biased. On the other hand, for P3 it was much harder to define bias P3 said "We did not know what unbiased would be. I suppose that is one of the challenges because it was hard to tell."

How to define bias?

Answers to the initial question give us an idea of how participants understood bias and how this was related to some wrong and questionable mathematical results. They linked bias to how well or bad the numbers looked in the datasets and how many error values the algorithms produce. This point of view was confirmed when we asked participants to describe the features that made a good dataset. P1 states, "For datasets, we were looking for kind of equal spread attributes, so we had a dataset which has race, gender and where people lived and we wanted it to be equal in terms of spread, I guess, so we wanted an equal amount of every race to be included to make sure it was not biased in any way or another in that kind of sets." Although, p1 just described a sample bias, P1 is also believing that if the data is statistically balanced then it is 'good'. However, sometimes 'balanced' data is not representative of the real-world population. Moreover, as P3 said "If it had one group that was more likely to smoke and we knew for a fact that was not true, that would be wrong. For example, we had a high percentage of the population smoking, and the algorithm assumed it was from the American Indian / Alaska native racial categories. This prediction is totally unfair…because, in fact, American Indian/Alaska Native were hugely overrepresented in the dataset as in reality, this group only makes up about 0.8% of the total population." This is not only a bias or fairness issue, but with a certain part of the population overrepresented in the dataset the result of the model or the prediction probably will be wrong. P2 said, "It is a hard question, but you just can tell bias when you see wrong values, or the results do not make sense, and there were overlapping and that cause bias because it is just not working, and it will not give right predictions".

Reason for choice of algorithms.

The teams used three algorithms to test bias (Neural Network, K-NN Classifier and Linear Regression) the reasons for selecting these were, because they had some experience with these algorithms, they were among the popular algorithms in ML and these algorithms gave a clear biased result according to the group. The participants accepted that the selection of algorithm was not based on absolute principles, but more like testing and exploring what works. P2 stated that the algorithms they used seemed to be very popular and P1 said "Honestly, we struggle to find a GOOD algorithm, and the ones we tested were very unreliable, and most of them had so many big issues that we really could not see them being used in the light of the application we tested", while P3 said "we tried the leaner regression, K-NN classifier and the Neural Network, and they were different from each other. Not so sure, but they were good enough." What was much more difficult for participants was to provide a coherent definition of what they meant by 'good'.

The outcome of the algorithm

We asked participants if the model produced results that fitted the goals of their project (which is to present bias) and two out of three agreed that they had the result they worked to achieve. P1 stated "yes, it showed how easy bias is added. Even if it not intended because before we start adding bias to things, we also tried to make an unbiased version of the algorithms and that was still biased" P2 also confirmed this point by saying "yes, this project can raise awareness about how people can be wrong in using certain algorithms with certain datasets. Just make people aware that high-end tools are not necessarily clever. If you get bad data, you will get the bad stuff out. It depends on what you feed your algorithm also what algorithm you are using as well." While p3 was expecting more obvious bias results and address some of the limitations that led to their output "The results were unexpected. I think we were imaging something that would have been achieved if we had a bigger dataset. So, I guess that is one of the main limitations (having small datasets) well, it was the largest we can find, but still, the model needed bigger." P3 continued by saying "We limited ourselves in the medical field where we could have much larger datasets, and it would have been useful. Mainly the bias we are finding was more down to the very limited size of the dataset rather than any other limitation regarding the data."

Conscious or unconscious bias?

Our last point to investigate was the possibility of producing the same result unconsciously, P2 said "I think our results can be produced unconsciously because some people are just not being aware that bias can be produced and then being naïve in which algorithm you chose for a certain dataset. I mean we did not even have to try to get this result. We all tried not to be biased in a way because everything was just falling"

Where does bias lie*?*

To highlight the reason for the bias or where it comes from, we asked the participants what caused this bias, algorithms or data? P1 said "In terms of who is most at fault, I think the model itself and the algorithms we used are very biased. They put their own biases on top of the dataset, well I do not think the datasets are perfect, but they are considered not too bad either. We did not find a lot of extra bias, but the algorithms would take this small amount of bias in the datasets and extrapolated it. In other words, the dataset is what causes the bias, but the algorithm allows it to make a big issue." P2 also agreed by saying "It is both. If you have a certain dataset you have to study it, and you did some tailor to select what algorithm you should use and vice versa, if you have an algorithm you have to tailor the dataset, so it worked the algorithm, so I would say." However, P3 had a very strong opinion that datasets are the cause of bias, "Definitely, it all depends on the data, if the dataset is not representative, then the output is not representative either. I think people produce these outputs because of bias dataset, and it is much difficult for the algorithms to be biased compared to the actual data you put in with"

The Frames

From figure 1 and our interviews, an initial observation is that the 'data' that forms the dataset for AI/ML is both 'data' and 'frame' for the developers, in that the choice of dataset (in terms of what it contains and the relationships it defines) creates the definition of the 'frame' (in terms of which algorithm to select, what questions could be asked, the dataset balances between factors etc.). To develop this further, we selected the elements of the frame based on the participants' answers to the detailed questions about the datasets and the algorithms. In this content analysis, we only used elements that were mentioned by 2 or more of the interviewees.

The first frame was the dataset frame, and it shows what was the main features of selecting the dataset from the participant's perspective as a team. As Figure 2 shows five features led them to select the Tobacco and the vision datasets, first, it had to be from the medical field, it should be big enough for the algorithm to execute, test and learn. Further, datasets must be diverse (has different entries to explore bias). It is also better to have an equal spread of attributes because there is no point exploring bias with a very bias dataset. Finally, it should contain lots of personal information, one reason for this feature is because personal information even if not explicitly used by the algorithm, affects the outcome. Another reason is lots of real-world companies use this info in their processes. Therefore, selecting datasets that got all these characteristics in their beliefs will affect the fairness of the model. According to P1, the data is a critical explanation for any biased system even if it was not the only reason.

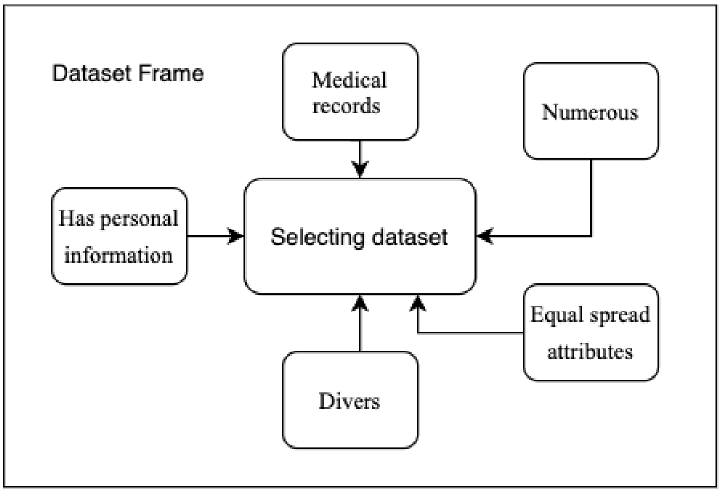

Figure2: dataset frame

The algorithm frame presents features of what makes a good algorithm and what are the reasons for selecting the three algorithms that were used in the project. It was most likely a straightforward process where lots of exploration was conducted to see what fits best. As Figure 3 shows, algorithms must be testable, used previously in ML, reliable and some members of the team (the participants) had an experience using the algorithm before the implementation. First let's make it clear that when the team decided to select an algorithm, they did not intend for a biased one, but the idea was to find a good algorithm to explore with the data they have. Unfortunately, even when not aiming for bias, the result suggests bias execution from the algorithms and unreliability, this was described earlier by the team and how they struggled to find a suitable algorithm, and how unreliable their algorithms were.

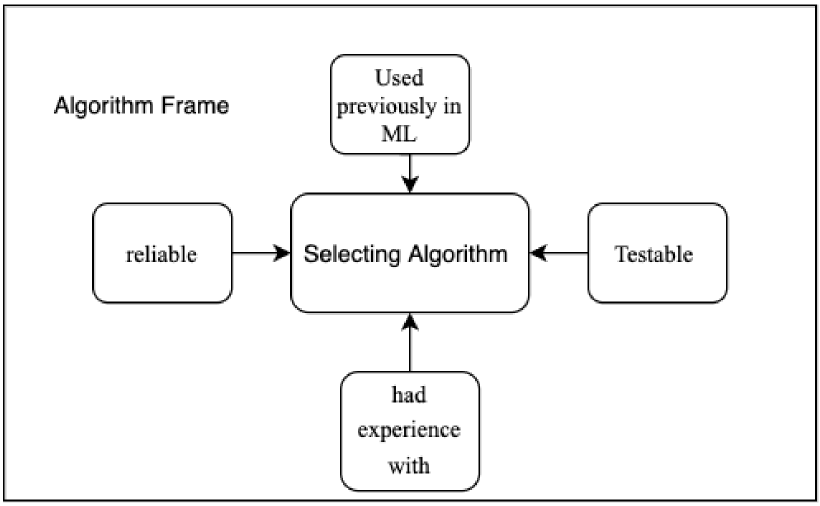

Figure3: algorithms frame

Figure 4 presents the initial goal of the model and how participants view bias. Figure 5 on the other hand, presents the final version of how participants interpret their results. These figures present different types of bias, and our argumentation in this comparison between these frames is that developers might not know the difference. This is because while working on the selection of the dataset and the algorithms, the questions that address the project is shifting when interacting with different process of cleaning data and testing algorithms. Seeing that developers often focus on Statistical, sampling, coverage bias (because they are using frames that address datasets and algorithms), at the expense of frames which address Epistemological bias.

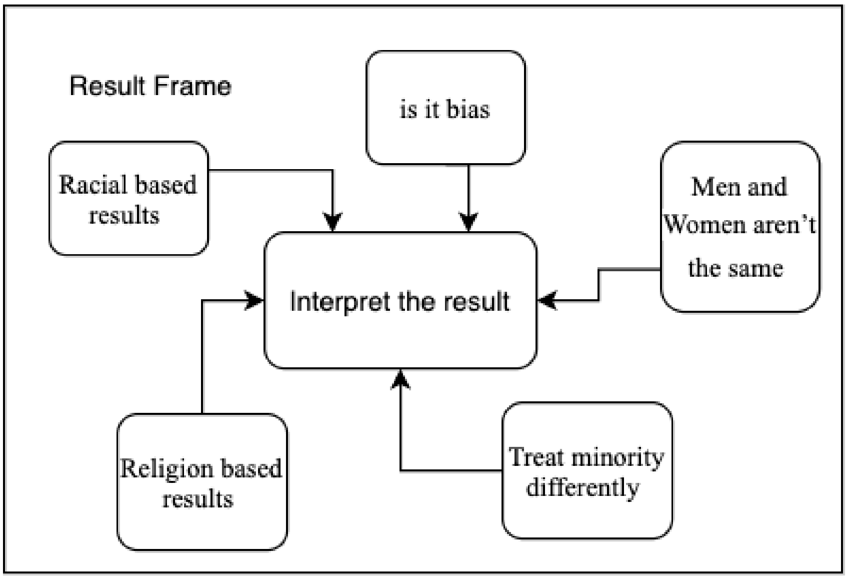

Figure 4: the initial result frame

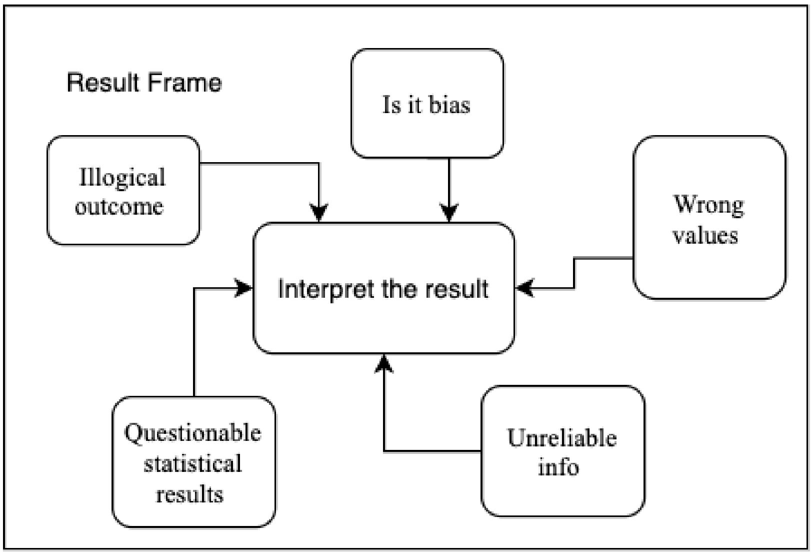

Figure 5: the final result frame 2

CONCLUSION

Conflict

We identified three frames (dataset frame, algorithm frame, and the resulting frame). When it comes to identifying bias in these frames, we believe that the data and the algorithms are usually considered by the developer, but it is not clear who would be responsible for the bias in the interpretation frame. Our results suggest that people change the initial version of the 'result frame' in the light of available datasets and algorithms without realising this. And this explains why we have two subframes for interpretation, the first one presents the 'expected result' at the beginning of the project which aims to present the ethical bias, and the second, a final result frame presents the actual outcome. This situation occurs when for example, the data need to be cleaned, adjusted and modified, resulting in differences between the original and prepared versions of the datasets. Moreover, the algorithm may need some data for training and some for testing, so this requires partitioning of the dataset. In other words, we end up with different conflicting interpretation frames. So, there is potential to introduce epistemological bias because the project changes slowly from project X to project Y (in that the original interpretation frame may have morphed into one that can be addressed using the revised dataset). Instead of project X, they have Project Y which has a nicely balanced dataset that executes well with the algorithm, but where the dataset might not reflect the real-world population, and the results did not address project X's problem.

These Frames presented three important components of any AI system (data, algorithms, developer), although. In figures (2-5) they seem to be different processes, they compete and complete each other in order to create a system. And when we say compete, we mean when you change the other components to fit one of them. For instance, when changing the original data to fit the algorithm, in this case, the algorithm has an important rule in the process. Oppositely, when the data is the critical one, every process to create the system will change according to the dataset.

Implications

This project shows how the developers of AI / ML might not take a narrow perspective on 'bias' (as a statistical problem rather than a social or ethical problem). At the end of the project, most of the participants' recommendations were based on technical implementation and awareness. This is not because they were unaware of these wider concerns but because the requirements relating to the management of data and the implementation of algorithms might narrow their focus into technical challenges. Consequently, bias outcomes can be produced unconsciously because developers are simply not attending to these broader concerns. There is a responsibility to think about bias, but it is not clear where in a work system that responsibility lays. We might think it is the responsibility of the programmer to consider bias when they build the program, but this study suggests that this is quite difficult for the programmer to think about social bias because they are too busy thinking about algorithms and data. The work system does not identify someone responsible to think about other types of bias. However, we can't always blame the programmer alone, we believe that someone else should be responsible for the result frame interpretation within the work system and find a way to link all these frames, so the bias in the result frame should be checked independently of the programmers and the programmers should be made to change what they were doing to minimize that bias. Therefore, creating accurate and effective model is important but so is ensuring that all races/ethnicities and socioeconomic levels are adequately represented in the data model (O'Neil 2016). Methods to debias machine learning algorithms are under development (Gianfrancesco et.al 2018) as are improvements in techniques to enhance fairness and reduce indirect prejudices that result from algorithm predictions (Kamishima et.al 2012).

REFERENCE:

Anjomshoae, S., Najjar, A., Calvaresi, D., & Främling, K. Explainable agents and robots: Results from a systematic.

Baber, C., McCormick, E. and Apperly, I., (2020), A framework for Explainable AI, *Contemporary Ergonomics 2020*

Baber, C., McCormick, E. and Apperly, I., (2021), A human-centered process model for Explainable AI, *Naturalistic Decision Making, 2021*

Borgo, R., Cashmore, M., & Magazzeni, D. (2018). Towards providing explanations for AI planner decisions. arXiv preprint arXiv:1810.06338.

High-Level Expert Group on ArtificialIntelligence (2019) 'High-Level Expert Group on Artificial Intelligence Set Up By the European Commission Ethics Guidelines for Trustworthy Ai', European Commission. Available at: https://ec.europa.eu/digital.

Hoffman, R. et al. (2018) Explaining Explanation, Part 4: A Deep Dive on Deep Nets, IEEE Intelligent Systems, 33(3), pp. 87–95.

Gianfrancesco, M. A., Tamang, S., Yazdany, J., & Schmajuk, G. (2018). Potential biases in machine learning algorithms using electronic health record data. JAMA internal medicine, 178(11), 1544-1547.

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012, September). Fairness-aware classifier with prejudice remover regularizer. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 35-50). Springer, Berlin, Heidelberg.

Klein, G. A., Calderwood, R., & Macgregor, D. (1989). Critical decision method for eliciting knowledge. IEEE Transactions on systems, man, and cybernetics, 19(3), 462-472.

Klein, G., Moon, B. and Hoffman, R.R., (2006 a). Making sense of sensemaking 1: Alternative perspectives. IEEE intelligent systems, (4), pp.70-73.

Klein, G., Moon, B. and Hoffman, R.R., (2006 b). Making sense of sensemaking 2: A macrocognitive model. IEEE Intelligent systems, 21(5), pp.88-92.

Hoffman, R. R. (Ed.). (2007). Expertise out of context: Proceedings of the sixth international conference on naturalistic decision making. Psychology Press.

Minsky, M. (1975). A Framework for Representing Knowledge, Reprinted in The Psychology of Computer Vision, P. Winston.

O'neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown.

Thomas, J. B., Clark, S. M., & Gioia, D. A. (1993). Strategic sensemaking and organizational performance: Linkages among scanning, interpretation, action, and outcomes. Academy of Management journal, 36(2), 239-270.

Thordsen, M. L. (1991, September). A comparison of two tools for cognitive task analysis: Concept mapping and the critical decision method. In Proceedings of the Human Factors Society Annual Meeting (Vol. 35, No. 5, pp. 283-285). Sage CA: Los Angeles, CA: SAGE Publications.