**Chapter 3 Part 1: News Rank, Recommendation and Targeted Ad Algorithms**

In Chapter 2, we suggested that there are problems of disempowerment due to the lack of understanding and control users have over ML algorithms used in social media platforms, mostly in news ranking and recommendation systems. To proceed, first we will refute an approach that is often taken by critics that platforms should simply make "fairer" algorithms. The two phenomenon that make this approach unproductive are that fairness is not a well-defined term, and comparing fairness of algorithms with that of our ideal does not hold, since human judgement, which is the incumbent judge in making decisions about how the platform should look to users, are infamously biased in all areas of life, making the comparison to ideals an unrealistic one. However, we cannot use this critique to deem the problem impossible and settle with the status quo. We will argue that even if making uncontested progress in ML-judgement is difficult, we can still create platforms that are more explainable so that users can be aware of the considerations that go into the decision-making process, and even have a way of influencing the algorithms by changing their history.

First of all, in their paper *Inherent Trade-Offs in the Fair Determination of Risk Scores,* Kleinberg et al show that the numerous definitions of "fairness" used while discussing the pros and cons of ML algorithms are mostly incompatible with one another, except for rare specialized cases (21). This means that even if developers make a conscious attempt to fulfill a fairness condition while implementing an algorithm, they are bound to be criticized by others who hold another definition of fairness. Of course, one could argue that the paper might have chosen one or more of the notions of fairness to be unreasonable ones, which makes the finding ineffective, even if accurate. However, the authors suggest and cite cases in which all three of the properties have been used as significant in judging an AI system's fairness. Furthermore, when we consider the three properties, we can see that they go hand in hand with our intuitive understanding of fairness and don't seem to contradict one another from the outset. The three properties, in short, say: (1) Scores (outputs) of the algorithm should have the same meaning regardless group, (2) The "negative class" should not be targeted in a way that is unwarranted by the properties of the class, (3) The "positive class" should not be awarded in a way that is unwarranted by the properties of the class.

Although the paper is written mainly for risk assessment systems, we can extrapolate what these "classes" would correspond to in news ranking or targeted ads. One such example from Chapter 2 is Tufekci's argument that an individual with onset of a mania period might be targeted because he is more likely to make impulsive purchases. Since she finds this unfair, we can say in this situation that this individual would be in the "negative class". Similarly, we can make such groupings for individuals with certain political affiliations, racial groups, etc. Therefore, their conclusion that these properties of a fair result cannot be simultaneously guaranteed is a powerful conclusion. Indeed, in social discourse, we generally take an agreement on definitions to be a necessary, if not sufficient, requirement in having a productive discussion. What the authors demonstrate here is that we do not yet have a common ground understanding of what fairness is, and in fact, we have definitions that yield incompatible results.

Discussing issues of fairness in a theoretical framework might require that we compare AI decision-making to an ideal. However, from a design framework, all that we have to work with is the in-practice status quo, which in the case of decision making, is human beings. Therefore, we must question how reliable humans are as judges when making claims about AI judgment. In *Extraneous Factors in Judicial Decisions,* Danziger et al show statistically significant differences in ruling depending on whether a judge took a food break before a hearing or not. In the paper, they extrapolate that numerous extraneous factors that should be irrelevant to the case make significant changes in courts of law, where the cost of being impartial is probably much higher than in a social media recommendation algorithm. Similarly, the National Bureau of Economic Research shows evidence that asylum judges, loan officers and baseball umpires consistently fall into "gambler's fallacy", which is the misconception that an outcome of a number of independent trials that has been produced few times is more likely to be produced in the next trial (25). Finally, numerous studies tell us that in many fields, AI decision making is more unbiased than human decision making (26). Therefore, from a design point of view, the argument for ideal algorithms may be too high of a bar, when the status quo is not that ideal to begin with.

With these difficulties in mind, designing a platform with a "fairer" algorithm cannot go much farther than a vaguely defined ideal. However, what is in our capacity today is to give information to the user about what considerations were taken into account for a specific decision to be made. In other words, we can challenge the supposed "Blackbox" nature of the algorithms. In their paper titled *Accountability of AI Under the Law: The Role of Explanation,* Finale Doshi-Velez et al tell usthat we could reasonably expect as much transparency in decision making of algorithms as we do from humans (22). In this work, the authors ask how we could continue to take advantage of the power of AI systems while still holding them accountable, and they respond: "explanation".

Taking the precedents of legal perspectives as what explanations we accept as valid in human judgment, they suggest that the same form of explanation can be expected of AI algorithms as well. The two methods they employ are "local explanation" and "constructing counterfactual scenarios". To understand these tools, first we must recognize that explanation and transparency are not equivalent. For example, a social media platform could give out their source code, but this would not necessarily translate to an explanation interpretable by humans without challenging calculations. However, the authors suggest that we do not need transparency in order to get explanations for particular decisions. With this in mind, giving local explanation is the ability of an algorithm to respond to questions such as: "What were the main factors in a decision?" and" Why did two similar-looking cases get different decisions?".

A local explanation is an "explanation for a specific decision, rather than an explanation of the system’s behavior overall". Even though algorithms look for an abundance of parameters and give them weights through non-linear (and therefore difficult to calculate) functions, the study shows that the algorithms could be questioned about particular decisions in these human-friendly ways. Similarly, constructing counterfactual scenarios is the algorithm's ability to respond to the question: "Would changing a certain factor have changed the decision?" By simulating the algorithm with an altered particular parameter, it could tell us whether this parameter was a tipping-factor in the judgement. Through these two tools, the authors suggest that "demanding explanation from AI systems in such cases is not so onerous that we should ask of our AI systems what we ask of humans".

Without much difficulty, we can imagine how designing a platform with such explanations regarding its news ranking or suggestions algorithm can help with the problems of disempowerment we described in Chapter 2. First of all, we explained that users can get inadvertently pigeonholed into a category which they are not happy with. For example, we can consider the videos Tufekci watches on YouTube for her research, and how she finds that she finds more extreme versions of the same category of videos in her "Up Next". With an algorithm that allows for explanations, Tufekci could ask whether a particular video would be in the Up Next if she had not seen another particular video at a past instance. Even as it is today, YouTube allows its users to delete any video from their watch history, so that the recommendations no longer take those data points into consideration; in a way, YouTube can "forget' that the user watched the video. In a platform with explanations, Tufekci could identify the videos that have pigeonholed her into a certain category and remove them from her watch history. Similarly, Mark Zuckerberg recently claimed that they are building a system that will allow the users to delete any portion of their profiles, in a similar method to YouTube, such that the recommendation algorithms "forget" about that past behavior (23).

At the same time, the explanation allows the users to have they type of first-order-input that we suggested is lacking in current systems. By tweaking their user history to determine which behavior they would like to affect the recommendation algorithms and news ranks, users get to make explicit decisions about their own experience of the platform, thus being empowered in a way that they currently are not. Thus, without having to give users physical control over the functioning of the algorithms, which might be a technical burden and might poorly affect the efficiency of the algorithms, the platform could still grant the users the ability to change the system's behavior in predictable and useful ways.