**Fairness – Expert**

**Module 1: Philosophical concepts of fairness (1 lessons):**

This module will expand on the intermediate course on fairness. It will further link philosophical concepts of fairness to technical models that can be captured algorithmically.

1. Linking philosophical concept of fairness (e.g., procedural vs substantive) to different models of fairness;

The machine learning community has become increasingly concerned with the potential for bias and discrimination in predictive models. This has motivated a growing line of work on what it means for a classification procedure to be “fair.” However, ensuring fairness is not equivalent to ensuring a lack of bias and non-discrimination.

Although most ethical AI regulations and guidance documents direct developers to focus on fairness, scholars have long debated what the definition of fairness is. In fact, there is a rich philosophical literature aimed at understanding and defining fairness. This literature began with the aim f understanding fairness from through a lens of identifying the characteristics of a fair society or fair political system. Nevertheless, the insights from these discussions illuminate the concepts of fairness in machine learning as well. In what follows we provide an overview of this discussion, the most persuasive conceptualizations of fairness including their conceptual advantages and disadvantages.

* 1. Procedural and substantive fairness

When we think about what it means to say that something is fair, we need to make an initial distinction. Conceptualizations of fairness can often be divided into procedural or substantive fairness. That is, some theories focus on establishing fair procedures whereas other focus on fair outcomes. Let’s examine these two categories in turn:

Procedural: Many accounts of fairness define the concept in terms of procedures. According to these theories, fairness is achieved when the procedures to determine an outcome are fair. For example, a coin-toss is considered a fair procedure to determine whether a sports team receives or maintains possession of the ball at the beginning of the game or the beginning of overtime. The reason that it is considered fair is that both sides have an equal chance of calling the side of the coin that lands facing up. It is important to note that “fair” procedures do not necessarily have to ensure equality of opportunity as a coin-toss does, so long as the relevant parties agree that the procedure is fair. Another example bears this out. Voting in a democracy is considered to be a fair, if not the fairest, way of determining a country’s political representatives. However, we know that not everyone always has an equal chance to vote: people who need to work on election days do not participate equally in an election. Further, someone may live in a state or area of the country that is heavily dominated by a political party with whom they staunchly disagree. Members of this party are almost always elected reflecting an inequality of opportunity for members of other parties to win. Nevertheless, this is an example of procedural fairness.

Substantive: For many philosophers, fairness refers to outcomes rather than procedures. Imagine that several of us are at a birthday party. At the end of the party, there is a significant amount of the birthday cake leftover. We agree to divide the cake so that each person takes something home. We decide that we will divide the cake so that each person takes home an equal share of the cake. However, unbeknownst to the other party-goers, I have attend two other birthday parties that same week. I was in the same situation and went home with several pieces of delicious cake, some of which I have already eaten and some of which are still waiting for me at home. Is it fair to divide the current cake equally, when I already have so much at home? Some would argue that this allocation violates substantive fairness. The outcome of the procedure does not produce substantive fairness for all parties because I have so much more than the others.

But even in such cases, we cannot conclude that the procedure has nothing to do with the fairness of the outcomes. In general, we can ask of a procedure whether it treats the people to whom it is applied fairly, for example by giving them adequate opportunities to advance their claims, not requiring them to provide personal information that they find humiliating to reveal, and so forth. Studies by social psychologists have shown that in many cases people care more about being treated fairly by the institutions they have to deal with than about how they fare when the procedure’s final result is known (Lind and Tyler 1988).

Distributive Justice Fairness:

The philosopher John Rawls is the most prominent defender of distributive justice as fairness. Rawls argues in favor of a just or fair distribution of primary goods and opportunities. Primary goods are those things that “someone wants whatever else they want”. By this he means that certain rights and liberties must be ensured without any tradeoffs allowed. For example, the right to freedom of expression is a primary good that anyone wants (or probably wants). If this is correct, then it has to be protected to ensure a fairness and justice.

However, we know that the protection of rights and liberties alone can still lead to unfair outcomes. Not everyone has equal access to exercise their rights owing to socio-economic status, poor health, or systematic prejudices. How do we ensure fairness of opportunities and not just fairness of rights and liberties? In an attempt to find a solution to this issues, Rawls argues that any inequalities in the distribution of opportunities must favor the least well-off. This is a complicate notion, so let’s spend some time to make sure we understand it. Suppose someone makes the argument that the fairest distribution of money in an society is egalitarian. This means wealth is redistributed so that each person has substantively the same amount. Another person counters this system by claiming that allowing for a free market provides the possibility for some individuals to improve their economic situation by working hard. And, they argue, both hard work and an opportunity for greater wealth are two things that people need in their lives to feel satisfied. Under Rawls’s theory someone could say that allowing for inequalities in the distribution of wealth is still fair as long as the opportunity to amass more wealth benefits the least well-off in society.

Reading: <https://journal.viterbo.edu/index.php/at/article/view/342/136>

Power and Fairness:

Many philosophers believed that Rawls’s theory of justice as fairness privileged white, male, capitalist thinking. Feminist philosophers in particular argued against Rawls’s theory by focusing on the role of power. They argued convincingly that power relations and power asymmetries must be considered in any discussions of fairness because they lead to oppression. Hierarchies of power and privilege threaten any notion of fairness unless they are identified and remedied.

The conception of power plays an important role in Iris Marion Young’s *Justice and* *the Politics of Difference*. She criticizes both how past theories of justice have conceptualized power as a possession as well as their distributive solution to the injustice of the abuse of power. Young’s deconstruction and analysis of power is significant in that it uncovers how distributive theories further entrench domination and oppression. This analysis of power is also important because Young’s remedy for lessening domination and oppression rests on *em*powering people by establishing a social structure that democratizes decision making. Analyzing power shows us not only how oppression is entrenched in society but also gives us one of the keys for lessening oppression. Along with power, Young’s discussion of the “5 faces of oppression” is helpful not only in diagnosing oppression but also in formulating a remedy for oppressions various instantiations. As such, any discussion of the nature of fairness has to contend with power relations and modes of oppression.

Oppression is the systematic constraints on groups that are not necessarily the result of the intentions of a tyrant. It is in this sense structural. It is not the result of a few peoples’ choices. Its causes are embedded in unquestioned norms, habits, and symbols. It is reproduced by well-meaning people, and, consequently, it is systematically reproduced in economic, political, and cultural institutions. As a result of diagnosing oppression as structural, we can take note that that an oppressed group need not have an oppressor.

Young understands a social group as a collective of persons differentiated from at least one other group by cultural forms, practices, or way of life. Members have an affinity with one another because of a shared way of life. Nevertheless, groups exists only in relation to another group. Thus, group identification arises when different groups interact and differences among groups are identified and experienced.

Young’s 5 faces of oppression are:

Exploitation, marginalization, powerlessness, cultural imperialism, and violence

What is exploitation?

Every commodity’s value is a function of the labor time necessary for its production. Oppression occurs through the transfer of the results of labor of one social group to the benefit of another social group. The capitalist extracts benefits from the workers. Servants enhance the status of the served. Relations of power and inequality arise. The energies of the have-nots are expended to maintain and augment the power of the wealthy.

Women’s oppression is not only the transfer of powers from women to men. Men’s freedom and power is possible because women work for them.

It is not merely distributive, it is structural. Workers live in a system that does not give them many options. Social processes, not just discrete exchanges, bring about a transfer of energies from one group to another that bring about unequal distributions.

What is marginalization?

Marginals are people the system of labor cannot or will not use.

This is perhaps the most dangerous form of oppression.

A whole group of people that cannot engage in social life and thus suffer material deprivation and even extermination.

Material deprivation can be addressed in small ways by welfare. But welfare can lead to further injustices.

Some freedoms are taken away. Being a dependent is being subject to the arbitrary and often demeaning treatment by bureaucrats. They enforce rules or face punitive measures. The experts claim to know what is best for the dependents.

But dependency is not inherently unjust. Feminists show that relations of dependency are a basic human condition.

It can be used as a reason to keep people from equal participation in social contexts

What is powerlessness?

Most people do not regularly participate in policy decision making that affect the conditions of their lives and actions. In this sense, they lack power.

But power is dispersed, so most people have power in relation to others and lack it to decide policies and results.

The powerless lack power even in this mediated sense.

They must take orders and can rarely give them.

The powerless don’t have the opportunity to develop and exercise skill; they have little or no work autonomy, exercise little creativity and judgment, have no technical expertise or authority, express themselves awkwardly, do not command respect.

This is mostly seen in non-professionals.

Professionals have an education, and develop their skills over time. They have day to day work autonomy. The privileges extend to a whole way of life.

White men are often treated with respect and women and black men without it, until the professionalism is uncovered.

What is cultural imperialism?

The previous three forms of oppression occur due to the division of labor.

Cultural imperialism is when the dominant meanings of a society render the particular perspective of one’s own group invisible at the same time as they stereotype one’s group and mark it out as the other.

It is the universalization of the dominant group’s experience and culture, and its establishment as the norm.

The dominant group projects their experience as that of humanity itself.

Difference becomes inferiority.

They are marked as Other.

They are marked as stereotypes and made invisible.

They are stamped with an essence. Stereotypes confine them to a nature which is somehow attached to a body, which cannot then be easily denied.

White males are individuals. Homosexuals are promiscuous, women are good with kids, Indians are alcoholics.

The oppressed are defined by the outside. They must internalize these stereotypes as they must react to them in one way or another.

The status of otherness does create a different cultural life. So it is self-perpetuating in a way.

The oppressed group has little influence, but the dominant culture imposes on the oppressed group its culture and interpretation of social life.

What is violence?

Violence is systematic when it is directed at members of a group simply because they are members of that group.

Members know they are liable to such violence even if they are direct victims of it.

Living under such a threat deprives people of dignity and freedom.

The possibility of violence occurs to members of the dominant group even if they don’t act on it, and vice versa.

Perpetrators often receive a light or no punishment

Sum up: Any true account of fairness must take these deeply rooted causes of structural unfairness into consideration. This discussion is not only theoretical. If police are creating an algorithm to help identify children who need safeguarding, developers must ensure that characteristics tof oppression that have become embedded in a person’s life and activities do not become perpetuated by the algorithm. We will cover more concrete examples later in the course.

Egalitarianism:

Finally, many scholars defend the position that fairness is when each person has the same amount of whatever it is that needs to be distributed fairly. Thus suppose we have limited supplies of a drug that can treat malaria, and a number of patients displaying symptoms of the disease, but lacking specialised medical knowledge we cannot tell whether one person’s condition is more serious than another’s; then by sharing out the drug equally, we can guarantee that each person at least receives the highest fraction of what they really need.

b. Explain the different formulations of fairness (e.g., calibration, equal error rates) and their ability to satisfy them in models;

Reading: <https://chrispiech.github.io/probabilityForComputerScientists/en/examples/fairness/>

<https://feministai.pubpub.org/pub/what-is-fairness-/release/1>

c. Understanding the impossibility to include all concepts of fairness in the same algorithmic model;

Given the variety of definitions of fairness as well as the importance to pay attention to structural unfairness, plus the challenge, if the not the impossibility, of formally modelling different conceptualizations of fairness, we can conclude that no single model can be considered “fair” full stop. We need to understand that fairness is context dependent, that unfairness is structurally embedded in datasets and that, consequently, there is no arrival point called fairness at which we can rest easy that we have created an ethical tool. Data scientists and ethics advisors must continually evaluate and question whether their models are fair by asking what does fairness mean in a particular context and what aspects of structural unfairness might be lurking in data. Achieving fairness in AI is not a destination but a continuous process.

d. Combine philosophical definitions of fairness with domain-specific notions/expectations of fairness (e.g. AI fairness in predictive policing vs AI fairness in job recruitment or AI fairness in financial sector).

In the policing or justice domain, fairness might mean ensuring that irrelevant characteristics of individuals do not inform a model’s predictions. For example, the COMPAS algorithm used in the US court system was designed to predict recidivism rates of criminal offenders. Pro Publica discovered that it based these predictions on non-criminal characteristics such as whether offenders were African-American or based on where they lived. This is a sobering example of how reporting bias and structural racism can be further entrenched in these domains.

In the health sector, fairness might mean providing medicine or care according to the level of a patient’s need. If one patient needs a scarce medicine more than another, for example to save his/her life, then perhaps it is fairest to provide this person the medicine. Can an algorithm used in the health sector be trained to be fair in this way?

In the education domain, fairness might mean providing pupils with outcomes based on what they deserve. If AI models are going to be used to mark and evaluate pupils’ and students’ work, then it might be fairest to ensure that each individual’s work is evaluated and marked accurately.

What we see from these examples, is that “fairness” takes on different meaning within different domains. Adding the lack of consensus among philosophers about what the concept means to the changing meaning of the term in varying contexts makes it difficult to ensure that artificially intelligent predictions and output will in fact be fair.

This means that it is essential for AI companies to include social science and humanities experts who understand the often hidden characteristics of societal domains and human interaction. This kind of interdisciplinary collaboration is essential to provide the opportunity for a deeper understanding of what fairness might mean and how unfairness can appear in varying contexts.

Reading: COMPAS example: <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>

Pre-Reading:

<https://www.scu.edu/ethics/ethics-resources/ethical-decision-making/justice-and-fairness/>

<https://www.psychologytoday.com/intl/blog/am-i-right/201205/its-not-fair-what-is-fairness>

<https://www.philosophytalk.org/blog/fairness-fixation>

Reading: Added to modules above.

Reflection Worksheet

Q&A:

1. What is Rawls’s definition of fairness? How could his theory help motivate creating a fair AI tool?
2. What are some objections to Rawls’s theory of fairness? Why are these objections important in our society permeated by AI tools?
3. What can fairness mean in differing contexts such as policing, health sector or education?

Multiple Choice Questions:

1. Substantive fairness means...
   1. The rules are fair
   2. The outcome is fair
   3. Fairness doesn’t matter
2. Iris Marion Young’t theory focused on the importance of understanding the nature of...
   1. Fairness
   2. Transparency
   3. Oppression
3. Can different conceptualizations of fairness be included in a single algorithmic model?
   1. No
   2. Yes
   3. Maybe
4. How do we achieve algorithmic fairness?
   1. Embed one of the definitions of fairness into the tool
   2. Don’t worry about it because it cannot be achieved
   3. Decide on which conceptualization of fairness best fits the use case and continually revisit the model and its use to see if it is fair.

Task:

Make notes to give a five-minute talk explaining to police why an AI tool you have developed for them is fair. Describe what the tool is for and how you have attended to and implemented fairness considerations into its design and use.

Task:

Re-read the COMPAS case study and answer these questions:

How could this have been avoided?

Using the principles learned in this topic, evaluate why, or why not, the COMPAS system is unfair.

Train the trainer:

For each theory on fairness, compile its advantages and disadvantages into a table. Use this table to teach the content. Ask the students to engage in a discussion of whether perfect fairness in a society is ever achievable?

Trainers should keep up to date with developments in the EU AI Act regarding the notion of fairness. What is the definition of fairness? How is it recommended to be implemented in AI? Who is accountable for a lack of fairness according to this regulation?

Inform yourself also about the psychology of fairness.

<https://www.psychologytoday.com/us/blog/the-mindful-self-express/201408/the-neuroscience-fairness-and-injustice>