**RDS #2: Fairness and Friends**

# **Cover:**

Data, Responsibly Comics, Volume 2: Fairness and Friends

By: Falaah Arif Khan, Eleni Manis and Julia Stoyanovich

(Line art portrait of the statue of the Woman of Justice, based on Themis, the Greek goddess of divine justice)

# **Terms of Use:**

All the panels in this comic book are licensed [CC BY-NC-ND 4.0](https://creativecommons.org/licenses/by-nc-nd/4.0/). Please refer to the license page for details on how you can use this artwork.

**TL;DR**: Feel free to use panels/groups of panels in your presentations/articles, as long as you

1.Provide the proper citation

2.Do not make modifications to the individual panels themselves

Cite as:

Falaah Arif Khan, Eleni Manis and Julia Stoyanovich. “Fairness and Friends”. *Data, Responsibly Comics*, Volume 2 (2021)<https://dataresponsibly.github.io/comics/vol2/fairness_en.pdf>

Contact:

Please direct any queries about using elements from this comic to themachinelearnist@gmail.com and cc stoyanovich@nyu.edu

## **Tenets of FairML**

Hello there! Ethics and Fairness has been all the rage of the AI news cycles recently. You must be wondering, what are all the pundits talking about?

(Picture the apostles with their wise words from ‘The Ten Commandments’)

Brace yourself for their holiness...Here are \*the Tenets of Fair-ML\*:

(A hooded figure bows down to an African woman apostle, who points to a holy text)

1. Be clear to follow the tenets in this guide

(Satya Nadella, dressed as an apostle, looks knowingly at the reader, and points to his holy text)

1. Be clear that there is no \***one\*** correct notion of fairness, and yet feel free to propose blanket software solutions for all datasets and applications

(The hooded figure, bowing down to the apostle has written across its body)

1. Be clear that ethics research is important insofar as it does not shed any bad light on the company and its products[[1]](#footnote-1) *#IstandwithTimnit*

(Yann LeCun, dressed as in golden robes, holds a holy decree in one hand, while pointing at text that reads)

1. Be clear that ML systems are biased when data is biased. To get an outcome that looks fair, simply train the exact same system on de-biased data

(Sundar Pichai, donned in golden robes, holds a mast in one hand, with both hands outstretched, while the text below him reads:)

1. Be clear that expertise in building unethical AI is a market advantage and can be launched as Ethics-as-a-Service.

(The protagonist walks in and signs the hand signal for ‘Time Out”)

Time out. Welcome to the Fair-ML Club.

(The protagonist bows her head, lowers her glasses to her nose and stares deeply at the reader, knowingly)

There’s only one tenet of the Fair-ML, and it’s that there are no tenets of Fair-ML.

Fairness is \*not\* a technical or statistical concept and there can never be a tool or software that can fully ‘de-bias’ your data or make your model ‘fair’. Fairness is an ethical concept, and a contested one at that. At best, we can select some ideal of what it means to be ‘fair’ and then make progress toward satisfying it in our particular setting.

Let’s back up further, shall we. What are we even trying to make ‘fair’ ? What are algorithms and when are they biased?

## **What is an Algorithm?**

Let’s throwback to the prehistoric days of early 2020. Remember the hobby that many of us attempted to master - with mixed results - during the pandemic lockdown: baking!

(Our second protagonist, let’s call her Mo, dressed in an apron, points proudly at the platter of bread that she holds in her hand)

The recipe is the algorithm: it lists the ingredients and their proportions, and the steps to take to transform them into a scrumptious loaf.

Akin to how each of us have our own cooking styles, algorithms are of different types.

#### **Rule based:**

(A recipe sheet flows in the back, while a collage of baking attempts hovers in the front)

The algorithm may be fully prescribed. For those of us who like to follow a recipe to the T, it lists exactly which ingredients to use, how much of each to take, how and in what order to combine them, how long to wait, and at what temperature to bake.

We call such algorithms “rule-based”.

(Mo pours milk into a beaker and meticulously measures the quantity being poured. She then pours in vinegar into a different beaker, her eyes hawk-like on the gradation on the beaker. Mo gazes intently at a mixture in a beaker, using her index finger to verify that the amount in the vessel exactly matches up with the gradation mark. After taking the loaf out, Mo measures the height and width of the loaf using a ruler)

If we know the rules well enough to write them down, and if we can always get exactly the same ingredients, then we will bake a delectable loaf every time.

(Mo blows a chef’s kiss into the air, clearly happy with the result of her baking endeavor of the day!)

#### **Data driven:**

But we may not always be so lucky...

We may have only ever eaten delicious sourdough, but may not know the recipe for making it ourselves.

We have an idea of what ingredients go into a loaf and have several data points of experience of what it’s supposed to taste like, and so we go about trying different combinations of ingredients and cooking techniques.

(Mo assembles a bunch of ingredients around her on the kitchen counter and proceeds to pour milk out of a bottle into her bowl, mixing it into the batter with a spoon. She then starts rigorously mixing the ingredients in the bowl. Mo takes the dough out and starts kneading it on the counter. She then takes a rolling pin and starts hitting the dough, trying to soften it into the texture she desires. Happy with her dough, Mo uses a cutter to make buns out of the dough.)

Each time we make a loaf, we ask ourselves: *do we like how the sourdough came out?* If so, we may keep this recipe. Or maybe we’ll try something slightly different, or a lot different and see which result we like better.

(Mo takes a loaf out of the oven. She tastes a bite and pulls a face at the taste. In a different attempt, Mo opens the oven frantically, only to see smoke rising out of a thoroughly burnt loaf. In yet another attempt, the dough doesn’t even reach the oven and we see Mo crying over her sunken labor, while pouring the batter into a bin.)

From this, we can figure out which culinary sorcery produces the yummiest results - closest to what we remember a good loaf tastes like.

This is how “data-driven” algorithms work.

(Mo holds up the perfect loaf of the bread, Lion-King Simba style!)

### **Data**

The recipe is the algorithm, now what about the data?

The \*input\* data is the ingredients and their relative proportions.

(An assortment of ingredients is spread on the counter, including flour, eggs, butter, salt, baking soda, etc. To the left is a variety of bowls and spoons, spatulas and rolling pins.)

Another form of data is the parameter settings of your cooking equipment, such as oven temperature or wait times. These are the knobs you can turn to adjust the recipe.

Then there’s data that describes the \*output\*: that scrumptious sourdough that we remember demolishing and are hoping to bake ourselves.

How much does it weigh? What is its nutritional value? How well-done is the crust and how chewy is the center?

(We zoom into one slice of wholewheat bread - analyzing the texture of the center, evaluating the shade of brown that the crust is and measuring the height and width of the slice. A measurement device displays the nutritional value of the slice at 32 calories)

These are all ‘objectively’ measurable factors.

The final kind of data is our reaction to the output: Does the loaf meet our expectations? Is it tasty? These factors boil down to personal preference and, more often than not, are more important than the numerically quantifiable properties of the output.

(Mo morphs into the food critic from ‘Ratatouille’, first staring ambivalently to the side. As Mo moves her fork, with the bread to her lips, her face displays an expression of deep sentiment)

### **Decisions**

What about decisions?

In the process we described, in the course of execution of the algorithm, we are faced with several decisions. Does the dough look good enough to put into the oven? Has the loaf risen enough and shall we take it out of the oven? Are we giving it a thumbs up or a thumbs down?

A more consequential decision is - now that we’ve tried a bunch of recipes, which will we consider a success? Will we say that it’s more important to have an appetizing-looking loaf or one that consistently comes out chewy on the inside and crusty on the outside? Will we decide to always - or never - use some specific ingredients or cooking techniques?

(Ten robot hands hold different slices of bread to Mo’s lips, awaiting her judgement on their cooking)

An even more important decision is - do we think that we’ve tried out enough recipes to pass our experience on to a machine, and trust it to bake on our behalf?

(Ten human hands hold different slices of bread to a robot’s lips, waiting to hear it’s judgement on her cooking)

What about making judgments on our behalf - deciding which loaves turned out well and which didn’t?

Can we trust that same machine - that we just taught how to make sourdough - to bake something different, like baguettes? And who must *pack up their knives and go home* if the baguettes are an utter failure?

(Mo, dressed in her battle gear of a chef’s hand and helmet, poses confidently with two loaves of delectable sourdough. She is supplemented by eight robotic arms that rise out of her back, holding several ingredients and cooking tools - from a whisk, spatula and rolling pin, to a bowl of flour, a stopwatch and a baking tin. Mo starts to resemble the Hindu goddess Durga, with ten arms)

Several \***moral\*** questions around \***agency**\*,\***autonomy**\* and **\*responsibility\*** naturally emerge:

How much autonomy do we give to a machine, a learning algorithm, an AI?

## **What is an ADS?**

So, an Algorithm is a recipe. Then, what is an Automated Decision System (ADS) ? Is it like a self-baking oven? Easy there, Musk-eteer.

(Elon Musk proudly points to a ‘Self-baking oven’, positively beaming with excitement and pride. The self-baking oven glows blue, as it automatically lowes its door and two baking shelves full of perfectly crafted sourdoughs lower themselves out)

We don’t really have a consensus on what an ADS actually is (or isn’t). The law seems to have taken a page out of the ‘Paula Abdul playbook of judging’, going overly lenient and vague in its definition. New York City’s Local Law 49[[2]](#footnote-2) defines an ADS as “*computerized implementations of algorithms, including those derived from machine learning or other data processing or artificial intelligence techniques, which are used to make or assist in making decisions*.”

(A floating-head caricature of Paula Adbul smiles sweetly at the reader)

Using this definition, one could argue that spreadsheets or even internet searches could be ADS, because they are, in fact, computerized and do, in fact, guide decision-making[[3]](#footnote-3).

A precise definition will be crucial for the efficacy of any attempt at regulating these systems. An alternate approach would be to define ADS by extension[[4]](#footnote-4).

*So You Think You’re An ADS?*

Do you:

1. Process data about people
2. Assist - either in combination with human decision making or autonomously - in making consequential decisions that impact people’s lives and livelihoods

Additionally, we would like it if you would

1. Have a specific, stated goal of improving and promoting equality and efficiency. At the very least, you must not hinder equitable access to opportunities
2. Be publicly disclosed and subject to legal audits.

(Three robot contestants of the talent show ‘So You Think You’re An ADS’ strike a pose for the reader. The first is a robot, dressed glamorously in a dress made up of cells from a spreadsheet. The second robot is dressed in a power suit and is performing facial recognition on two job applicants, attempting to evaluate their suitability for the job, based on their facial gestures. The third is dressed up fully as a calculator and striking a pose with its hands crossed in front of it. )

Is a formula in a spreadsheet an ADS? Perhaps - depends on what it’s used for! Is an automated hiring tool? Definitely. But is a calculator an ADS? No!

## **Bias**

With that in mind, let’s look at what we mean by ‘bias’ in an ADS and how it arises[[5]](#footnote-5).

Systematic discrimination by an algorithm is termed ‘bias’. In the context of data-driven systems, biases are ‘harmful’ associations picked up by the algorithm - either from the data itself, or from how the algorithm is designed, or from the objectives that we specified for it, or from how we use it.

#### **Pre-existing - in the data**

Pre-existing biases exist in society and come ‘pre-baked’ into the model as a result of the underlying discriminatory system that the data was generated from. These would be the flavor notes that will seep into your bread if you don’t prioritize the purity/freshness of your ingredients or if you decide to use premixed off-the-shelf batter. A notorious example is the gender and racial stereotypes that language models pick up when trained on data from social media platforms.

(Mo is out shopping in the supermarket, picking up ‘Bread Mix’ from the ‘Ready to Cook’ isle)

These days, most people have no idea about what’s mixed into the synthetic, mass produced, preservation-focused batter they’ve been ingesting and feeding others.

#### **Technical - in the technical system:**

Technical bias are those imperfections that will seep into your bread if you use the wrong equipment. Think about what would happen if your oven temperature is miscalibrated or if your baking equipment is the wrong size.

(Mo seems to have overfilled her baking tray - we see batter overflowing out of its tin as it rises in the oven. While attempting to make cupcakes, we see Mo’s oven seems to be faulty: it’s only heating on one side and so all her cupcakes come out lopsided - risen on the right and completely flat on the right)

In the context of algorithms, these include hardware limitations, incorrect choices of representation and strong modeling assumptions that are not satisfied in the real world.

#### **Emergent - due to decisions:**

The patterns that emerge as a result of your baking comprise ‘emergent’ bias.

What if you become such a maestro at baking that you inadvertently make bread a steady part of your diet! Or make it so often, that you turn everyone around you off the thought of ever eating another slice!

(Oh no! Mo seems to have gotten obsessed with her bread! We see her stuffing an entire bun into her mouth, while holding a tray of several others in her hand)

Or think about how your idea of ‘what bread should taste like’ is shaped by the popularity of products like ‘Wonder Bread’.

(Mo goes to Paris! We see her in a red beret and striped blue tee-shirt, holding a bunch of baguettes close to her. For the American palette, we throwback to an old commercial for ‘Wonderbread’ - with a young, rosy-cheeked boy taking a bite of a slice of ‘Wonderbread’.)

In the algorithmic sense, these are the behavioral patterns that emerge in society as a result of algorithmic decision-making.

## **Intro to fairness - mirror**

Data is a mirror reflection of the world.[[6]](#footnote-6)

All we have is a distorted reflection. Without any knowledge or assumptions about the properties of the mirror or of the world it reflects, we cannot know whether we are looking at a distorted reflection of a perfect world or a perfect reflection of a distorted world or whether these distortions compound.[[7]](#footnote-7)

(Our protagonist looks up at the reflection of the earth. There’s a planar mirror between the earth and the reflection, obscuring the protagonist from looking at the real Earth - all she can see is the reflection. Within the reflection, there are areas that are missing and ‘corrupted’ by imperfections, while other areas are completely distorted due to the curvature of the mirror.)

### **What is Algorithmic Fairness?**

Algorithmic fairness is the corrective lens that we wear to see the world closer to what we want it to look like than what it actually is.

Corrective lenses are tailored to the wearer and, similarly, different individuals judge different fairness ideals to matter, for different reasons. Based on our worldview (beliefs about what the ideal world should look like), we apply corrective measures in the form of different statistical measures of ‘fairness’.

However, wearing these lenses only changes how we view the reflection - It does not and cannot fix distortions in the mirror or fix distortions in the world. Unless such fixes are supplemented by systemic change, we can quickly confuse the world seen through rose-colored glasses with the real world.

(Our protagonist scratches her chin in wonder while looking at the reflection from the mirror. A mystical monocle appears between the reflection and the protagonist, that corrects for how the protagonist sees the reflection - all the distortions and imperfections in the reflection no longer exist!.)

### **How does Algorithmic Decision-Making work?**

Algorithmic decisions are mappings between three ‘spaces’, namely - the construct space (the real world), the observed space (the reflection) and the decision space (the outcomes or allocations).[[8]](#footnote-8)  
Constructs are the ‘desired’ or ‘true’ features that we want to take into account while making our decision. (Einstein points to his forehead and winks at the reader knowingly)

The observations are the features that we are actually able to measure. (An adolescent Einstein is begrudgingly taking a standardized test, frowning in frustration as he writes his answers)

The decision space is the space of outcomes/allocations. (Graduation caps flying the air, while two robotic thumbs signal a thumbs up and a thumbs down)

#### **WYSIWYG vs WAE:**

In a perfect world, where there is neither a distortion in the world nor in the reflection, our constructs and our observations would be the same. In reality, the construct space is unobservable and so we need to make assumptions about its nature and about the mapping from construct to observation to get an allocation that is ‘fair’ (by some specific notion).

Different worldviews give rise to different notions of ‘fairness’.

(Picture three spaces - a globe full of data, from which we can see three individuals. The first, a white male in a dapper three-piece suit. The second, a woman of color, dressed plainly. And the third, a white woman dressed up in a stylish dress. This view is the construct- how we expect the people actually are. The next space is the ‘observed’ space. Let’s see how the observed space looks different based on which worldview we decide to operate under)  
The \*‘What You See Is What You Get’\* worldview assumes that relevant characteristics are correctly captured in the data and that differences among people’s abilities (by some task-specific distance metric) are preserved from the construct space to the observed space. Consider an algorithm that evaluates people’s abilities in some dimension.   
(Under this worldview, the three people look identical in the construct space and in the observed space - race, gender, dress sense and all!)  
A suitable combination of the observations will correctly represent the person’s true ability and this ensures that the decisions are ‘fair’.  
(Decisions are represented by thumbs up or thumbs down given by the algorithm)

On the other hand, the \*‘We are All Equal’\* worldview is based on the idea that differences in people’s abilities should not be attributable to factors outside of their control. In so far that people’s abilities are measured in a manner that is independent of their protected characteristics such as sex and race, the decisions are ‘fair’.   
(Here, the observed space is such that people all look the same color, are all dressed in the same attire and have the same body structure and haircut. We can no longer tell people’s gender or race from their appearance.)

#### **Group vs individual:**

Individual fairness advocates that ‘similar individuals must be treated similarly’[[9]](#footnote-9). Mathematically, if the distance between two people, based on some task-relevant metric, is small, then they should both be allocated the same outcome. The “What You See Is What You Get” worldview tracks individual fairness insofar that it will object to two individuals who are \*truly\* similar in the construct space, to appear to be dissimilar in the observed space.  
However, the converse need not be true – people who are \*truly\* dissimilar in the construct space can end up looking similar in the observed space.  
(Picture Michael Jordan, pulling his signature ‘air’ pose, and pulling a slam dunk from the ‘individual’ side to the ‘group’ side)

Group fairness tries to ensure some notion of parity in outcomes for members of different protected groups. Mathematically, we would aim to equalize some statistical measure - such as positive outcomes, error rates or false positive/false negative rates - across groups.

Think of it as two different coaching styles – Are you the Doug Collins of the ’86-’88 Bulls, designing your entire offense around your most talented player - eager to see him earn his place among the all-time greats?

(Larry Bird, Michael Jordan and Magic Johnson smile during a photoshoot, dressed in their ‘Dream Team’ Olympics uniforms)

Or are you the Phil Jackson of the Bulls, identifying the different strengths of different players and organizing the triangle offense to perfection, thereby taking the Bulls – led by the inimitable Jordan, of course - to their first championship victory.

(Jordan stands next to Phil Jackson, both holding a championship trophy in their hands and beaming with joy)

In principle, individual and group fairness need not be incompatible[[10]](#footnote-10) – you can pull off two ‘threepeat’ Championship wins, while having Jordan win league MVP each year!

#### **Procedural vs outcome:**

A second dichotomy arises from the way in which we arrive at a ‘fair’ decision. Procedural fairness emphasizes that the same process be applied to all individuals, irrespective of the societal factors that might advantage some and disadvantage others in getting a ‘fair’ shot in the selection process.

Outcome fairness, on the other hand, aims to ensure that outcomes (positive or negative) meet some requirement, such as positive outcomes being distributed equally among different groups. This ensures that members from certain groups are not systematically discriminated against, but might come at the cost of procedural fairness - correcting for systemic inequalities might require a different procedure to be applied to candidates from different groups.

(In the culmination of the tension between the Avengers, Iron Man lunges at Captain America, who tries to block Tony’s punch with his shield)

This dichotomy tracks two doctrines from US Anti-Discrimination law - Disparate Treatment and Disparate Impact. Disparate treatment prohibits procedural unfairness - intentional discrimination through the use of different formal procedures or making decisions based explicitly on protected characteristics is illegal. Disparate impact, on the other hand, prohibits unjustified and avoidable disparity in outcomes for people of different protected groups.

This very disagreement almost broke up the mighty Avengers!

(Tony Stark/Iron Man and Steve Rogers/Captain America both stare at each other solemnly)

On one hand, you have Team Stark, who believe in signing the accords and operating under a prescribed mandate and procedure. And then there are those who, like Cap, believe in the fidelity of the outcome, even if it requires preferential treatment.

## **Impossibility Results**

The famous Impossibility Results[[11]](#footnote-11) [[12]](#footnote-12) have decreed that different statistical measures of ‘fairness’ are mutually incompatible.

For example, say you need to reward your hungry hungry helpers. And say your helpers are of different ages and culinary expertise. How do you go about making this allocation?

(Picture Chef Remy, from the 2007 Pixar movie ‘Ratatouille’ and his buddies. Remy is the executive chef, while two of his aproned friends are the sous chefs. Six chubby novice line chefs close out the group of ‘hungry helpers’)

If you decide that the ‘fair’ way to do this would be to ensure that you will split the pie into three equal parts - one for each level of culinary expertise, then each rookie would get less than each executive chef - purely due to that fact that there are more rookies!

(Since you have three sets of chefs - executive, sous and line, you must first split the wheel of cheese into three equal parts. ⅓ goes to Remy, since he is the only executive chef. Each of the sous gets half of their one-third, so ⅙ each. The line chefs are aplenty - 6 to be exact and so they must split their one-third amongst themselves. Each line chef gets only 1/18 of the wheel each - far less than Remy’s ⅓)

If you decide then you must instead give each chef the same amount of food, then it would be impossible to have parity in outcomes for all groups - there would be much more food overall given to the rookie group.

(A different way to split the wheel of cheese is to give everyone an equal piece - we simply split it into 9 pieces, and each chef - irrespective of whether they are an executive chef or line chef or sous chef gets 1/9. Now, if we check for outcomes for different groups, then the line chefs totally have 6/9 of the wheel (since there are 6 line chefs and each got 1/9), the sous have 2/9 of the wheel and the executives only get 1/9 of the total.)

See, how different metrics are inherently incompatible?

And so, before we rush to fulfill some statistical measure of ‘Fairness’, we must ask ourselves what view of the world are we operating under? What do we mean by ‘Fairness’? Is it ‘Non-Discrimination’ (from legal doctrines)? Is it distributive justice (from political philosophy)? Is it equality in the distribution of some commodity/outcome (in the economic sense)?

(Picture a variation of Mount Rushmore, instead with different figureheads for different conceptions of ‘Fairness’, namely - The Statue of Liberty, Justice Ruth Bader Ginsberg, Martin Luther King and the Lady of Justice)

The most popular character in the Fair-ML multiverse seems to be Fairness as ‘Equality of Opportunity’ (EoP). Let’s read through its origin story, shall we?

## 

## **Age of EoP**

It’s the age of EoP!

(Picture an Empire of different EOP villages. Let’s take a stroll through EOP-ville)

1. Libertarian:

(Picture an assortment of different buildings, some majestically built, others raggedly perched on top of marsh land, some more modern art than architecture and other boring vanilla constructions)

The Libertarians are a deeply individualistic people, as evidenced by the unique designs of the settlements in their village. Any holding or opportunity acquired honestly- without theft or cheating - is claimed fairly, even if it means that some end up with significantly lesser claims than others.

1. Formal:

(Nearby is a Colosseum-type dome. Four gladiators strike a pose on a stage in front of the entrance)

The theatre of Formal EOP has its doors open to all talents. Everyone willing and able is welcome, but you come with what you need - no special treatment once you’re in.

1. Rawlsian:

(A wall with minarets separates the Formal EOP villages, from the Substantive EOPs. To the left is a mighty bouncy castle, with a huge trampoline at the entrance. Settlers are seen jumping on it and launching themselves to higher stories of the castle.)

Here stay the Rawlsians, in their bouncy castle of social security. Strategically placed trampolines ensure that no matter people’s starting points in life, individuals with the same talents and willingness to use them have the same opportunities for success

1. Luck-Egalitarians:

(A river flows between the castle and a campsite. There’s a bridge that takes us over to the other side. Here we see a huge communal campfire, with settlers gathered around, playing music and singing merrily. Around them roam unicorns - enjoying the warmth of the fire or drinking the freshwater of the river.)

And here assemble the luck-egalitarians, forsaking all disparities endowed by Mother Luck, even those in talent and effort, in favor of outcomes that reflect only individuals’ responsible choices, and unicorns on rainbows!

**(Deep Dive)**

### **Libertarian:**

* Motto: “You do you!”
* The Libertarian view focuses on the individual’s freedoms and liberties.

(Picture a massive Monopoly board, the pieces that players are moving are their unique settlements that we saw in EOP-ville. ‘Community Chest’ has been replaced with ‘Solo chest’, while other rules such as the ‘Chance’ cards, ‘Free Parking’ and ‘Go to Jail’ remain.)

* Like a game of Monopoly, players are free to capitalize on whatever opportunities they have access to - such as rolling doubles and getting to move twice, or picking up that Chance card that advances you to Boardwalk! - provided they gain such access fair and square - no cheating by rolling biased dices, stealing from the bank or forcing players to trade properties.
* All players are free to decide which property to chase. Whether they actually get the opportunity to buy and develop on that spot is not entirely devoid of chance, but the game does not attempt to correct for it. Instead, the emphasis is on respect for players’ liberty to buy and sell property and their freedom to exercise their individual skills of negotiation and dice-throwing.

(The mascot for the Libertarians is Ayn Rand, dressed as the Monopoly Man, waving her pen at the reader, while leaning against a walking stick, as money showers down onto her)

* From a political philosophy perspective, this is not a form of EOP at all: there’s nothing being equalized. A Libertarian ADS is only concerned about ensuring a very limited notion of Procedural Fairness.

### Formal EOP:

* Motto: “Careers open to talents”
* Formal EOP tries to ensure equality at the point of competition - in any contest, the most qualified person wins.

(Picture different military training drills - cadets jump over hurdles, crawl under barbed wire and try to scale high walls.)

* This is a view that rejects hereditary privilege as the basis for winning positions: being an aristocrat won’t get you the job. Still, formal EOP makes no attempt to correct for inequalities of birth that can lead to a disparity in opportunities to build qualifications

(The mascot for Formal EOP is a muscular Gandhi, dressed as a colonel - his aviators and military cap over his signature dhoti and shawl. Gandhi is seen chastising his officer, while his three monkeys are covering the eyes, mouth and ears of his officers.)

Formal EOP advocates ‘See nothing Irrelevant, Speak nothing Irrelevant, Hear nothing Irrelevant’. Decision makers are taught to ignore inequalities of birth and focus only on relevant qualifications in adjudicating a contest

* In Fair-ML, this has been codified as ‘Fairness through Blindness’, where any protected attributes - those that can identify group membership - are stripped away from the data.
* But there’s more to formal EOP, if we consider its motivation. A test that is more inaccurate for members of a protected class - that badly mismeasures the qualifications of women candidates compared to men, for example - also violates the spirit of Formal EOP, even if the test does not take gender into account.

### (Substantive) Rawls’ Fair EOP:

* Motto: “Equally endowed babies must be given equal life prospects”

(Picture the portrait of ‘God and Adam’, but instead with John Rawls and three different babies. A ‘god-like’ Rawls extends his finger to touch the foreheads of the three babies, peering into their minds, to see their talents and aspirations)

* Rawls’s Fair EOP[[13]](#footnote-13) says all people, regardless of how rich or poor they are born, should have opportunities to develop their talent, so that at the point of competition, individuals with equal ‘native’ talents and equal motivation have equal prospects of success.
* Rawls wants to ensure that your privileged birth doesn’t snowball into a lifetime of privileges that allow you to outcompete kids whose disadvantage at birth has led to compounded dis-privilege.
* Rawls’s view is targeted to opportunities to develop qualifications from childhood onward. But Fair-ML has reinterpreted his view to mean that at the point of a competition, competitors should be measured according to their talents and motivation, in recognition of competitors’ unequal opportunities to develop qualifications.

(The scale of justice is holding two different buildings. The lighter building - the one with more measurably ‘unjust’ has a ladder at its top. Two youths hold onto the ladder and try to scale the building. Another youth stands at the end of the ladder, his arms outstretched, to help anyone who might want to climb to better opportunities)

* Along these lines, Fair-ML formulations of Rawlsian Fair EOP include statistical parity and equality of odds[[14]](#footnote-14). Assuming talents and motivation are equally distributed among subpopulations and that competitions are won on the basis of talents and motivation, each subpopulation should have the same success rate as any other.
* However, \*these measures are not in the spirit of Rawlsian EOP at all\*, because at the point where an ADS is making a decision it is already too late to provide people with opportunities to build qualifications. Such a narrow focus of EOP immediately bypasses any notion of developmental equality of opportunity and goes to equality of opportunity at the point of competition.

### (Substantive) Luck Egalitarian EOP

* Motto: “Nothing that you did not choose for yourself should affect your life prospects”
* The Luck Egalitarian says that Rawls doesn’t go far enough in controlling for factors that provide unfair advantage or disadvantage.
* Our outcomes should only be affected by our “choice luck” (responsible choices); no effects of “brute luck” (from having rich parents to getting struck by lightning) should be allowed to stand.

(Ronald Dworkin and John E. Roemer stand as the mascots of the Luck-Egalitarians and pet their majestic unicorns)

* How do we separate one’s circumstances of luck from those of responsible choices is an open question...
* One popular formulation in Fair-ML is that of Roemer’s EOP[[15]](#footnote-15), which measures a person’s effort by their rank in the effort distribution of their type/circumstance, rather than by the absolute level of effort that they expend.[[16]](#footnote-16)
* This dials back on the idea of controlling for all brute luck. Instead, we focus on a few brute luck factors, such as race and sex, that track significant undeserved privilege and disprivilege and affect people’s opportunities to develp qualifications.
* We create brackets based on matters of brute luck and then compare candidates to others in their own brackets.
* This formulation in Fair-ML captures the essence of luck egalitarianism and is appealing because it also meets the decision maker’s objective to find qualified candidates - the ADS considers all of a candidate’s qualifications, not just those that are attributable to native talent/motivation (rawls) or responsible choices (other luck egalitarians)

## **Revisiting Impossibility results**

## The impossibility results in Fair-ML are commonly interpreted to mean that ‘fairness is impossible’. But, if we look at different statistical measures as promoting different conceptions of EOP - Formal vs Substantive, then this incompatibility is wholly unsurprising. We would not expect a world view that only looks at ‘relevant’ qualifications at the point of competition (Formal EOP) to be compatible with one that aims to provide comparable developmental opportunities for individuals and, at the point of competition, seeks to correct for inequalities in candidates' developmental opportunities (Substantive). We can interpret this incompatibility as the difference in philosophical viewpoints and incentives of decision makers. (Picture a fencing match between two armored opponents. One has the word ‘Formal’ in its armor, while the other says ‘Substantive’)

This grounding gives us some much needed guidance in choosing a suitable ‘fairness’ measure for our given context. If we believe that inequalities of birth do not affect a person’s qualifications, then the Formal approach might be sufficient to model a ‘fair’ footrace between a king and a peasant.  
(A stout man, in full regal garb - velvet robes, velvet cape, leather boots, golden crown and sword hanging from belt - tries to catch up to a peasant who is running barefooted.)   
Similarly, Formal EOP, in the form of ‘Blind Auditions’, might be desirable in talent contests that wish to judge contestants solely on their singing chops and not on ‘irrelevant’ traits like gender, race, appearance.

(Two celebrity judges sit with their backs turned to a contestant. The contestant is a non-binary, person of color who is belting out operatic vocals. The female judge seems taken by the performance and is about the press her button, while the other (male) judge is listening to the performance intently, unable to decide whether to turn around or not)

When we have reason to believe that structural inequalities preclude certain demographics from having a fair shot at a competition, we might decide to model a more Substantive conception of EOP.

(Three hurdle jumpers are competing in a race)

In a footrace, if hurdles – in the form of systemic discrimination and inequitable access – abound in the path of certain candidates, we might want to compare the hurdle jumpers with other hurdler jumpers and the smooth-track runners with smooth-track runners. The substantive approach is desirable here in order to ensure that we don’ t end up inflicting a Sisyphean struggle upon certain candidates by overlooking their inequalities of birth.

(A woman of color is pushing a massive boulder up a steep hill)

## **Broader view of Justice**

Our stroll through EOP-ville has shown how Fair-ML has only codified narrow notions of EOP so far. The impossibility results could be interpreted to imply that EOP is not sufficient by itself in the creation of a Just society. This tracks well with broader doctrines of Justice in political philosophy, such as Rawls’ broader theory of Justice[[17]](#footnote-17), which speaks to how society must be structured and how any social inequalities must be distributed in a manner that is ‘’just”.

Rawls Theory of Justice posits the following hierarchical principles:

(Picture a fashion magazine spread of photos of John Rawls putting on a three-piece suit. He begins by putting on a crisp white shirt and poses seductively to the camera while buttoning it up. He then puts on a waistcoat and poses midway through putting on his blazer)

1. [Rights and Liberties] Everyone has the same inalienable right to equal basic liberties

2. a. [Rawlsian Fair EOP] All offices and positions must be open to all under conditions of fair equality of opportunity.

b. [Difference Principle] Any social inequality must be applied in such a manner that they be of the greatest benefit to the least advantaged

In the Rawlsian system, these principles are hierarchically ordered – Fair EOP has less priority than the principle of basic rights and liberties. Akin to how incredibly counterintuitive it would be to put on a blazer, without wearing a shirt first, we cannot go about creating Fair EOP in such a manner that would encroach on a fundamental right or liberty.

## This broader perspective is helpful to see how an ADS that is (statistically) ‘fair’ can go on to infringe on basic rights and liberties, and in effect, be unjust. (A dapper John Rawls, in a navy-blue three-piece suit sits on a stool and poses to the camera)

## Take the example of “fair” hiring of people with disabilities. “Disability” would be treated as a protected class and removed from explicit consideration, but algorithms could still infer disability from other proxy variables. If social media information is used, the ADS could infer disability status—for example, based on membership in certain social groups or on posting about disability-related issues—then a scheme that discriminates on the basis of “inferred” disability would incentivize people against joining such groups and speaking about such topics. (A blind woman is surfing social media on her laptop. The bits rise out of her computer and construct the sign for disability - a person on a wheelchair) Such an ADS could satisfy group fairness through statistical parity and yet be fundamentally unjust: it would violate a commitment to equal basic rights and liberties such as freedom of speech, and freedom of association.

## **Broader view of EOP**

## Another idea from political philosophy that is overlooked in fair-ML is the distinction between equality of developmental opportunities, EOP over a lifetime, and EOP at a decision point (Fair-ML’s focus). It might be worth exploring fairness over the course of a lifetime – Do people have comparable/equally desirable sets of life opportunities available to them? Does developmental EOP compound over the course of a lifetime? Or does a disadvantage of birth snowball into a lifetime of disadvantage? (A young Beth Harmon, from hit television show ‘Queen’s Gambit’ sits across Mr Shaibel and plays chess. An older Beth Harmon looks piercingly at the reader)

Equality of developmental opportunities is about making sure people have the opportunity to hone any talent they wish to, instead of being disadvantaged by circumstances of birth that preclude them from certain opportunities. This is motivated by the idea that what matters from the point of view of justice is people having genuine opportunities to realistically achieve goals (e.g., being a track athlete), not merely formal opportunities to compete for jobs (e.g., to be allowed to compete in a race, even though one has no realistic opportunity to finish competitively).

(We see an ongoing race between (the Queen’s Gambit’s) Beth Harmon and (American track and field star Allyson Felix. Felix has already finished the race and is shooting the victory sign, while Beth struggles to even finish, already crying from exhaustion. Interspersed with this picture are scenes from Felix’s childhood, showing her conditioning her body and practicing sprints)

This is exceedingly difficult to model. Current formulations, using economic models of EOP[[18]](#footnote-18) pose the fairness task as a mapping from circumstance and effort directly to utility. This formulation bypasses any notion of equality of \*opportunity\* to achieve equality of \*outcomes\*. Think about the prospects open to a front office clerk versus those of a waitress at a diner. While their utility at this point in time might be the same, their opportunities for growth (and utility in the future) are vastly different. Utility just isn’t the right metric to model EOP with.

(We see 3 pictures of Tracey Armstrong, President and CEO of Copyright Clearance Center (CCC). First, we see a 21-year-old Tracey as a front office clerk - her actual first position at CCC. Then we picture Tracey as a waitress at a diner in a parallel universe. To the right, we see Tracey, in a power suit, confidently smiling at the reader, having risen to the position of CEO at CCC in 2007.)

## **Open Problems**

#### **Interpretability**

There are limitations to what answers we might get from EOP doctrines and overlooking these can embolden their application in spheres in which theory provides little to no guidance. These doctrines do not give us any direction about \*where\* to apply ‘Fairness’ - in the procedure or at the outcome. The guidance is only about \*how\* a ‘fair’ test should behave.

When applying this test to black box ADS, we run into issues of interpretability and can only infer details about how the test is behaving by looking at which inputs have been fed into the algorithm, or by systematically studying the outcomes for a variety of candidates.

(Throwback to the children’s book ‘The Little Prince’. We see two sheep and a ram, with the words ‘Justice’, ‘Non-Discrimination’ and ‘Equality’ written in them. Below we see a box, with three small holes.)

The ‘fairness’ that you asked for is inside this box.

#### **Brute luck vs choice luck**

Substantive EOP seeks to provide all individuals with realistic opportunities to develop qualifications and hence a realistic shot at competing for a broad range of positions. If we decide that the only way that we can operationalize the Substantive view is to separate qualifications into matters of brute luck/circumstance (to be controlled for) and choice luck/effort (that the individual can be held accountable for), then we must decide how to make this separation! Which characteristics can we hold one accountable for? And which matters are completely out of their control?

(Picture a 11-year-old Harry Potter, looking around nervously, as the Sorting Hat sits on his head and debates which Hogwarts house, he should put Harry into)

Sounds like what we need is a Sorting Hat!

From a practical perspective, it is obvious that we cannot separate out a person’s talents from circumstances of birth. And yet, we have tests like the SAT and GRE, which supposedly gauge intelligence and academic excellence, and are used to make admission decisions.

Just registering for such a test - forget about getting access to study materials - is prohibitively expensive. Such standardized tests do not evaluate native talent, but instead systematically discriminate on the basis of inequalities of birth. Always. (A sullen Severus Snape looks deeply into the reader’s eyes) *#TransWomenareWomen*

#### **Intersectional Fairness**

ADS are broadly used socio-techno-political systems. One area in which social dynamics of power and oppression become especially salient is problems of intersectionality[[19]](#footnote-19). Intersectionality analyzes the overlapping dimensions of protected attributes such as sex, race, class, disability, etc.   
(We see a 2D plot of two Bell- curves intersecting in the middle. To the left of the first plot stands a White male, with a physical disability. Near the center of the plot, where both curves intersect stands an able-bodied Black man. To his left, near the tails of both curves stands a blind Black woman and a wheelchaired Black man)   
An example is the study of facial recognition software on black women[[20]](#footnote-20) (the intersection of race and gender). At times intersectionality is causal in nature[[21]](#footnote-21) – Take the intersection of race and disability. Due to unequal access to healthcare, people with disability are disproportionately black. Modelling intersectional fairness in the face of composite protected classes and to study how different forms of biases interact and compound is a hard open problem[[22]](#footnote-22).

#### **Missing Data**

What do we do about gaps in data? Many demographics are poorly represented in data, due to issues of inequitable access or distrust in the data collection mechanism itself. Data gaps make it hard to view the big picture and manifest as a disparity in model performance (error rates, false positives, false negatives) for under-represented demographics.

(We have an incomplete jigsaw puzzle of Banksy’s ‘Girl with Balloon’. Several pieces are missing, especially the piece with the balloon, and so we aren’t able to decipher what the girl is gesturing towards)

Then there’s the problem of observability. In most ‘fairness’ related tasks we are modelling for ‘risks’ - ‘risk of loan default’, ‘risk of recidivism’, ‘risk of college dropout’. We seldom get to observe whether the person actually goes on to do any of those things.

#### **Exploration vs Exploitation**

This leads us to the tradeoff between exploration and exploitation.

(Picture the multi-armed bandit! Dressed in his checked red shirt and leather waistcoat, red bandana and cowboy hat, this robotic bandit has 4 arms, in each of which he holds a pistol.)

In order to test the efficacy of an ADS we might need to put it into the real world to gather more data. This poses a difficult ethical conundrum- Is it justified to forgo the wellbeing of individuals who will be impacted by the current (perhaps sub-optimal) ADS for the potential future wellbeing of individuals? Are these costs borne disproportionately by a certain demographic? Does this lead to new forms of ‘unfairness’?[[23]](#footnote-23)

## **Bias - dragon**

Before we depart, let us heed an important warning about the very nature of this tale… Bias is a three-headed dragon, each head a formidable opponent in its own right.

(Meet Bias-the three headed dragon. Each of its heads has a name for themselves! From left to right: pre-existing, technical and emergent!)

It’s incredibly difficult to detect bias in data, even more so in the output of a black-box ML algorithm. Or when that model is asked to make predictions on data that is different from what it was trained on, possibly even as a side-effect of that very model’s use.

(Look at this armoured knight, deep into his combat practice - repeatedly delivering blows to a straw dragon hanging from the ceiling. Do you think dragons in the wild are also made of straw?)

This complexity compounds when you think about the \*incentives\* that ADS create. It’s not just some abstract prediction coming out of an algorithm anymore - it’s being used to make a decision in the real world.

..And these decisions determine critical social allocations such as jobs, grades and loan decisions. This creates incentives for people to behave in a way that maximizes their allocation from the ADS. This ‘new’ behavior in turn reflects in the data and affects the subsequent prediction from the algorithm.

Playing in the arena of Fair-ML is not only like facing a three-headed dragon, but then having a new, ever-evolving, dynamically-generated opponent each time. Devise a method to cut one head of pre-existing bias off, and two new heads of emergent biases grow out..

(Look at that! Our knight has gone intrepidly into battle with Bias- the three headed dragon. Our knight heroically pierces the middle head of technical bias with her sword. Oh dear! What is that? The other two heads are breathing fire onto the knight. Let’s hope her armour is fire-proof)

Then, there’s the nature of current scholarship. Codifying fairness in algorithms is a technical fix to a societal problem. Fair-ML has emerged as a specialized sub-field of ML, with only a certain group of researchers taking it upon themselves to slay the dragon and rescue the princess.

For a while we made progress on the technical front, but eventually were taken down by the triple threat of the socio-techno-political nature of bias. Yet, we seem to have blood to spare and so we keep rushing into battle with new metrics and methods..

(Wow! Look at that, our knight has defeated the bias-dragon! Here she stands with her greatspear, on top of the dragon’s head that she hath just slain. The princess from the castle looks down at the knight with adoration and gratitude.)

At the end of the day, the question we really should be asking ourselves is - What do we do about a society that locks up princesses in castles, in the first place?

# About

A computer scientist, artist and philosopher join a zoom room. This happens!

‘Fairness and Friends’ is the second volume of the [Data, Responsibly Comic series](https://dataresponsibly.github.io/comics/). We hope that it will serve as the computer scientist’s guide to political philosophy!

[Falaah](https://falaaharifkhan.github.io/research/) is a scientist/engineer by training and an artist by nature and the creator of [MachineLearnist Comics](https://www.patreon.com/machinelearnist) - a collection of webcomics about the current AI landscape.

[Eleni](https://www.linkedin.com/in/elenimanis) is the Research Director at the [Surveillance Technology Oversight Project](https://www.stopspying.org/). She began her career as an Assistant Professor of Philosophy at Franklin & Marshall College, focused on justice in democracies, and now works at the intersection of her expertise in ethics, democratic justice, and technology policy.

[Julia](http://stoyanovich.org/) is an Assistant Professor of Computer Science and Engineering and of Data Science at NYU and the founding Director of the [Center for Responsible AI @ NYU](http://airesponsibly.com/). She leads the ‘Data, Responsibly’ project, the latest offering of which is the inimitable interdisciplinary course on [Responsible Data Science](https://dataresponsibly.github.io/courses/).

1. <https://www.wired.com/story/timnit-gebru-exit-google-exposes-crisis-in-ai/> [↑](#footnote-ref-1)
2. <https://www1.nyc.gov/assets/buildings/local_laws/ll49of2019.pdf> [↑](#footnote-ref-2)
3. Automated Decision Systems Task Force. New York City Automated Decision Systems Task Force Report. [https://www1.nyc.gov/assets/adstaskforce/ downloads/pdf/ADS-Report-11192019.pdf, 2019.](https://www1.nyc.gov/assets/adstaskforce/) [↑](#footnote-ref-3)
4. [*Julia Stoyanovich, Bill Howe, and H. V. Jagadish. 2020. Responsible data management. Proc. VLDB Endow. 13, 12 (August 2020)*](http://www.vldb.org/pvldb/vol13/p3474-asudeh.pdf) [↑](#footnote-ref-4)
5. [*Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. ACM Trans. Inf. Syst. 14, 3 (July 1996*](https://nissenbaum.tech.cornell.edu/papers/Bias%20in%20Computer%20Systems.pdf)*)* [↑](#footnote-ref-5)
6. [*Julia Stoyanovich, Bill Howe, and H. V. Jagadish. 2020. Responsible data management. Proc. VLDB Endow. 13, 12 (August 2020)*](http://www.vldb.org/pvldb/vol13/p3474-asudeh.pdf) [↑](#footnote-ref-6)
7. Falaah Arif Khan and Julia Stoyanovich. “Mirror, Mirror”. Data, Responsibly Comics, Volume 1 (2020) <https://dataresponsibly.github.io/comics/vol1/mirror_en.pdf> [↑](#footnote-ref-7)
8. [Sorelle A. Friedler and Carlos Scheidegger and Suresh Venkatasubramanian. 2016. On the (im)possibility of fairness](https://arxiv.org/abs/1609.07236) [↑](#footnote-ref-8)
9. [*Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. ITCS 2012*](https://arxiv.org/abs/1104.3913) [↑](#footnote-ref-9)
10. [*Reuben Binns. 2020. On the apparent conflict between individual and group fairness. FAT\* 2020*](https://arxiv.org/pdf/1912.06883.pdf) [↑](#footnote-ref-10)
11. [Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 2017](https://arxiv.org/abs/1703.00056). [↑](#footnote-ref-11)
12. [Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. 2016](https://drops.dagstuhl.de/opus/volltexte/2017/8156/pdf/LIPIcs-ITCS-2017-43.pdf) [↑](#footnote-ref-12)
13. John Rawls. A theory of justice. Harvard University Press, 1971 [↑](#footnote-ref-13)
14. [Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. NeurIPS 2016.](https://papers.nips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html) [↑](#footnote-ref-14)
15. [John. E. Roemer. Equality of opportunity: a progress report. Social Choice and Welfare. 2002](http://econdse.org/wp-content/uploads/2012/02/roemer-eop.pdf) [↑](#footnote-ref-15)
16. [Hoda Heidari, Michele Loi, Krishna P. Gummadi, and Andreas Krause. A Moral Framework for Understanding Fair ML through Economic Models of Equality of Opportunity. FAT\* 2019](https://arxiv.org/abs/1809.03400) [↑](#footnote-ref-16)
17. John Rawls. A theory of justice. Harvard University Press, 1971 [↑](#footnote-ref-17)
18. [Hoda Heidari, Michele Loi, Krishna P. Gummadi, and Andreas Krause. A Moral Framework for Understanding Fair ML through Economic Models of Equality of Opportunity. FAT\* 2019](https://arxiv.org/abs/1809.03400) [↑](#footnote-ref-18)
19. Kimberle Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. University of Chicago Legal Forum (1989). [↑](#footnote-ref-19)
20. [Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. FAT\* 2018](http://gendershades.org/index.html) [↑](#footnote-ref-20)
21. [Yang, Ke & Loftus, Joshua & Stoyanovich, Julia. (2020). Causal intersectionality for fair ranking.](https://arxiv.org/abs/2006.08688) [↑](#footnote-ref-21)
22. [Alexandra Chouldechova and Aaron Roth. 2020. A snapshot of the frontiers of fairness in machine learning. Commun. ACM 63, 5 (May 2020)](https://dl.acm.org/doi/10.1145/3376898) [↑](#footnote-ref-22)
23. [Sarah Bird, Solon Barocas, Kate Crawford, Fernando Diaz and Hanna Wallach. Exploring or exploiting? Social and ethical implications of autonomous experimentation in AI. In Proceedings of Workshop on Fairness, Accountability, and Transparency in Machine Learning. ACM, 2016.](https://www.fatml.org/schedule/2016/presentation/exploring-or-exploiting) [↑](#footnote-ref-23)