DATA, RESPONSIBLY

#2

MachineLearnist COMICS

FAIRNESS & FRIENDS

TERMS OF USE

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

ACCESSIBILITY STATEMENT

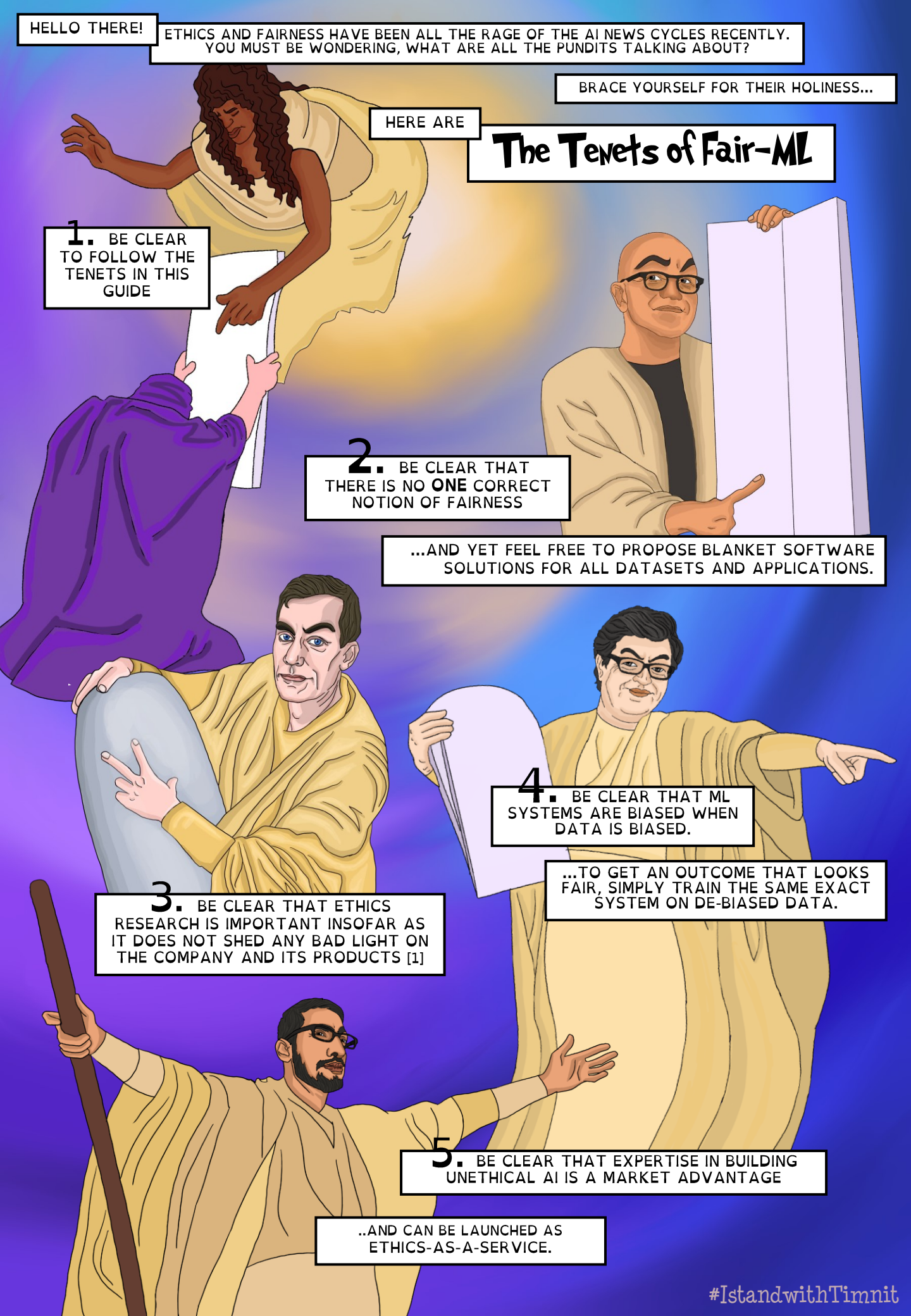
The purpose of scientific publication is the presentation of ideas and dissemination of findings. In the course of our (ongoing) work on creating a comic series about Responsible AI, we have found that relatable cartoons and visual humor are a rich but underappreciated source of clarity and accessibility that enable effective communication to a broad audience. Comic books are a particularly prescient medium for literature reviews and critical surveys, and for bridging insights from different disciplines such as philosophy, law, sociology, and computer science. Given the inherently interdisciplinary nature of machine learning, we see comics and other technical artwork as a promising new medium of scholarship. We hope to demonstrate their utility through our work and to popularize their adoption more broadly in the scientific community.

We care deeply about making our comics as digitally accessible as possible. Towards this end, we have taken the following measures:

1. We've chosen a typeface that was developed specially for dyslexic readers. All of the major text in the comic is in the "Open Dyslexic" font.
2. The comic book is fully alt-texted and can be read entirely using a screen reader. We are also releasing a complete transcript of the comic book, including all of the text and image descriptions.
3. We will be translating the comic into different languages to cater to speakers of languages other than English, as we have done with previous volumes of the Data, Responsibly comic series.

We would like to thank Amy Hurst and Chancey Fleet for guiding us on the Accessibility front.

Please feel free to reach out to us if you have any recommendations on how we can further improve the accessibility of our comics.


HELLO THERE!
ETHICS AND FAIRNESS HAVE BEEN ALL THE RAGE OF THE AI NEWS CYCLES RECENTLY. YOU MUST BE WONDERING, WHAT ARE ALL THE PUNDITS TALKING ABOUT?
BRACE YOURSELF FOR THEIR HOLINESS...
HERE ARE
The Tenets of Fair-ML
1. BE CLEAR TO FOLLOW THE TENETS IN THIS GUIDE
2. 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.
3. BE CLEAR THAT ETHICS RESEARCH IS IMPORTANT INSOFAR AS IT DOES NOT SHED ANY BAD LIGHT ON THE COMPANY AND ITS PRODUCTS [1]
4. BE CLEAR THAT ML SYSTEMS ARE BIASED WHEN DATA IS BIASED.
...TO GET AN OUTCOME THAT LOOKS FAIR, SIMPLY TRAIN THE SAME EXACT SYSTEM ON DE-BIASED DATA.
5. BE CLEAR THAT EXPERTISE IN BUILDING UNETHICAL AI IS A MARKET ADVANTAGE
..AND CAN BE LAUNCHED AS ETHICS-AS-A-SERVICE.
#IstandwithTimnit

TIME OUT. WELCOME TO THE FAIR-ML CLUB.

THERE'S ONLY ONE TENET OF 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 TOWARDS 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?

HERE'S A 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!

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 WE EACH HAVE OUR OWN COOKING STYLES, ALGORITHMS ARE OF DIFFERENT TYPES...

THE ALGORITHM MAY BE FULLY PRESCRIBED.

FOR THOSE OF US WHO LIKE TO FOLLOW A RECIPE TO THE T, IT LISTS EXACTLY WHAT INGREDIENTS TO GET, 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"

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 GREAT LOAF OF SOURDOUGH EVERY TIME!


BUT WE MAY NOT ALWAYS BE SO LUCKY... WE MAY ONLY EVER HAVE 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 THE INGREDIENTS AND COOKING TECHNIQUES.
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.
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.

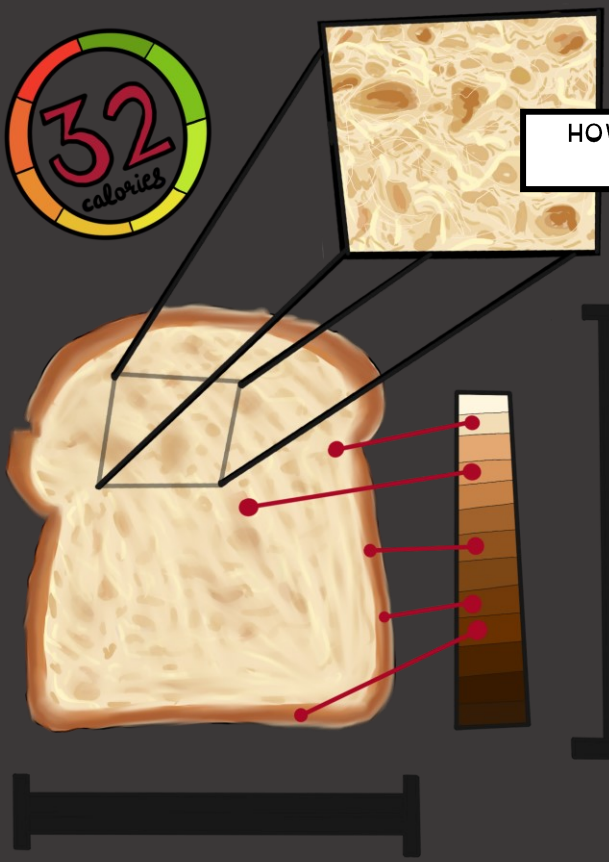
THE RECIPE IS THE ALGORITHM, NOW WHAT ABOUT **THE DATA?**

THE **INPUT** DATA IS THE INGREDIENTS AND THEIR RELATIVE PROPORTIONS.

ANOTHER FORM OF DATA IS THE PARAMETER SETTINGS OF YOUR COOKING EQUIPMENT SUCH AS OVEN TEMPERATURE OR WAIT TIMES.

THEY 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 CHEWY IS THE CENTER?

HOW WELL-DONE IS THE CRUST?

WHAT IS IT'S NUTRITIONAL VALUE?

HOW MUCH DOES IT WEIGH?

THESE ARE ALL 'OBJECTIVELY' MEASURABLE FACTORS.

THE FINAL KIND OF DATA IS OUR **REACTION TO THE OUTPUT**

IS IT TASTY?

DOES THE LOAF MEET OUR EXPECTATIONS?

THESE FACTORS BOIL DOWN TO PERSONAL PREFERENCE AND, MORE OFTEN THAN NOT, ARE MORE IMPORTANT THAN THE NUMERICALLY QUANTIFIABLE PROPERTIES OF THE OUTPUT.

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?

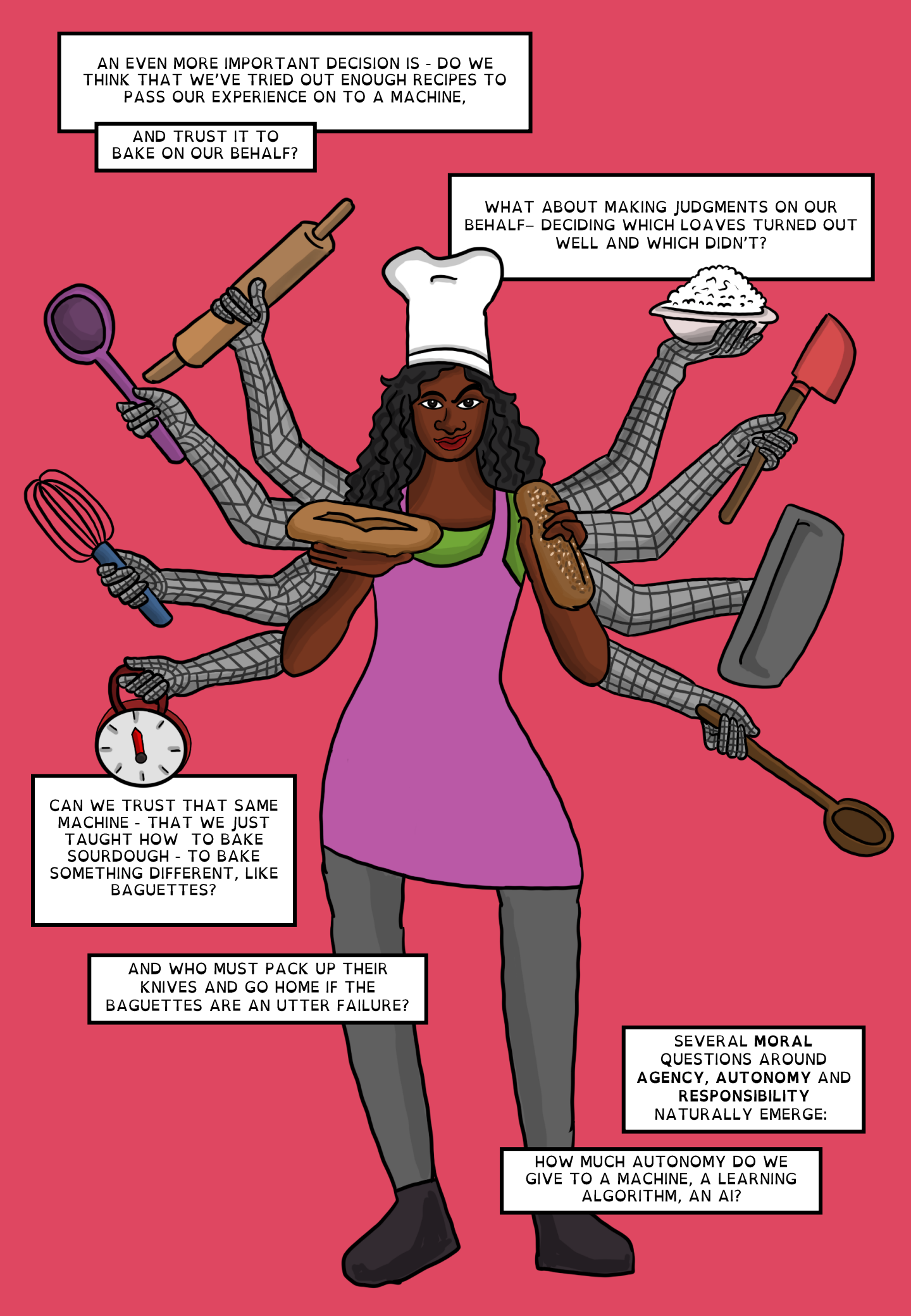
IS THE RESULT INSTAGRAM-WORTHY?

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?


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?
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 BAKE SOURDOUGH - TO BAKE SOMETHING DIFFERENT, LIKE BAGUETTES?
AND WHO MUST PACK UP THEIR KNIVES AND GO HOME IF THE BAGUETTES ARE AN UTTER FAILURE?
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.

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 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."* [2]

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]

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]

SO YOU THINK YOU'RE AN ADS?

DO YOU:

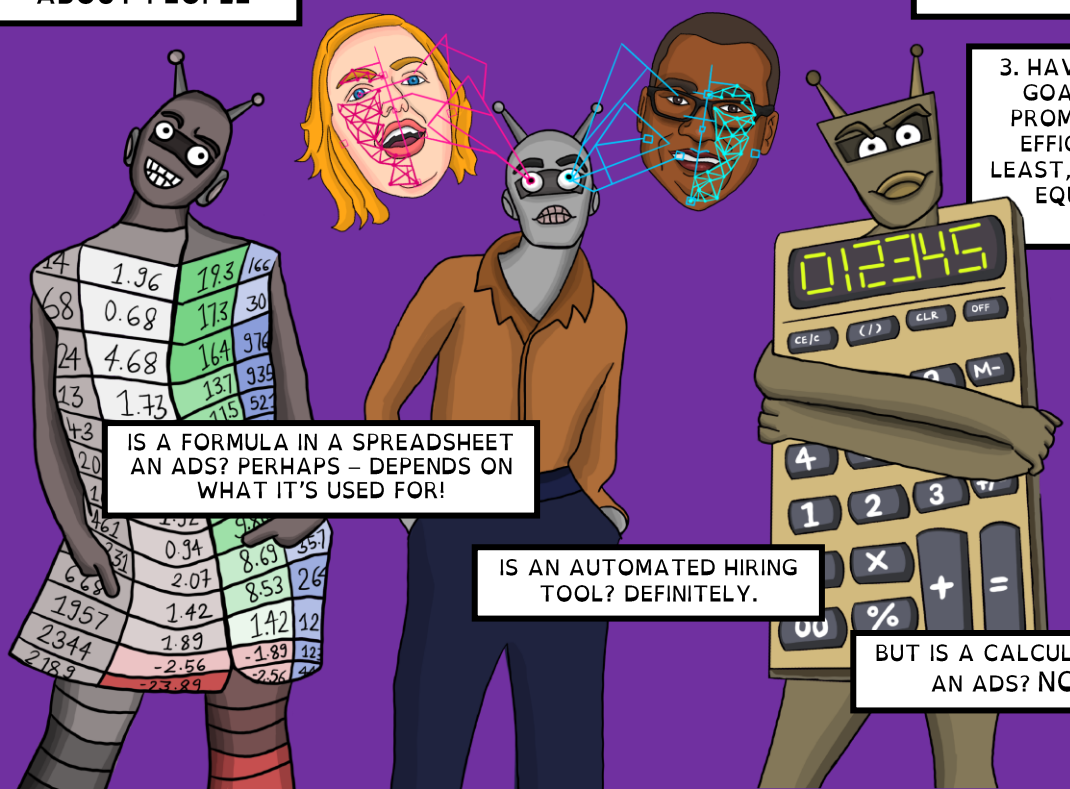
1. PROCESS DATA ABOUT PEOPLE

2. ASSIST - EITHER IN COMBINATION WITH HUMAN DECISION MAKING OR AUTONOMOUSLY - IN MAKING CONSEQUENTIAL DECISIONS THAT IMPACT PEOPLE'S LIVES.

ADDITIONALLY, WE WOULD LIKE IT IF YOU WOULD:

3. HAVE A SPECIFIC, STATED GOAL OF IMPROVING AND PROMOTING EQUALITY AND EFFICIENCY. AT THE VERY LEAST, YOU MUST NOT HINDER EQUITABLE ACCESS TO OPPORTUNITIES

4. BE PUBLICLY DISCLOSED AND SUBJECT TO LEGAL AUDITS.



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!

ALL ABOUT THAT BIAS...

WITH THAT IN MIND, NOW LET'S LOOK AT WHAT WE MEAN BY BIAS IN AN ADS AND HOW IT ARISES. [5]

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.

SYSTEMATIC DISCRIMINATION BY AN ALGORITHM IS TERMED 'BIAS'.

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.

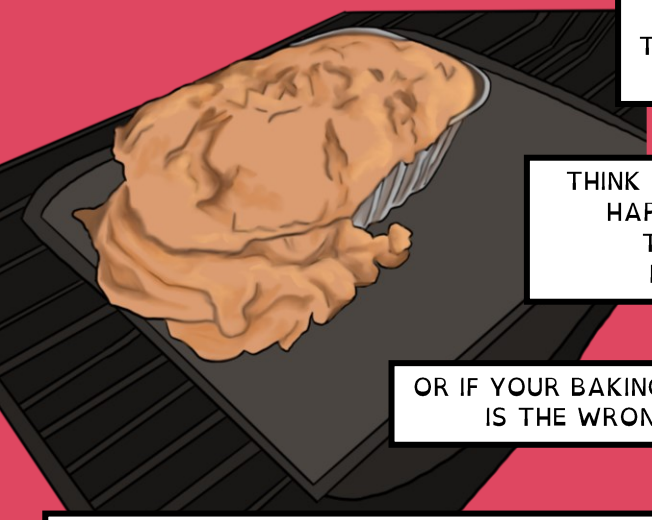
THESE WOULD BE THE FLAVOR NOTES THAT WILL SEEP INTO YOUR BREAD IF YOU DON'T PRIORITIZE THE PURITY/FRESHNESS OF YOUR INGREDIENTS OR IF YOU DECIDE TO USE PREMIXED OFF-THE-SHELF BATTER.

A NOTORIOUS EXAMPLE IS THE GENDER AND RACIAL STEREOTYPES THAT LANGUAGE MODELS PICK UP WHEN TRAINED ON DATA FROM SOCIAL MEDIA PLATFORMS.

TECHNICAL

(IN THE TECHNICAL SYSTEM)

TECHNICAL BIASES 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.

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!

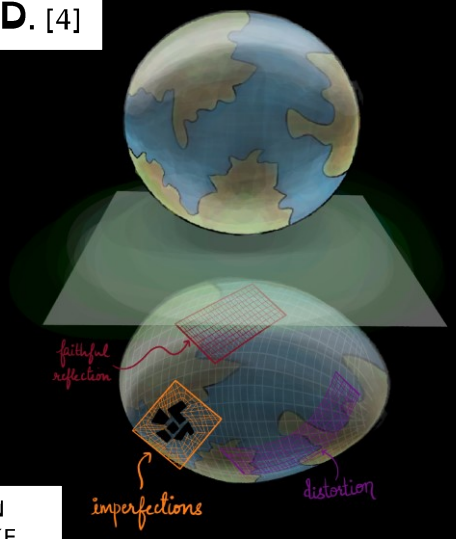
OR MAKE IT SO OFTEN, THAT YOU TURN EVERYONE AROUND YOU OFF THE THOUGHT OF EVER EATING ANOTHER SLICE!

OR THINK ABOUT HOW YOUR IDEA OF 'WHAT BREAD SHOULD TASTE LIKE' IS SHAPED BY THE POPULARITY OF PRODUCTS LIKE 'WONDER BREAD'.

DATA IS A MIRROR REFLECTION OF THE WORLD. [4]

ALL WE HAVE IS A DISTORTED (BIASED) REFLECTION.

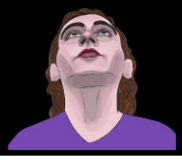
WITHOUT KNOWLEDGE OR ASSUMPTIONS ABOUT THE PROPERTIES OF THE MIRROR AND OF THE WORLD IT REFLECTS, WE CANNOT KNOW WHETHER WE ARE LOOKING AT A DISTORTED REFLECTION OF A PERFECT WORLD OR A PERFECT REFLECTION OF A DISTORTED WORLD OR WHETHER THESE DISTORTIONS COMPOUND. [6]



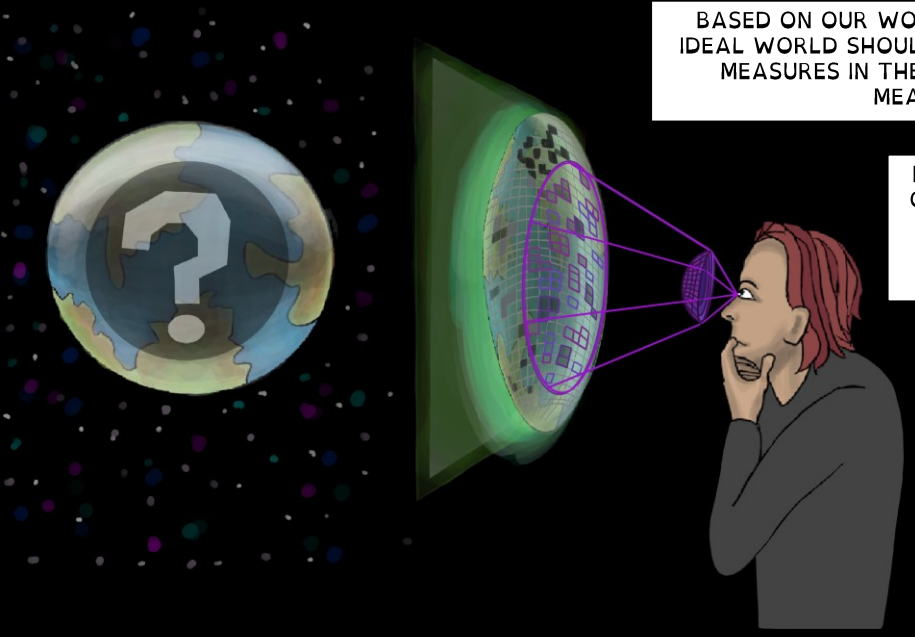
WHAT IS ALGORITHMIC FAIRNESS?

ALGORITHMIC FAIRNESS IS THE CORRECTIVE LENS THAT WE WEAR IN ORDER 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'.

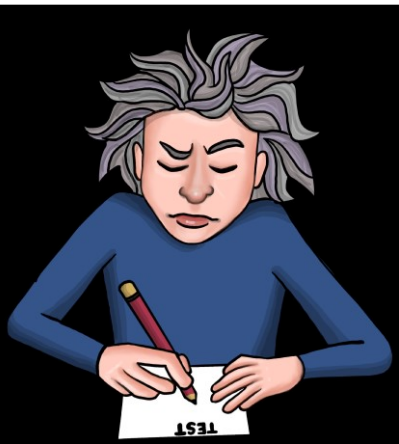


HOWEVER, WEARING THESE LENSES ONLY CHANGES HOW WE VIEW THE REFLECTION - IT DOES NOT AND CANNOT FIX DISTORTIONS IN THE MIRROR OR FIX DISTORTIONS IN THE WORLD.

UNLESS SUCH FIXES ARE SUPPLEMENTED BY SYSTEMIC CHANGE, WE CAN QUICKLY CONFUSE THE WORLD SEEN THROUGH ROSE-COLORED GLASSES WITH THE REAL WORLD.

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). [7]

"INTELLIGENCE" IS THE CONSTRUCT.



TEST SCORES ARE THE OBSERVATIONS THAT WE ARE ACTUALLY ABLE TO MEASURE.

THE DECISION IS WHETHER OR NOT TO CERTIFY ONE'S INTELLECTUAL ABILITY BY CONFERRING UPON THEM A DIPLOMA

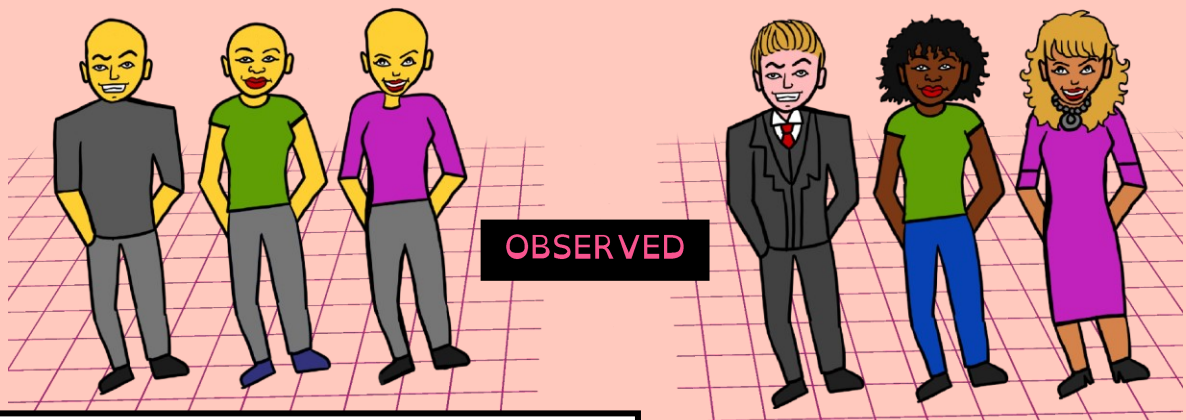
IN A PERFECT WORLD, WHERE THERE IS NEITHER A DISTORTION IN THE WORLD NOR IN THE REFLECTION, OUR CONSTRUCTS AND OUR OBSERVATIONS WOULD BE THE SAME.

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. THESE ASSUMPTIONS COLOR OUR JUDGMENTS ABOUT WHETHER ALLOCATIONS OF BENEFITS ARE 'FAIR' (BY SOME SPECIFIC NOTION).

DIFFERENT WORLDVIEWS AFFECT OUR INTUITIONS ABOUT 'FAIRNESS'. [7]

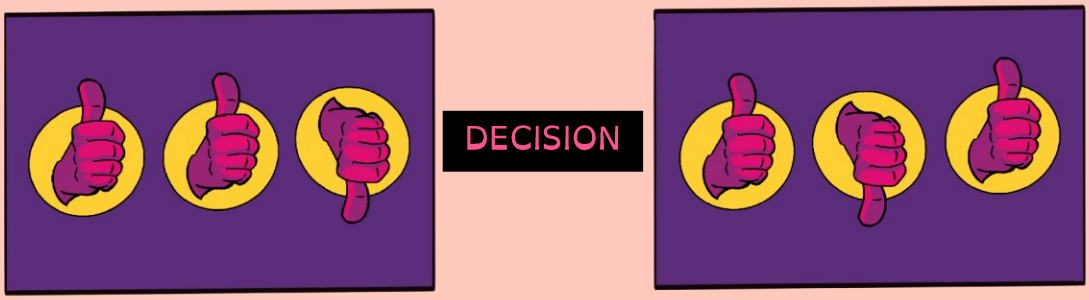
THE **'WHAT YOU SEE IS WHAT YOU GET'** WORLDVIEW ASSUMES THAT RELEVANT CHARACTERISTICS ARE CORRECTLY CAPTURED IN THE DATA

AND THAT DIFFERENCES AMONG PEOPLE'S ABILITIES (BY SOME TASK-SPECIFIC DISTANCE METRIC) ARE PRESERVED FROM THE CONSTRUCT SPACE TO THE OBSERVED SPACE.



ON THE OTHER HAND, THE **'WE ARE ALL EQUAL'** WORLDVIEW IS BASED ON THE IDEA THAT DIFFERENCES IN PEOPLE'S OBSERVED ABILITIES ARE ATTRIBUTABLE TO FACTORS OUTSIDE OF THEIR CONTROL.

THIS IS GOOD NEWS! IF WE CAN CORRECTLY MEASURE PEOPLE'S *TRUE* ABILITIES, WE CAN MAKE 'FAIR' DECISIONS.



IN SO FAR THAT PEOPLE'S ABILITIES CAN BE MEASURED IN A MANNER THAT IS INDEPENDENT OF THEIR PROTECTED CHARACTERISTICS SUCH AS SEX AND RACE, WE CAN MAKE 'FAIR' DECISIONS.

INDIVIDUAL V/S GROUP

INDIVIDUAL FAIRNESS ADVOCATES THAT 'SIMILAR INDIVIDUALS MUST BE TREATED SIMILARLY'. [8]

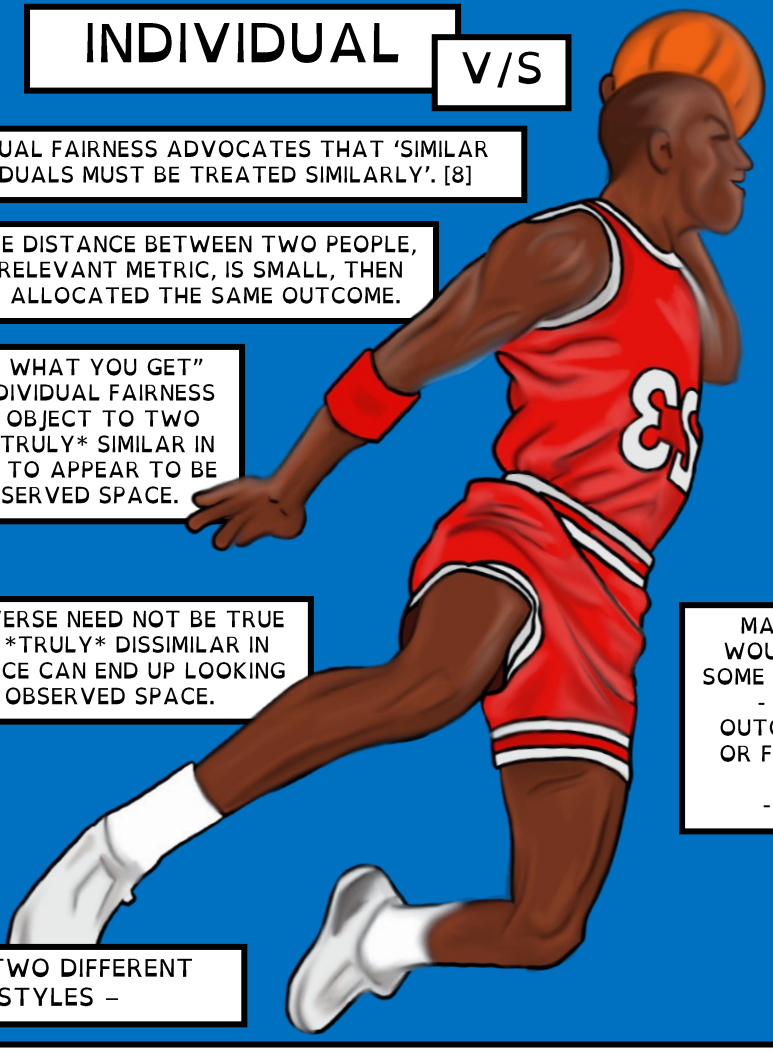
MATHEMATICALLY, IF THE DISTANCE BETWEEN TWO PEOPLE, BASED ON SOME TASK-RELEVANT METRIC, IS SMALL, THEN THEY SHOULD BOTH BE ALLOCATED THE SAME OUTCOME.

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.

GROUP FAIRNESS TRIES TO ENSURE SOME NOTION OF PARITY IN OUTCOMES FOR MEMBERS OF DIFFERENT PROTECTED GROUPS.

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.

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?

OR ARE YOU THE PHIL JACKSON OF THE BULLS, IDENTIFYING THE DIFFERENT STRENGTHS OF DIFFERENT PLAYERS AND ORGANIZING THE TRIANGLE OFFENSE TO PERFECTION,

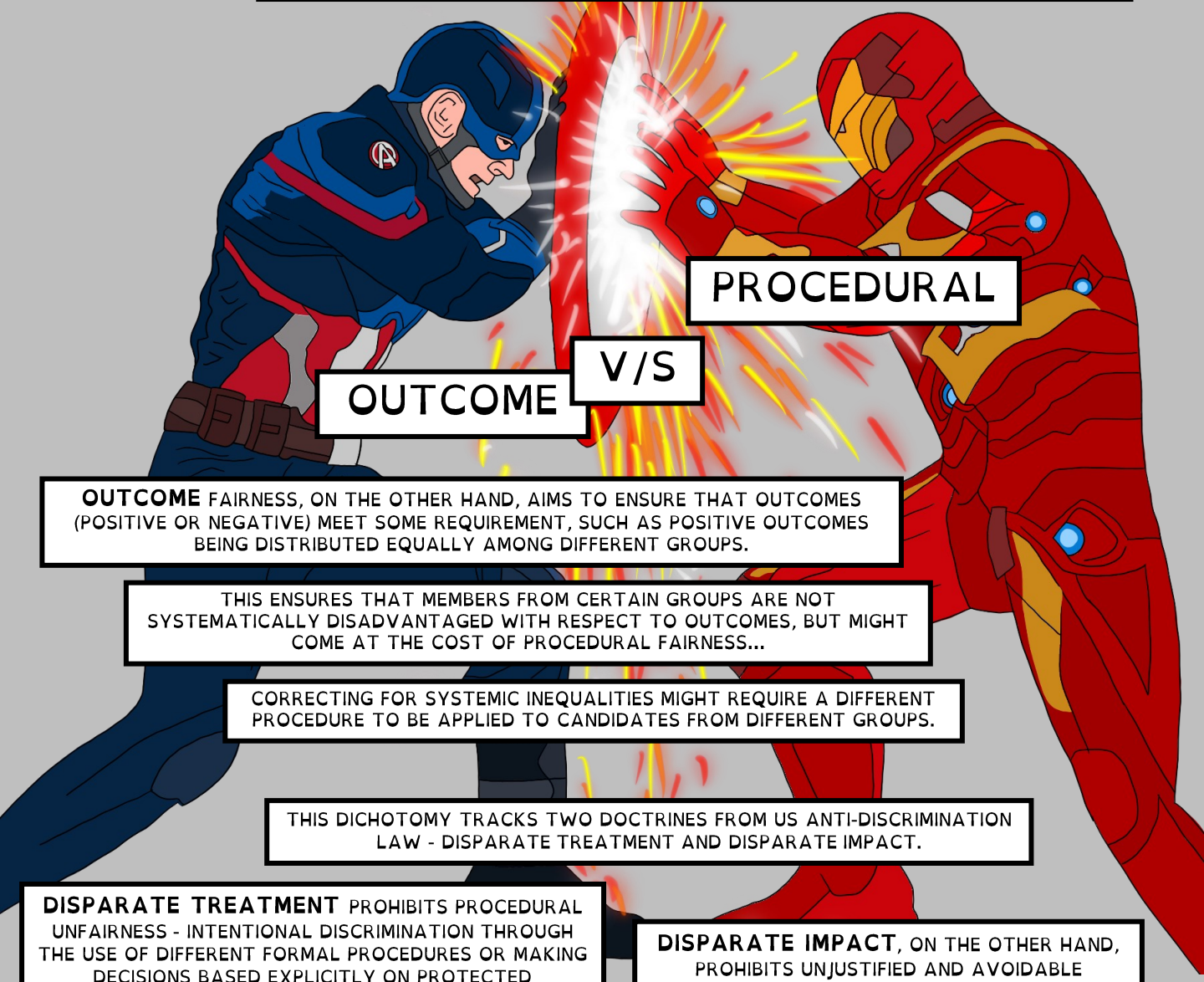
...THEREBY TAKING THE BULLS – LED BY THE INIMITABLE JORDAN, OF COURSE - TO THEIR FIRST CHAMPIONSHIP VICTORY.

IN PRINCIPLE, INDIVIDUAL AND GROUP FAIRNESS NEED NOT BE INCOMPATIBLE [9] – YOU CAN PULL OFF TWO 'THREEPEAT' CHAMPIONSHIP WINS, WHILE HAVING JORDAN WIN LEAGUE MVP EACH YEAR.

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 DISADVANTAGED WITH RESPECT TO OUTCOMES, 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.

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 DISPARITIES IN OUTCOMES FOR PEOPLE OF DIFFERENT PROTECTED GROUPS.

THIS VERY DISAGREEMENT ALMOST BROKE UP THE MIGHTY AVENGERS!

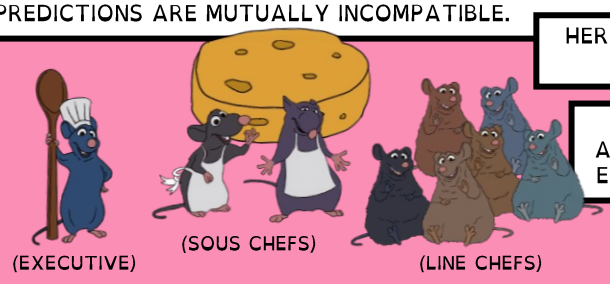
ON ONE HAND, YOU HAVE TEAM STARK, WHO BELIEVE IN SIGNING THE ACCORDS AND OPERATING UNDER A PRESCRIBED MANDATE AND PROCEDURE.

AND THEN THERE ARE THOSE WHO, LIKE CAP, BELIEVE IN THE EFFICACY OF THE OUTCOME, EVEN IF IT REQUIRES PREFERENTIAL TREATMENT.

THE FAMOUS IMPOSSIBILITY RESULTS [10, 11] HAVE DECREED THAT DIFFERENT MEASURES OF 'FAIRNESS' IN PREDICTIONS ARE MUTUALLY INCOMPATIBLE.

HERE'S AN EXAMPLE OF IMPOSSIBILITY IN 'FAIR' RESOURCE ALLOCATION-

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?



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

(OR)

- PURELY DUE TO THAT FACT THAT THERE ARE MORE ROOKIES!

IF YOU DECIDE INSTEAD TO 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.

SEE, HOW DIFFERENT METRICS ARE INHERENTLY INCOMPATIBLE?

AND SO, SINCE WE CANNOT SIMULTANEOUSLY SATISFY DIFFERENT 'FAIRNESS' IDEALS, WE MUST BE CONSCIENTIOUS IN SELECTING A SUITABLE FAIRNESS METRIC FOR OUR PARTICULAR PROBLEM.

WHAT 'WRONG' ARE WE TRYING TO CORRECT?

WHAT DO WE MEAN BY '**FAIRNESS**'?

IS IT **NON-DISCRIMINATION** (FROM LEGAL DOCTRINES)?

IS IT **EQUALITY** IN THE DISTRIBUTION OF SOME COMMODITY/ OUTCOME (IN THE ECONOMIC SENSE)?

IS IT SOME NOTION OF **DISTRIBUTIVE JUSTICE** (FROM POLITICAL PHILOSOPHY)?

THE MOST INFLUENTIAL CHARACTER IN THE FAIR-ML MULTIVERSE SEEMS TO BE **FAIRNESS** AS **EQUALITY OF OPPORTUNITY (EOP).**

LET'S READ THROUGH ITS ORIGIN STORY, SHALL WE?

It's the

Age of EOP!

Libertarian

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.

Formal

The theatre of Formal EOP has its doors open to all talents.

Everyone willing and able is welcome, but you compete with what you have - no special treatment once you're in.

Substantive

Rawlsian

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.

Luck-Egalitarian

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!

Libertarian

"YOU DO YOU!"

THE LIBERTARIAN VIEW FOCUSES ON THE INDIVIDUAL'S FREEDOMS AND LIBERTIES.

LIKE IN 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 DICE, 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.

THIS DOESN'T APPEAR TO BE 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

"CAREERS OPEN TO TALENTS"

(Substantive)

Rawls' Fair EOP

"Equally talented babies must be given equal life prospects"

RAWLS'S FAIR EOP [13] SAYS ALL PEOPLE, REGARDLESS OF HOW RICH OR POOR THEY ARE BORN, SHOULD HAVE OPPORTUNITIES TO DEVELOP THEIR TALENT,

SO THAT PEOPLE WITH THE SAME TALENTS AND MOTIVATION HAVE THE SAME EDUCATIONAL AND EMPLOYMENT OPPORTUNITIES.

RAWLS WANTS TO ENSURE THAT YOUR PRIVILEGED BIRTH DOESN'T SNOWBALL INTO A LIFETIME OF PRIVILEGE THAT ALLOWS YOU TO OUTCOMPETE KIDS WHOSE DISADVANTAGE AT BIRTH HAS LED TO COMPOUNDED DISPRIVILEGE.

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

ALONG THESE LINES, FAIR-ML FORMULATIONS OF RAWLSIAN FAIR EOP INCLUDE STATISTICAL PARITY AND EQUALITY OF ODDS [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 DISTORT RAWLSIAN EOP, WHICH IS FUNDAMENTALLY CONCERNED WITH PROVIDING DEVELOPMENTAL OPPORTUNITIES **BEFORE** COMPETITIONS.

AT THE POINT WHERE AN ADS IS MAKING A DECISION IT IS ALREADY TOO LATE TO PROVIDE PEOPLE WITH OPPORTUNITIES TO BUILD QUALIFICATIONS.

INSTEAD, FAIR-ML FORMULATIONS OF RAWLSIAN EOP MIGHT MEASURE HOW EQUITABLY A COMPETITION DISTRIBUTES DEVELOPMENTAL OPPORTUNITIES **IN ADVANCE OF LATER COMPETITIONS.**

(Substantive)

Luck-Egalitarian EOP

"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.

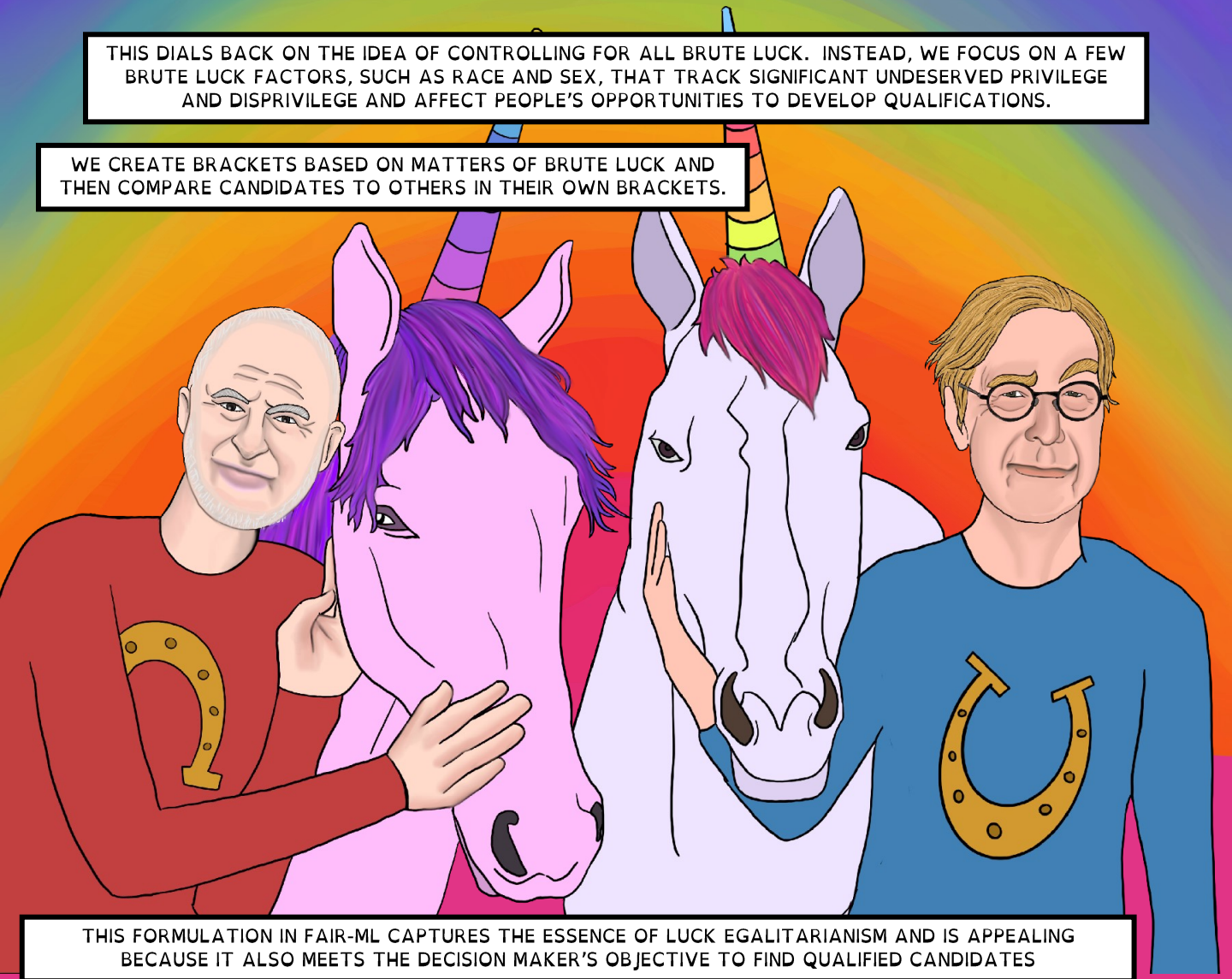
OUR OUTCOMES SHOULD ONLY BE AFFECTED BY OUR "CHOICE LUCK" (RESPONSIBLE CHOICES); NO EFFECTS OF "BRUTE LUCK" (FROM HAVING RICH PARENTS TO GETTING STRUCK BY LIGHTNING) SHOULD BE ALLOWED TO STAND.

HOW DO WE SEPARATE THE EFFECTS OF **LUCK** FROM THE EFFECTS OF **RESPONSIBLE CHOICES?**

ONE POPULAR FORMULATION IN FAIR-ML IS **ROEMER'S EOP** [15], WHICH MEASURES A PERSON'S EFFORT COMPARED TO OTHERS IN SIMILAR CIRCUMSTANCES. [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 DEVELOP QUALIFICATIONS.

WE CREATE BRACKETS BASED ON MATTERS OF BRUTE LUCK AND THEN COMPARE CANDIDATES TO OTHERS IN THEIR OWN BRACKETS.



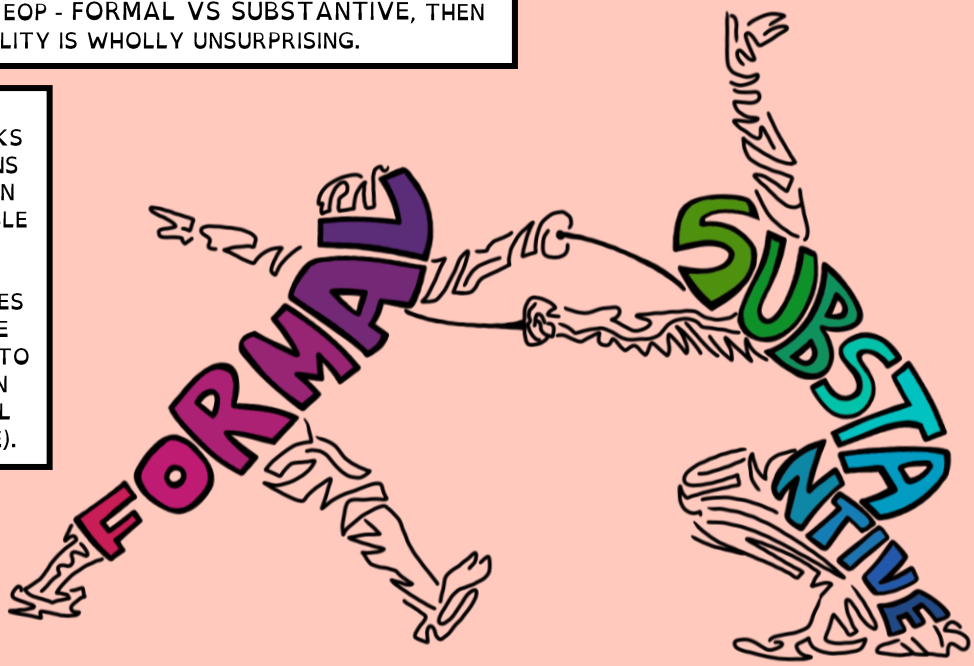
THIS FORMULATION IN FAIR-ML CAPTURES THE ESSENCE OF LUCK EGALITARIANISM AND IS APPEALING BECAUSE IT ALSO MEETS THE DECISION MAKER'S OBJECTIVE TO FIND QUALIFIED CANDIDATES

- 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)

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.

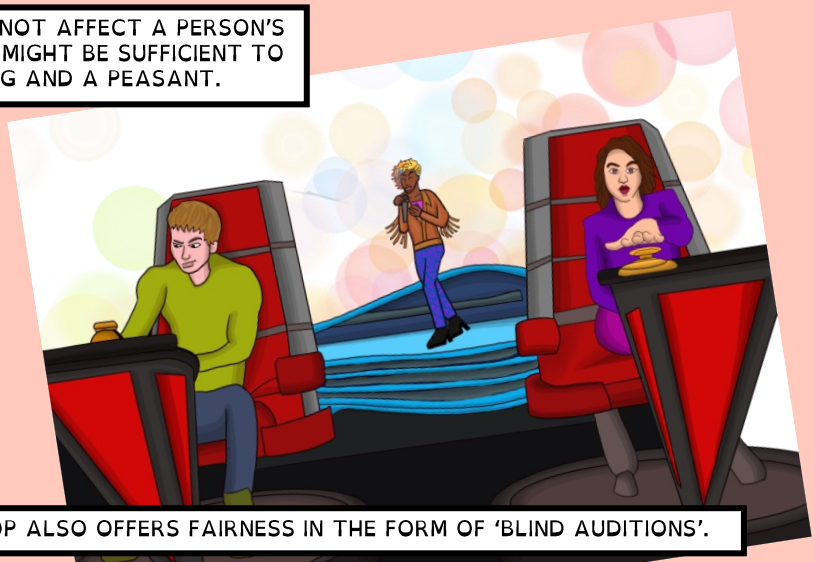
WE WOULD NOT EXPECT A WORLD VIEW THAT ONLY LOOKS AT 'RELEVANT' QUALIFICATIONS AT THE POINT OF COMPETITION (FORMAL EOP) TO BE COMPATIBLE WITH ONE THAT AIMS TO PROVIDE COMPARABLE DEVELOPMENTAL OPPORTUNITIES FOR INDIVIDUALS AND, AT THE POINT OF COMPETITION, SEEKS TO CORRECT FOR INEQUALITIES IN CANDIDATES' DEVELOPMENTAL OPPORTUNITIES (SUBSTANTIVE).



WE CAN INTERPRET THIS INCOMPATIBILITY AS THE DIFFERENCE IN PHILOSOPHICAL VIEWPOINTS AND INCENTIVES OF DECISION MAKERS.

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.



FORMAL EOP ALSO OFFERS FAIRNESS IN THE FORM OF 'BLIND AUDITIONS'.

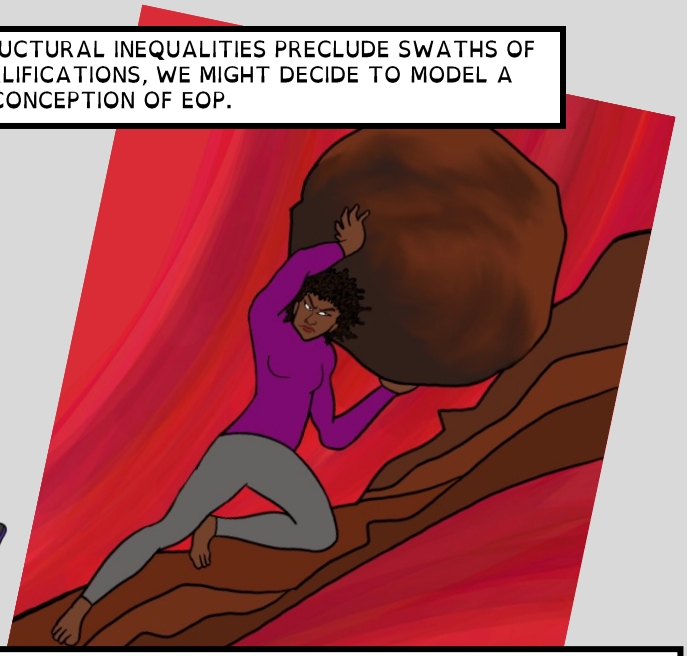
WHEN WE WORRY THAT JUDGES WILL BE SWAYED BY IRRELEVANT TRAITS LIKE GENDER, RACE AND APPEARANCE, BLIND AUDITIONS FORCE JUDGES TO EVALUATE CONTESTANTS SOLELY ON THEIR SINGING CHOPS.

SIMILARLY, MAKING EMPLOYERS BLIND TO JOB APPLICANTS' CREDIT SCORES OR CRIMINAL CONVICTIONS DURING INITIAL APPLICANT SCREENINGS CAN HELP PEOPLE OVERCOME STUBBORN OBSTACLES TO EMPLOYMENT!

WHEN WE HAVE REASON TO BELIEVE THAT STRUCTURAL INEQUALITIES PRECLUDE SWATHS OF PEOPLE FROM DEVELOPING COMPETITIVE QUALIFICATIONS, WE MIGHT DECIDE TO MODEL A MORE SUBSTANTIVE CONCEPTION OF EOP.

IN A FOOTRACE, IF HURDLES – IN THE FORM OF SYSTEMIC DISCRIMINATION AND INEQUITABLE ACCESS – ABOUND IN THE PATH OF CERTAIN COMPETITORS,

WE MIGHT WANT TO COMPARE THE HURDLE JUMPERS WITH OTHER HURDLE 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 DISADVANTAGES THEY'VE HAD RELATIVE TO OTHER COMPETITORS.

ONE IMPORTANT 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).

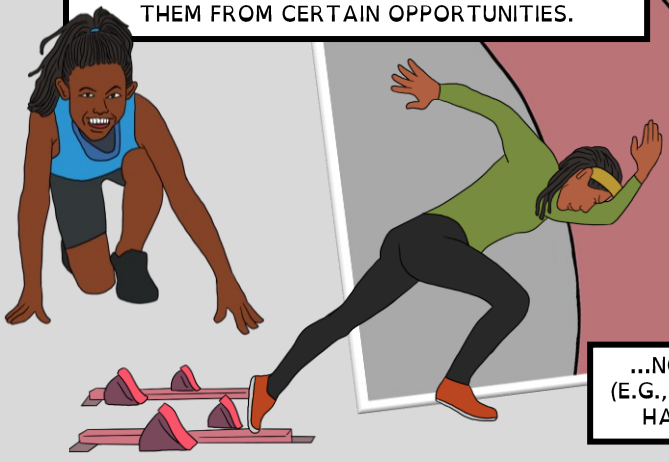
IT MIGHT BE WORTH EXPLORING FAIRNESS OVER THE COURSE OF A LIFETIME – DO PEOPLE HAVE COMPARABLE/EQUALLY DESIRABLE SETS OF LIFE OPPORTUNITIES AVAILABLE TO THEM?

DOES EOP COMPOUND OVER THE COURSE OF A LIFETIME?

OR DOES A DISADVANTAGE OF BIRTH SNOWBALL INTO A LIFETIME OF DISADVANTAGE?

EQUALITY OF DEVELOPMENTAL OPPORTUNITIES IS ABOUT MAKING SURE PEOPLE HAVE COMPARABLE OPPORTUNITIES TO HONE THEIR TALENTS,

INSTEAD OF BEING DISADVANTAGED BY CIRCUMSTANCES OF BIRTH THAT PRECLUDE THEM FROM CERTAIN OPPORTUNITIES.

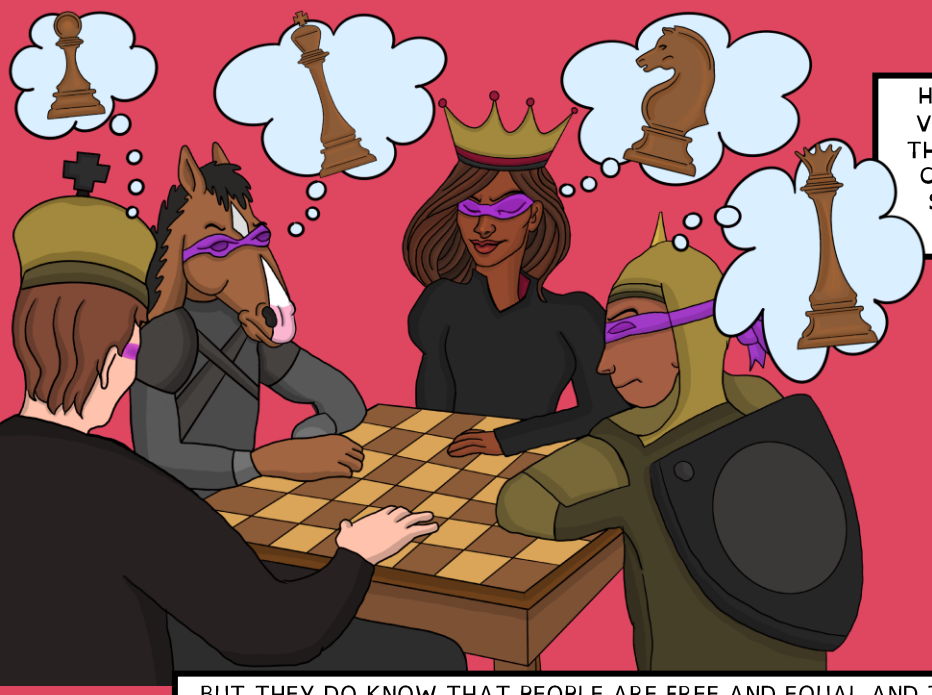


THIS IS MOTIVATED BY THE IDEA THAT WHAT MATTERS FROM THE POINT OF VIEW OF JUSTICE IS PEOPLE HAVING GENUINE OPPORTUNITIES TO REALISTICALLY ACHIEVE GOALS (E.G. BEING A TRACK ATHLETE),

...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).

OUR STROLL THROUGH EOP-VILLE HAS SHOWN US A RANGE OF INTERPRETATIONS OF 'FAIRNESS'. BUT IS 'FAIRNESS' ALL THAT'S REQUIRED FOR AN ALGORITHM TO BE 'JUST'?

RAWLS SANDWICHES HIS EOP PRINCIPLE BETWEEN TWO OTHER PRINCIPLES THAT ALSO MUST BE SATISFIED FOR A DEMOCRATIC SOCIETY TO BE 'JUST'.



HE ARRIVES AT THESE PRINCIPLES VIA **THE ORIGINAL POSITION**- A THOUGHT EXPERIMENT ABOUT HOW CITIZENS WOULD NEGOTIATE THE SET-UP OF SOCIETY, UNDER THE **'VEIL OF IGNORANCE'**

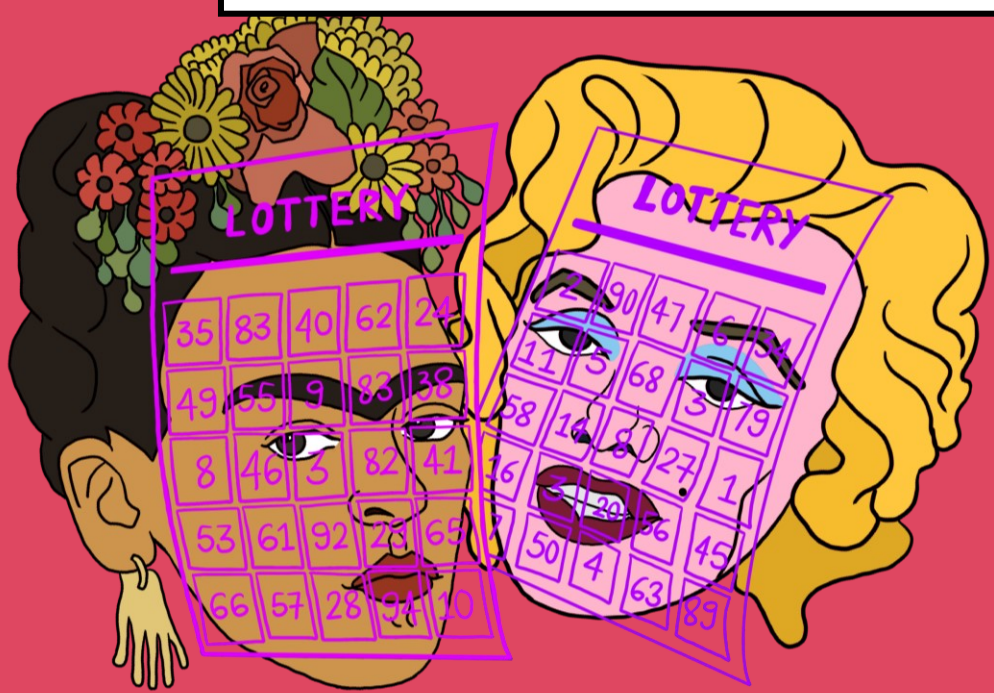
- IF CITIZENS DO NOT KNOW THEIR RACE, CLASS, SEX, TALENTS, SOCIAL POSITION (OR ANY OTHER CHARACTERISTICS THAT MIGHT CAUSE THEM TO FAVOR PEOPLE LIKE THEMSELVES), THEY WILL ADVOCATE FOR ALL SOCIAL POSITIONS AND THEIR ATTACHED PRIVILEGES TO BE DISTRIBUTED 'FAIRLY'.

BUT THEY DO KNOW THAT PEOPLE ARE FREE AND EQUAL AND THAT THEY HAVE THE ABILITY TO CHOOSE A CONCEPTION OF THE GOOD LIFE AND THE ABILITY TO ABIDE BY RULES OF JUSTICE.

AND SO, RAWLS POSITS THAT THE PRINCIPLES OF SOCIAL COOPERATION THAT PEOPLE ARRIVE AT THROUGH SUCH A NEGOTIATION WILL BE APPROPRIATE FOR A FREE AND DEMOCRATIC SOCIETY.

RAWLS USES THE NOTION OF THE **"NATURAL LOTTERY"** TO DESCRIBE THE MORALLY ARBITRARY DISTRIBUTION OF TALENTS, FAMILY CIRCUMSTANCES, AND OTHER AT-BIRTH FORTUNE AND MISFORTUNE TO PEOPLE.

FROM THE ARBITRARINESS OF THE NATURAL LOTTERY, RAWLS CONCLUDES THAT WE DON'T DESERVE OUR STARTING POINTS IN LIFE,



...AND ARRIVES AT THE **DIFFERENCE PRINCIPLE** - WHICH HARNESSES THE ARBITRARY DISTRIBUTION OF TALENTS TO GENERATE A SOCIAL SYSTEM THAT SERVES EVERYONE.

RAWLS' THEORY OF JUSTICE

POSITS THE FOLLOWING HIERARCHICAL PRINCIPLES: [13]

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.

2. (b) **[DIFFERENCE PRINCIPLE]** SOCIAL AND ECONOMIC INEQUALITIES MUST BE OF THE GREATEST BENEFIT TO THE LEAST ADVANTAGED

IN THE RAWLSIAN SYSTEM, THESE PRINCIPLES ARE HIERARCHICALLY ORDERED –

FAIR EOP CAN'T BE SATISFIED AT THE EXPENSE OF CITIZENS' EQUAL BASIC RIGHTS AND LIBERTIES,

AND THE DIFFERENCE PRINCIPLE CAN'T BE SATISFIED AT THE EXPENSE OF EOP

- AKIN TO HOW INCREDIBLY COUNTERINTUITIVE IT WOULD BE TO PUT ON A BLAZER, WITHOUT WEARING A SHIRT FIRST!

FOR EXAMPLE, TAKE THE CHILDREN OF RICH PARENTS -

IN TRYING TO GIVE PEOPLE ACCESS TO EQUAL DEVELOPMENTAL OPPORTUNITIES, ONE MIGHT END UP PREVENTING PARENTS FROM RAISING KIDS ACCORDING TO THEIR VALUES,

BECAUSE THIS WOULD MEAN THAT SOME KIDS GET BETTER DEVELOPMENTAL OPPORTUNITIES THAT OTHERS.

IN TRYING TO SATISFY RAWLS'S FAIR EOP, WE MIGHT END UP INFRINGING ON RICH PARENTS' BASIC LIBERTIES.

IN THE CONTEXT OF ALGORITHMS, 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.

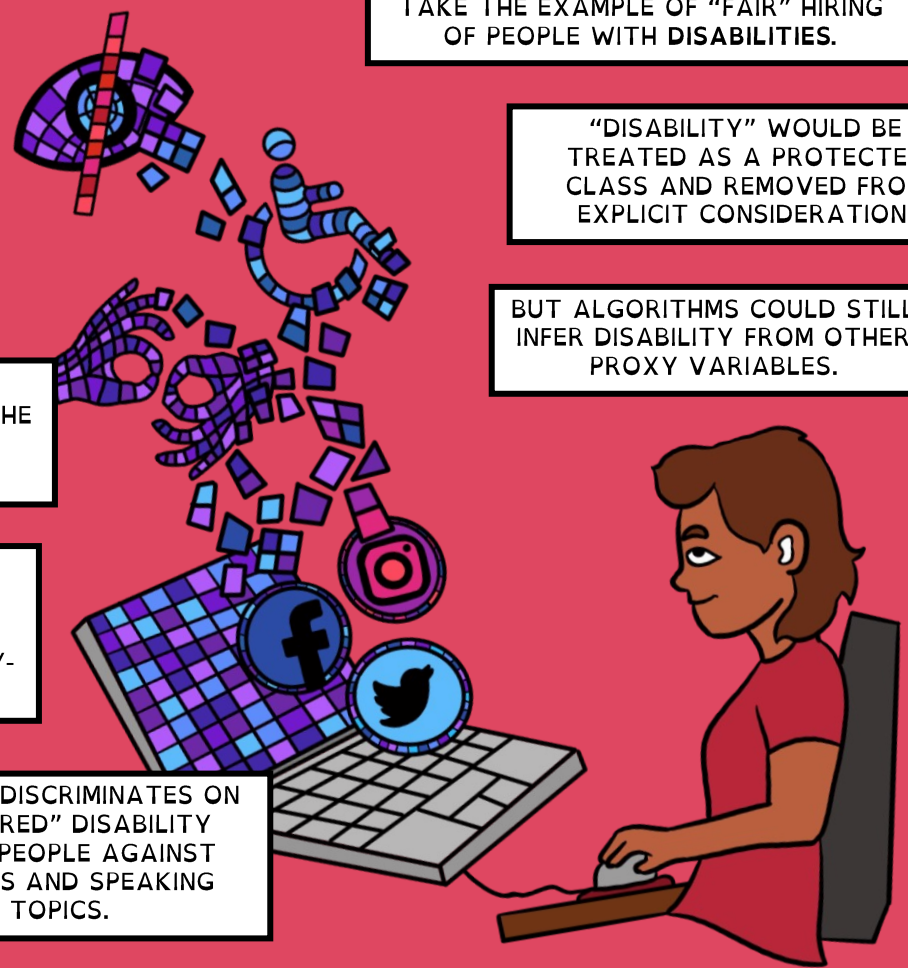
TAKE THE EXAMPLE OF "FAIR" HIRING OF PEOPLE WITH **DISABILITIES.**

"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.

SUCH AN ADS COULD SATISFY SOME CONCEPTION OF 'FAIRNESS' AS EOP AND YET BE FUNDAMENTALLY UNJUST: IT WOULD VIOLATE A CANDIDATE'S FREEDOM OF SPEECH AND FREEDOM OF ASSOCIATION.

THERE ARE LIMITATIONS TO WHAT ANSWERS WE CAN GET FROM EOP DOCTRINES,

AND OVERLOOKING THESE CAN EMBOLDEN THEIR APPLICATION IN SPHERES IN WHICH THEORY PROVIDES LITTLE TO NO GUIDANCE...

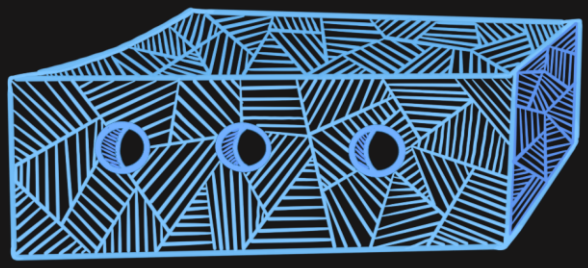
THESE DOCTRINES DO NOT GIVE US ANY DIRECTION ABOUT *WHERE* TO APPLY 'FAIRNESS' - IN THE PROCEDURE OR AT THE OUTCOME.

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.



THE FAIRNESS YOU ASKED FOR IS INSIDE THIS BOX!

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 CIRCUMSTANCE (TO BE CONTROLLED FOR) AND EFFORT (THAT THE INDIVIDUAL CAN BE HELD ACCOUNTABLE FOR), THEN WE MUST DECIDE HOW TO MAKE THIS SEPARATION!

WHICH OUTCOMES CAN WE HOLD ONE ACCOUNTABLE FOR?

AND WHICH MATTERS ARE COMPLETELY OUT OF THEIR CONTROL?

SOUNDS LIKE WHAT WE NEED IS A SORTING HAT!

FROM A PRACTICAL PERSPECTIVE, IT IS OBVIOUS THAT WE CANNOT SEPARATE A PERSON'S NATIVE TALENTS FROM THEIR CIRCUMSTANCES, OR THEIR RESPONSIBLE CHOICES FROM THEIR BRUTE LUCK.

AND YET, WE HAVE TESTS LIKE THE SAT AND GRE, WHICH SUPPOSEDLY GAUGE INTELLIGENCE AND ACADEMIC EXCELLENCE, AND ARE USED TO MAKE UNIVERSITY ADMISSIONS DECISIONS.

#TransWomenareWomen

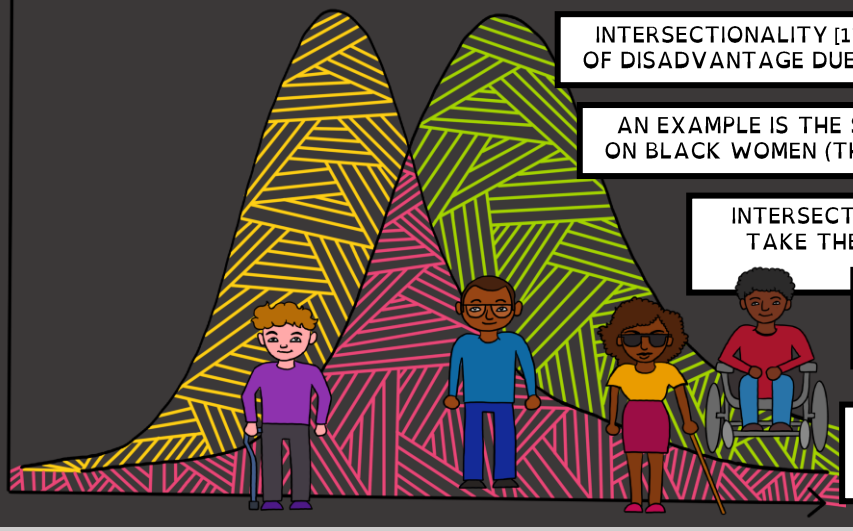
JUST REGISTERING FOR SUCH A TEST - FORGET ABOUT GETTING ACCESS TO STUDY MATERIALS - IS PROHIBITIVELY EXPENSIVE.

SUCH STANDARDIZED TESTS DO NOT EVALUATE NATIVE TALENT, BUT INSTEAD DISCRIMINATE ON THE BASIS OF SOCIAL ADVANTAGES AND DISADVANTAGES.

ALWAYS.

ADS ARE BROADLY USED SOCIO-TECHNO-POLITICAL SYSTEMS.

SOCIAL DYNAMICS OF POWER AND OPPRESSION ARE HIGHLIGHTED BY PROBLEMS OF INTERSECTIONALITY



INTERSECTIONALITY [17] ANALYZES OVERLAPPING DIMENSIONS OF DISADVANTAGE DUE TO SEX, RACE, CLASS, DISABILITY, ETC.

AN EXAMPLE IS THE STUDY OF FACIAL RECOGNITION SOFTWARE ON BLACK WOMEN (THE INTERSECTION OF RACE AND GENDER). [18]

INTERSECTIONALITY CAN BE CAUSAL IN NATURE [19] – TAKE THE INTERSECTION OF RACE AND DISABILITY.

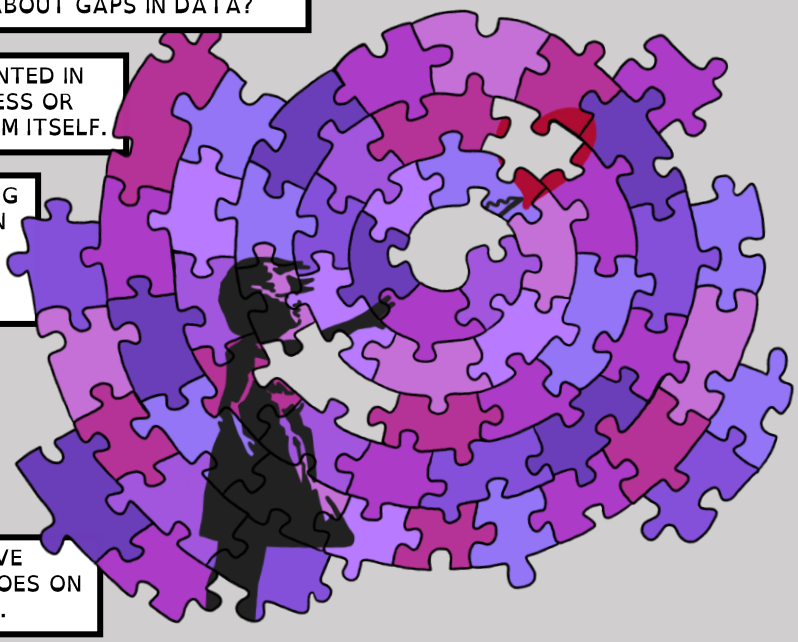
DUE TO UNEQUAL ACCESS TO HEALTHCARE, BLACK INDIVIDUALS ARE MORE LIKELY TO BECOME DISABLED.

MEASURING HOW BIASES INTERACT AND COMPOUND IS A HARD OPEN PROBLEM. [20]

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.



THEN THERE'S THE PROBLEM OF OBSERVABILITY [20]. IN MOST 'FAIRNESS' RELATED TASKS WE ARE MODELLING FOR 'RISKS' - 'RISK OF LOAN DEFAULT', 'RISK OF RECIDIVISM', 'RISK OF COLLEGE DROPOUT'.

SELDOM DO WE GET TO OBSERVE WHETHER THE PERSON ACTUALLY GOES ON TO DO ANY OF THOSE THINGS.

THIS LEADS US TO THE TRADEOFF BETWEEN EXPLORATION AND EXPLOITATION.

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?

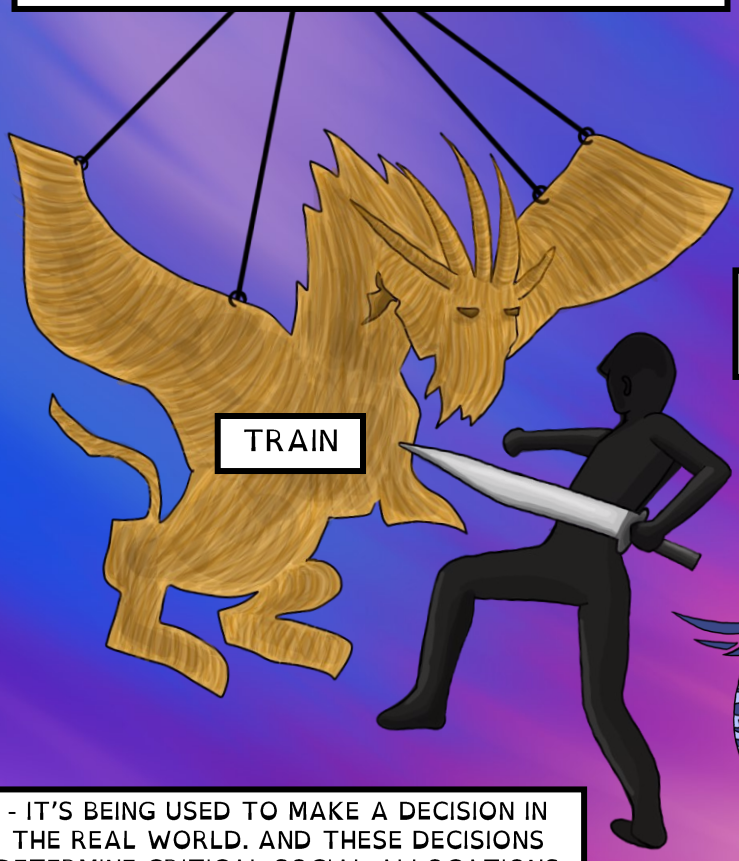
ARE THESE COSTS BORNE DISPROPORTIONATELY BY A CERTAIN DEMOGRAPHIC?

DOES THIS LEAD TO NEW FORMS OF 'UNFAIRNESS'? [21]

BEFORE WE DEPART, LET US HEED AN IMPORTANT WARNING ABOUT THE NATURE OF THIS TALE...

BIAS IS A THREE-HEADED DRAGON, EACH HEAD A FORMIDABLE OPPONENT IN ITS OWN RIGHT. 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.



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 LOANS.

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 OFF ONE HEAD OF PRE-EXISTING BIAS, AND TWO NEW HEADS OF EMERGENT BIASES GROW OUT.

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...

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?

FIN.

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. We hope that it will serve as the computer scientist's guide to political philosophy!

Falaah Arif Khan,
Co-Creator, Author, Artist

Falaah is a scientist/engineer by training and an artist by nature, and the creator of MachineLearnist Comics - a collection of webcomics about the current AI landscape.

Eleni Manis,
Author

Eleni is the Research Director at the Surveillance Technology Oversight Project. 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 Stoyanovich,
Co-Creator, Author

Julia is an Assistant Professor of Computer Science and Engineering and of Data Science and the founding Director of the Center for Responsible AI at New York University. She leads the 'Data, Responsibly' project, the latest offering of which is the inimitable interdisciplinary course on Responsible Data Science.

REFERENCES

1. https://www.wired.com/story/timnit-gebru-exit-google-exposes-crisis-in-ai/
2. https://www1.nyc.gov/assets/buildings/local_laws/ll49of2019.pdf
3. New York City Automated Decision Systems Task Force Report.
4. Julia Stoyanovich, Bill Howe, and H. V. Jagadish. (2020). Responsible data management. Proc. VLDB Endow. 13, 12 (August 2020)
5. Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. ACM Trans. Inf. Syst. 14, 3 (July 1996)
6. Falaah Arif Khan and Julia Stoyanovich. "Mirror, Mirror". Data, Responsibly Comics, Volume 1 (2020)
7. Sorelle A. Friedler and Carlos Scheidegger and Suresh Venkatasubramanian. (2016). On the (im)possibility of fairness
8. Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. (2012). Fairness through awareness. ITCS 2012
9. Reuben Binns. (2020). On the apparent conflict between individual and group fairness. FAT* 2020
10. Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, (2017).
11. Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. (2016)
12. Fishkin, Joseph. Bottlenecks: A New Theory of Equal Opportunity. Oxford: Oxford University Press, 2014.
13. John Rawls. A theory of justice. Harvard University Press, (1971)
14. Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. NeurIPS 2016.
15. John. E. Roemer. Equality of opportunity: a progress report. Social Choice and Welfare. (2002)
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)
17. 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).
18. Joy Buolamwini and Timnit Gebru. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. FAT* 2018
19. Ke Yang, Joshua Loftus & Julia Stoyanovich (2020). Causal intersectionality for fair ranking.
20. Alexandra Chouldechova and Aaron Roth (2020) A snapshot of the frontiers of fairness in machine learning. CACM 63, 5 (May 2020)
21. 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.