Introduction and Algorithmic Fairness (Part 1)

Responsible Data Science
DS-UA 202 and DS-GA 1017

Weeks 1–2

Instructors: Julia Stoyanovich and George Wood

This reader contains links to online materials and excerpts from selected articles on introduction to responsible data science and on algorithmic fairness. For convenience, the readings are organized by course week. Please note that some excerpts end in the middle of a section. Where that is the case, the partial section is not required reading.

Week 1: Introduction

DATA, RESPONSIBLY

#1

MachineLearnist COMICS

MIRROR, MIRROR

TERMS OF USE

All the panels in this comic book are licensed CC 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 and Julia Stoyanovich. "Mirror, Mirror".
Data, Responsibly Comics, Volume 1 (2020)
https://dataresponsibly.github.io/comics/vol1/mirror_en.pdf

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

HEY THERE!
YOU MADE IT!

WELCOME TO **OPTOPIA!** (1)

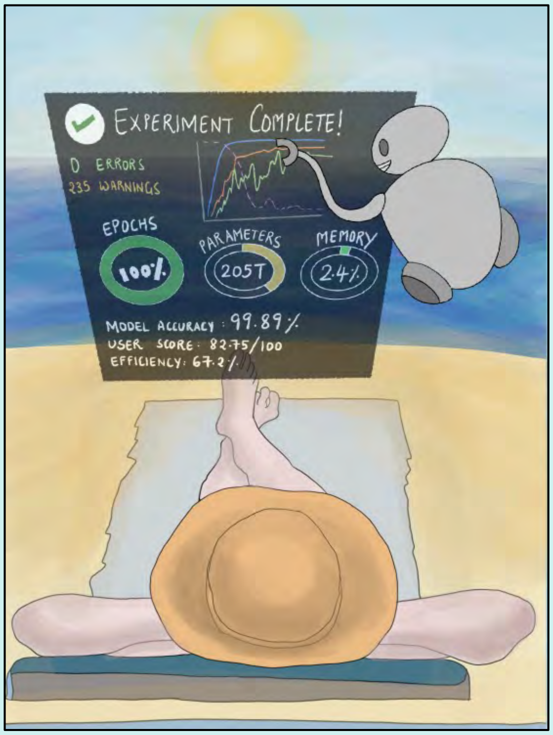
IT'S THE LAND OF ALGORITHM DRIVEN UTOPIA!

REMEMBER ALL THOSE CRAZY SCIENTISTS TALKING FOR DECADES ABOUT CREATING ARTIFICIAL INTELLIGENCE? WELL, THIS IS IT.

WE ALL LAUGHED AT THEM AND SAID IT WAS IMPOSSIBLE (2), BUT YOU KNOW WHAT...

THEY WERE RIGHT. THEY DID IT.

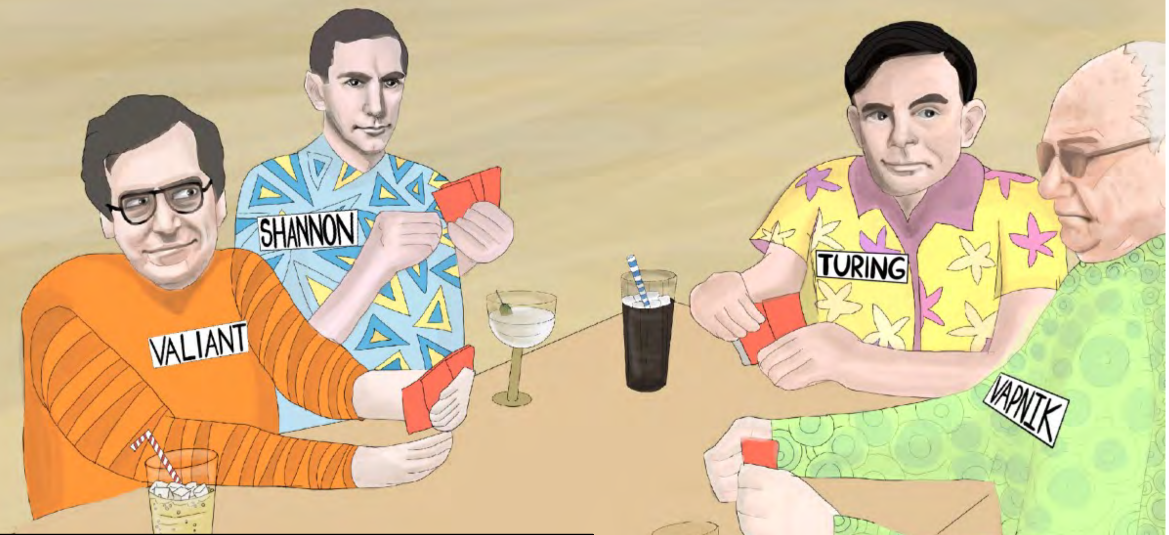
AND NOW THEY JUST SIT BACK AND RELAX WHILE THEIR REPLICAS DO ALL THE WORK.



LOOK AT THIS GUY, HE JUST PUBLISHED A NEW PAPER, ALL WHILE SIPPING A NICE GLASS OF WINE.

I KNOW WHAT YOU'RE THINKING..

IS THIS YET ANOTHER WHITEWASHED HOLLYWOOD PRODUCTION?

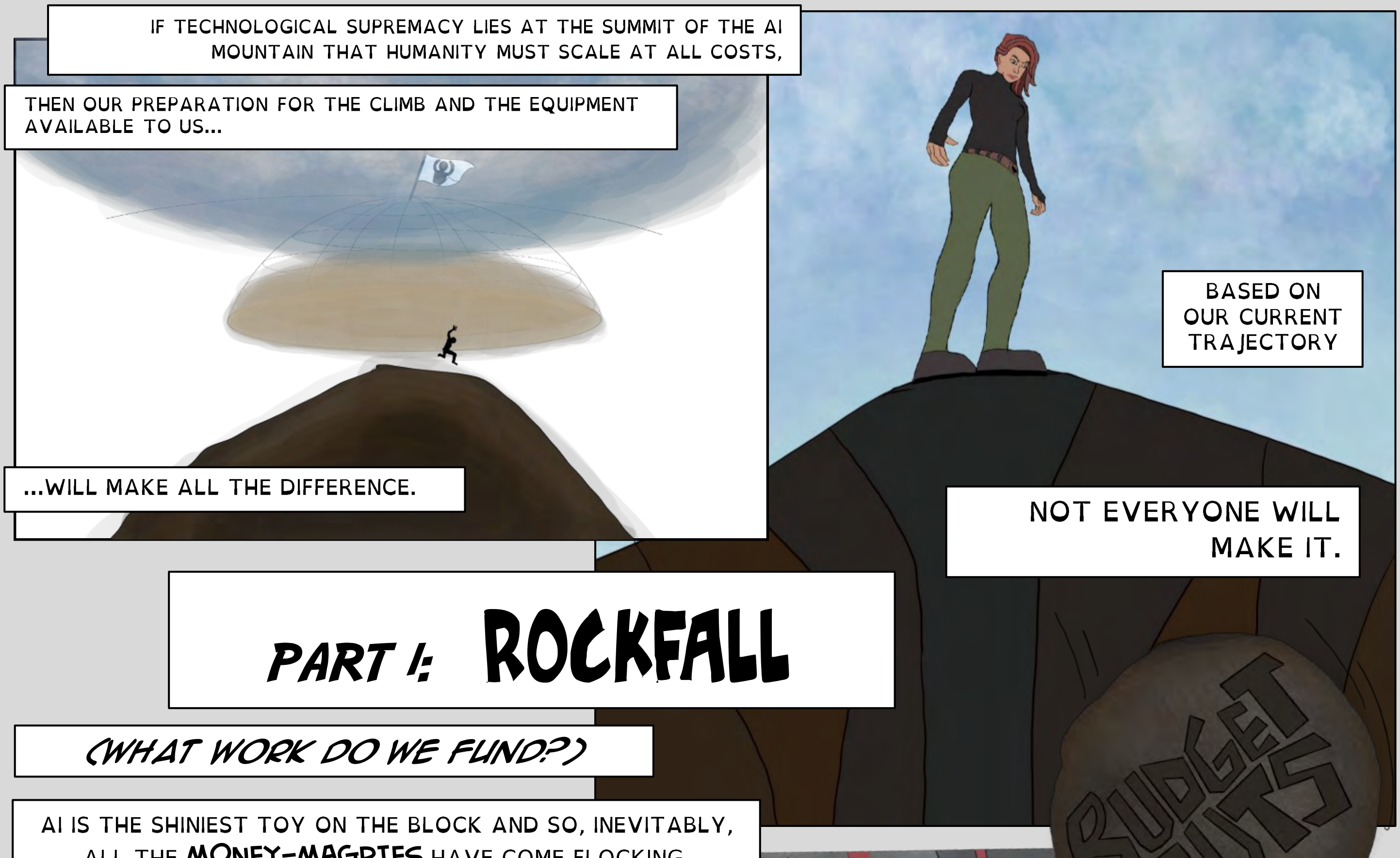


WHERE ARE ALL THE **WOMEN** AND **PEOPLE OF COLOR**?

IF TECHNOLOGICAL SUPREMACY LIES AT THE SUMMIT OF THE AI MOUNTAIN THAT HUMANITY MUST SCALE AT ALL COSTS,

THEN OUR PREPARATION FOR THE CLIMB AND THE EQUIPMENT AVAILABLE TO US...

...WILL MAKE ALL THE DIFFERENCE.



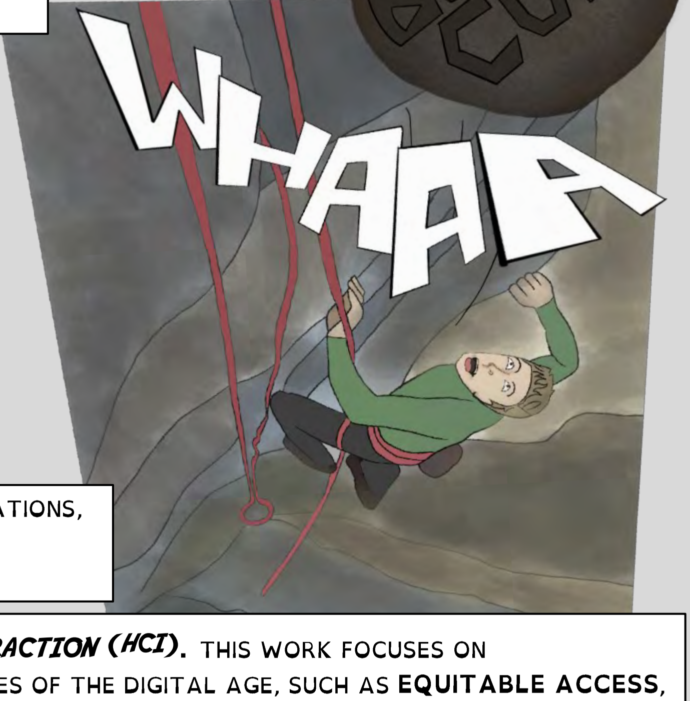
BASED ON OUR CURRENT TRAJECTORY

NOT EVERYONE WILL MAKE IT.

PART I: ROCKFALL

(WHAT WORK DO WE FUND?)

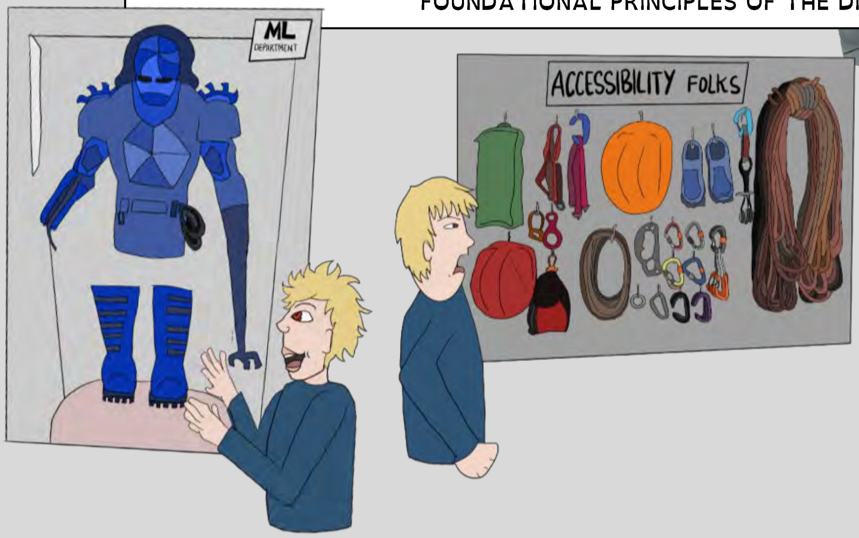
AI IS THE SHINIEST TOY ON THE BLOCK AND SO, INEVITABLY, ALL THE **MONEY-MAGPIES** HAVE COME FLOCKING.



HOWEVER, BEYOND THE USUAL SLEW OF POPULAR APPLICATIONS, SUCH AS **VISION** AND **LANGUAGE MODELING**, THE MONEY SELDOM TRICKLES DOWN.

FOR EXAMPLE, TAKE ***HUMAN-COMPUTER INTERACTION (HCI).*** THIS WORK FOCUSES ON FOUNDATIONAL PRINCIPLES OF THE DIGITAL AGE, SUCH AS **EQUITABLE ACCESS**,

AND YET IT SELDOM SEES THE KIND OF ECONOMIC BACKING OR MEDIA COVERAGE AS MACHINE LEARNING (ML) DOES.



LET'S GIVE **HCI** A MOMENT IN THE **SPOTLIGHT**, SHALL WE?

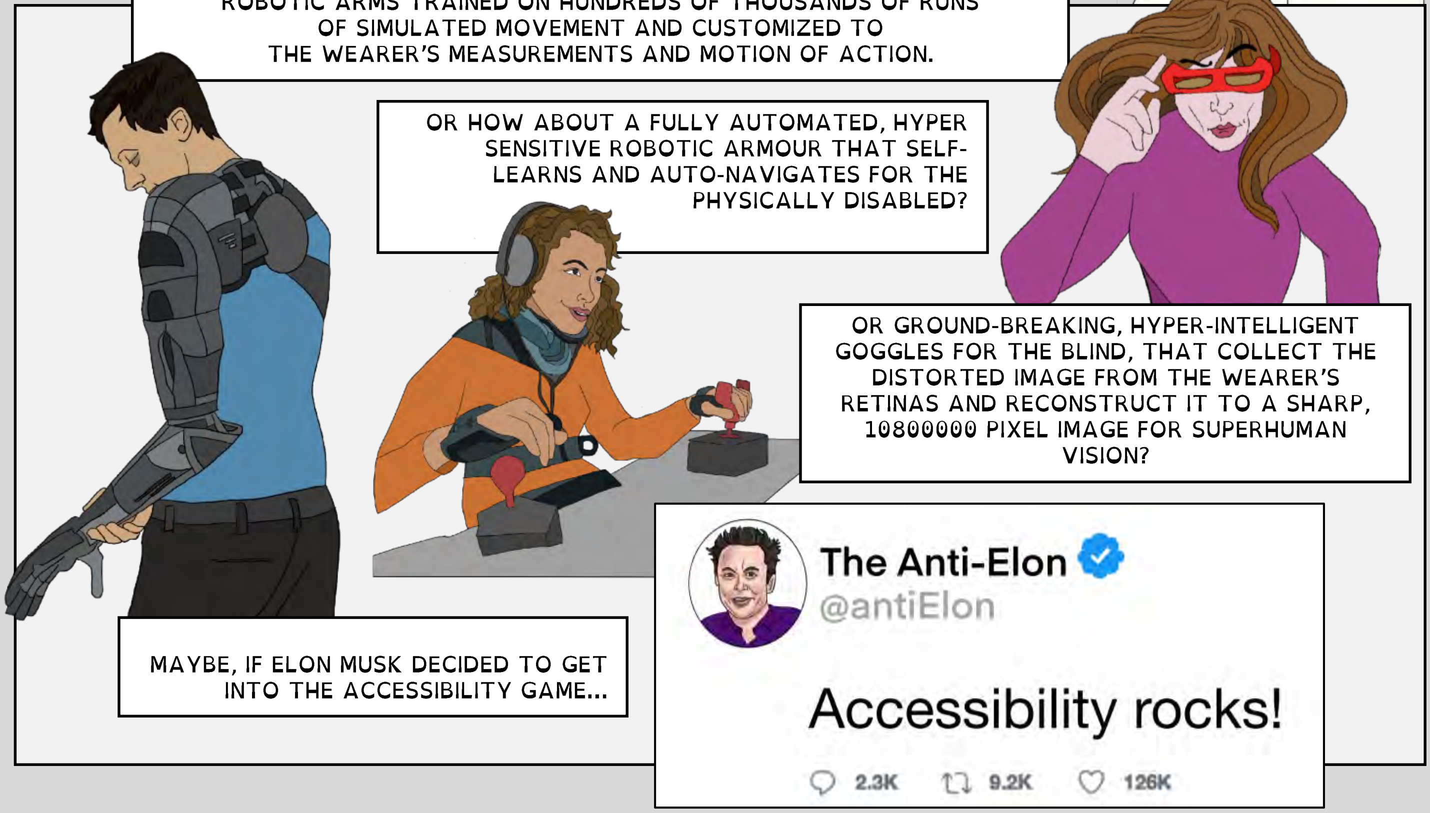
DIGITAL ACCESSIBILITY

DID YOU KNOW?

15% OF THE ENTIRE POPULATION EXPERIENCE SOME FORM OF DISABILITY- VISUAL, AUDITORY, MOTOR OR COGNITIVE. (3)

> "THE POWER OF THE WEB IS IN ITS UNIVERSALITY. ACCESS BY EVERYONE REGARDLESS OF DISABILITY IS AN ESSENTIAL ASPECT"
>
> -TIM BERNERS-LEE

SO, WHAT IS **DIGITAL ACCESSIBILITY**?
THIS VOLUME IS ABOUT ML AND DATA, SO YOU'RE PROBABLY IMAGINING ROBOTIC ARMS TRAINED ON HUNDREDS OF THOUSANDS OF RUNS OF SIMULATED MOVEMENT AND CUSTOMIZED TO THE WEARER'S MEASUREMENTS AND MOTION OF ACTION.



OR HOW ABOUT A FULLY AUTOMATED, HYPER SENSITIVE ROBOTIC ARMOUR THAT SELF-LEARNS AND AUTO-NAVIGATES FOR THE PHYSICALLY DISABLED?

OR GROUND-BREAKING, HYPER-INTELLIGENT GOGGLES FOR THE BLIND, THAT COLLECT THE DISTORTED IMAGE FROM THE WEARER'S RETINAS AND RECONSTRUCT IT TO A SHARP, 10800000 PIXEL IMAGE FOR SUPERHUMAN VISION?

MAYBE, IF ELON MUSK DECIDED TO GET INTO THE ACCESSIBILITY GAME...

The Anti-Elon ✓
@antiElon

Accessibility rocks!

2.3K 9.2K 126K

IN OUR REALITY, DIGITAL ACCESSIBILITY IS FOCUSED ON MAKING SURE WEB PLATFORMS ARE EASILY NAVIGABLE AND USABLE BY PEOPLE WITH ANY KIND OF DISABILITY

IT IS THIS VERY WORK THAT MAKES SURE THAT THE IMAGE YOU JUST POSTED ON INSTAGRAM HAS CAPTIONS

SO THAT THE BLIND USERS OF THE PLATFORM CAN ALSO PARTAKE IN YOUR TRIUMPH OVER THAT SOURDOUGH RECIPE.

OR WHEN YOU DROP A NEW TUTORIAL VIDEO FOR ALL ONE SQUILLION OF YOUR SUBSCRIBERS TO ENJOY,

IT IS THIS WORK THAT CONVERTS YOUR VOCAL PEARLS OF WISDOM INTO TEXT FOR YOUR DEAF FOLLOWERS.

NEW PATH!
AAAAA

ACCESSIBILITY NEEDS TO BE A FUNDAMENTAL DESIGN PRINCIPLE FOR BUILDING WEBSITES AND SOFTWARE,

BUT IN OUR QUEST FOR OPTOPIA, IT IS USUALLY OVERLOOKED.

WITHOUT **A11IES** (4), THE DEMOGRAPHIC THAT WAS HOLDING ON TO THE ACCESSIBILITY ROPE IS NOW CUT OFF.

LET'S GET RID OF THE **MAGPIE MENTALITY**?

FOR YOUR NEXT FUN DATA SCIENCE PROJECT, INSTEAD OF SOME COMMUNITY-OVERFITTED IMAGE RECOGNITION CHALLENGE, MAYBE CHOOSE AN **OPEN PROBLEM IN DIGITAL ACCESSIBILITY**, SUCH AS AUTOMATIC VIDEO CAPTIONING. THEN HOPEFULLY ONE DAY THERE WILL BE **"NO MORE CRAPTIONS"** (5)

PART 2: GHOSTS IN THE SHELL

(WHO ARE WE BUILDING MODELS FOR?)

WE HAVEN'T YET FIGURED OUT HOW TO MAKE EXISTING DIGITAL PLATFORMS ACCESSIBLE TO EVERYONE, YET WE'RE ALREADY JUMPING TO FORGE A NEW "INTELLIGENT" CLASS OF WEB APPLICATIONS.

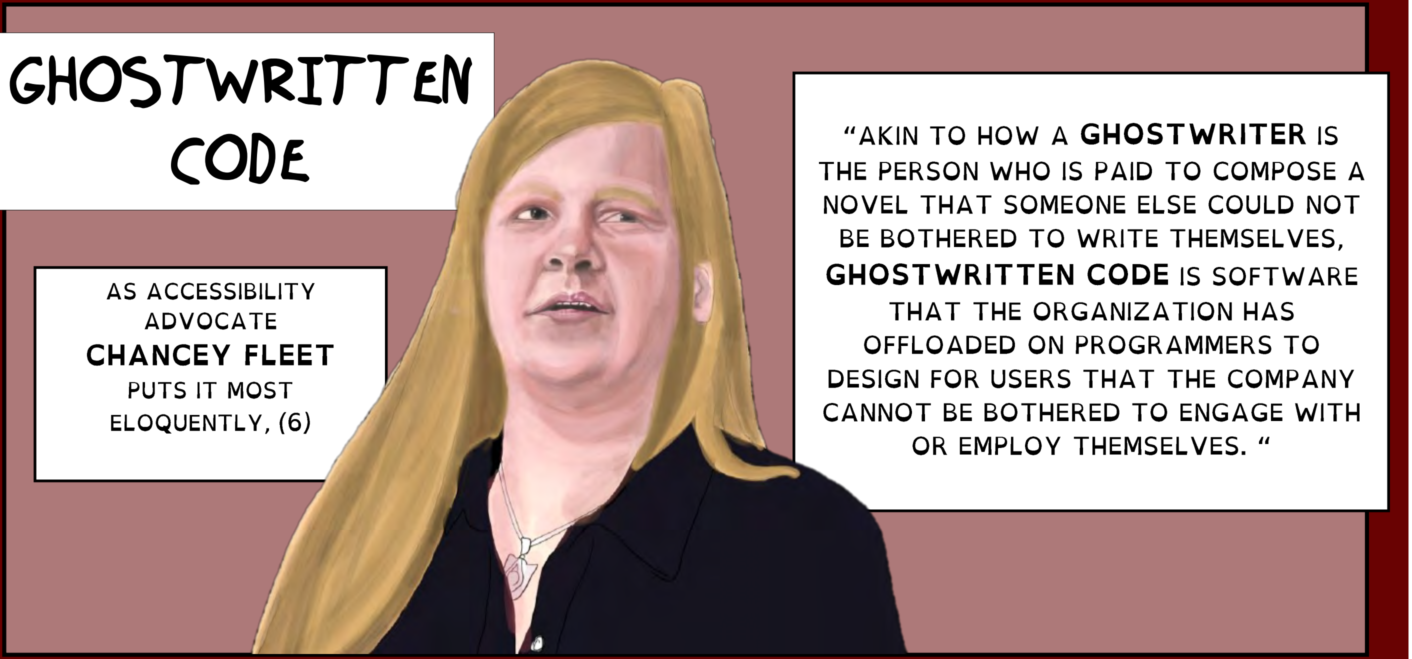
WE'RE SO CAUGHT UP IN THE **"HOW"** (USING ML/AI/DL/DS !!!) THAT WE FORGET TO ASK, **"FOR WHOM''?**

WHEN PLATFORMS ARE NOT DESIGNED FOR EVERYONE, THEY GIVE OFF THE STENCH OF "**ENCODED INHOSPITALITY**" (6).

SEEMINGLY INNOCUOUS THINGS SUCH AS **POP-UPS** AND **EXPIRING FORMS** ON WEBSITES COMPLETELY HIJACK THE ONLINE EXPERIENCE OF USERS WITH DISABILITIES WHO RELY ON SCREEN READERS.

Expectation
89

GHOSTWRITTEN CODE



AS ACCESSIBILITY ADVOCATE **CHANCEY FLEET** PUTS IT MOST ELOQUENTLY, (6)

"AKIN TO HOW A **GHOSTWRITER** IS THE PERSON WHO IS PAID TO COMPOSE A NOVEL THAT SOMEONE ELSE COULD NOT BE BOTHERED TO WRITE THEMSELVES, **GHOSTWRITTEN CODE** IS SOFTWARE THAT THE ORGANIZATION HAS OFFLOADED ON PROGRAMMERS TO DESIGN FOR USERS THAT THE COMPANY CANNOT BE BOTHERED TO ENGAGE WITH OR EMPLOY THEMSELVES. "

THESE GHOSTS ARE MAKING THEIR WAY INTO DATA-DRIVEN PRODUCTS AS WELL.

TAKE THE INFAMOUS FACIAL RECOGNITION SOFTWARE THAT HAS BEEN ALL OVER THE NEWS RECENTLY. RACIAL INJUSTICES ARE PROBLEMATIC ENOUGH, BUT HAVE YOU CONSIDERED HOW THESE MODELS DISCRIMINATE AGAINST BLACK DISABLED PEOPLE?

AS DISABILITY RIGHTS ADVOCATE **HABEN GIRMA** EXPLAINS (7),

"MY EYES MOVE INVOLUNTARILY, EACH ONE SWINGING TO ITS OWN MUSIC. THEY'VE DANCED THIS WAY FOR AS LONG AS I CAN REMEMBER."

HOW WELL DO YOU THINK **FACIAL RECOGNITION** WOULD PERFORM ON **BLIND BLACK PEOPLE**?

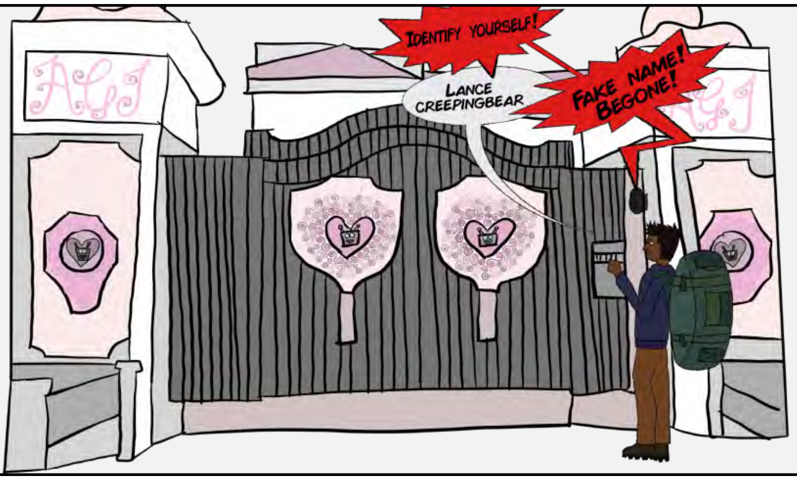
HAVING BEEN TRAINED ON THE FACIAL DYNAMICS OF SIGHTED WHITE PEOPLE, FACIAL RECOGNITION TECHNOLOGY PEDDLES AN ABLEIST AND RACIST NARRATIVE.

THE ATYPICAL, ASYMMETRIC MECHANISMS OF THE EYES OF SOME BLIND PEOPLE ARE PERCEIVED AS ABNORMAL, ANOMALOUS AND THREATENING BY THESE SYSTEMS.

HOW IS IT THAT WE CAN **FORGET** TO CONSIDER **ENTIRE DEMOGRAPHICS** WHILE DESIGNING PRODUCTS?

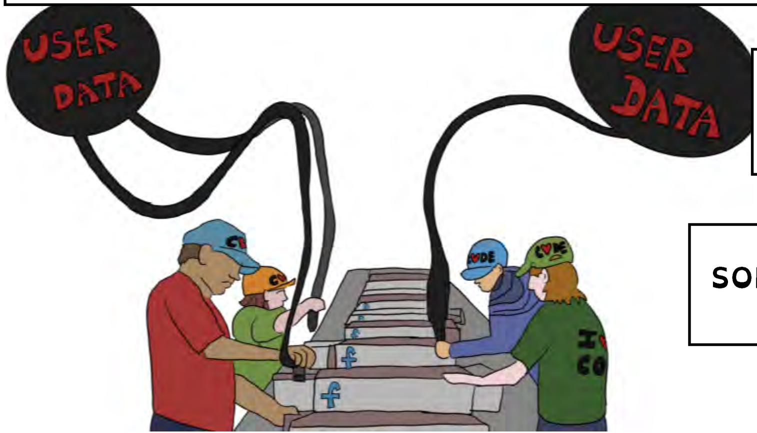
TAKE FACEBOOK'S **"REAL NAME" POLICY** THAT INDISCRIMINATELY TARGETED NATIVE AMERICANS (8)

THE LARGEST SOCIAL NETWORK IN THE WORLD SURE OVERLOOKED THE **CULTURAL AND LINGUISTIC DIFFERENCES** IN NAMES ACROSS THE GLOBE



AND ENDED UP DEPLOYING A BIGOTED ALGORITHM THAT BLOCKED USERS WHOSE NAMES DID NOT CONFORM WITH THE WESTERN ARCHETYPE OF NAMES

IN ADDITION TO COMPLETELY OVERLOOKING **WHO** WE ARE BUILDING A PRODUCT FOR, HAVE WE ALTOGETHER DONE AWAY WITH THE QUESTION OF WHETHER A CERTAIN PRODUCT ***SHOULD*** EVEN BE BUILT?



SURE, YOU HAVE SEVERAL HUNDRED TERABYTES OF USER DATA AND A FLEET OF ENGINEERS WAITING TO DIP THEIR HANDS INTO THE ML PIE,

BUT, IS YOUR PRODUCT A **SOLUTION** TO AN ACTUAL PROBLEM OR SIMPLY **SOLUTIONISM**

PART 3: THE POISONING

(WHAT PROBLEMS ARE WE TRYING TO SOLVE?)

TECHNOLOGY IS SUPPOSED TO DRIVE INNOVATION AND MOVE US TOWARDS A MORE SOPHISTICATED AND ADVANCED FUTURE, RIGHT?

AND SO WHEN THE NEW FLAVOR OF TECHNOLOGICAL ADVANCEMENT COMES TO MARKET, WHAT ELSE MUST WE DO BUT EAGERLY LAP IT UP?

WELL, IF THERE'S ANY MENTION OF "INTELLIGENCE" ON THE PRODUCT BEING HANDED TO YOU...

YOU MIGHT NOT WANT TO DRINK THAT!

IT'S SNAKE OIL!

-**ARVIND NARAYANAN**
PROFESSOR OF COMPUTER SCIENCE
AT PRINCETON UNIVERSITY (9)

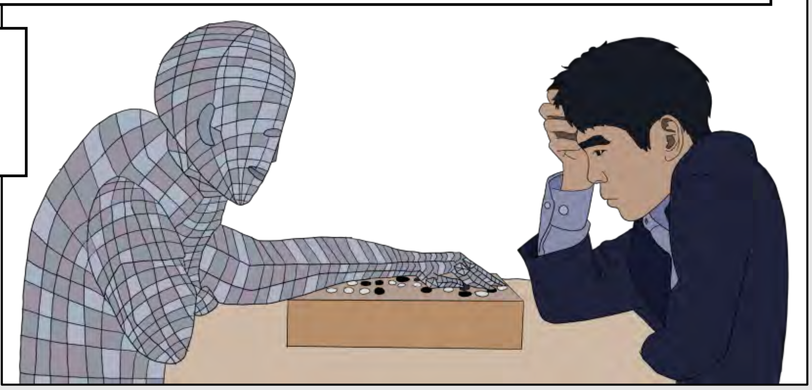
WHAT IS AI-SNAKE OIL?

SNAKE OIL IS THE MYSTICAL SUBSTANCE THAT IS CREATED BY TAKING EQUAL PARTS MEDIA HYPE AND PUBLIC MISINFORMATION AND STIRRING THEM INTO A POTION, WITH AN IRRESISTIBLE LABEL THAT SCREAMS "DATA" AND "INTELLIGENCE"

... AND AFTER YEARS OF EXPERIMENTATION, THE TECH INDUSTRY HAS FINALLY PERFECTED THE RECIPE!

DEVELOPMENTS SUCH AS **ALPHA-GO** (THE GO PLAYING AI) AND **SHAZAM** (THE MUSIC RECOGNITION APP) ARE INDICATIVE OF GENUINE SCIENTIFIC PROGRESS AND DO DEMONSTRABLY MORE GOOD THAN HARM.

WHY? BECAUSE THE RULES OF GO DON'T CHANGE WHETHER THE PLAYER IS **MALE/FEMALE**, **BLACK/WHITE**, **RICH/POOR**!



PERCEPTION TASKS, SUCH AS **FACIAL RECOGNITION**, THAT ARE INTERTWINED WITH THE **SOCIAL, POLITICAL AND CULTURAL UNDERPINNINGS** OF THE DATA ON WHICH THEY WERE TRAINED, ARE FAR MORE TOXIC.

THINGS START TO GET REALLY TOXIC IN SETTINGS SUCH AS **HIRING, MODERATION OF HATE SPEECH OR ALLOCATION OF GRADES** (10), WHEN WE TRY TO IMPOSE OBJECTIVITY (FIT A MATHEMATICAL FUNCTION ONTO THE DATA) ON **HUMAN JUDGMENT**, WHICH IS INHERENTLY SUBJECTIVE

WE GET REALLY CREATIVE WITH WHAT WE THINK WE CAN ACHIEVE WITH TECHNOLOGY WHEN WE START **PREDICTING SOCIAL OUTCOMES** USING ALGORITHMS, SUCH AS **COMPAS FOR CRIMINAL SENTENCING.** (11)

WE LOOK AROUND AND SEE THE HARDEST PROBLEMS KNOWN TO US AND DECIDE THAT, SINCE WE CANNOT SOLVE THEM, WE MUST INSTEAD GET A MACHINE TO DO IT FOR US.

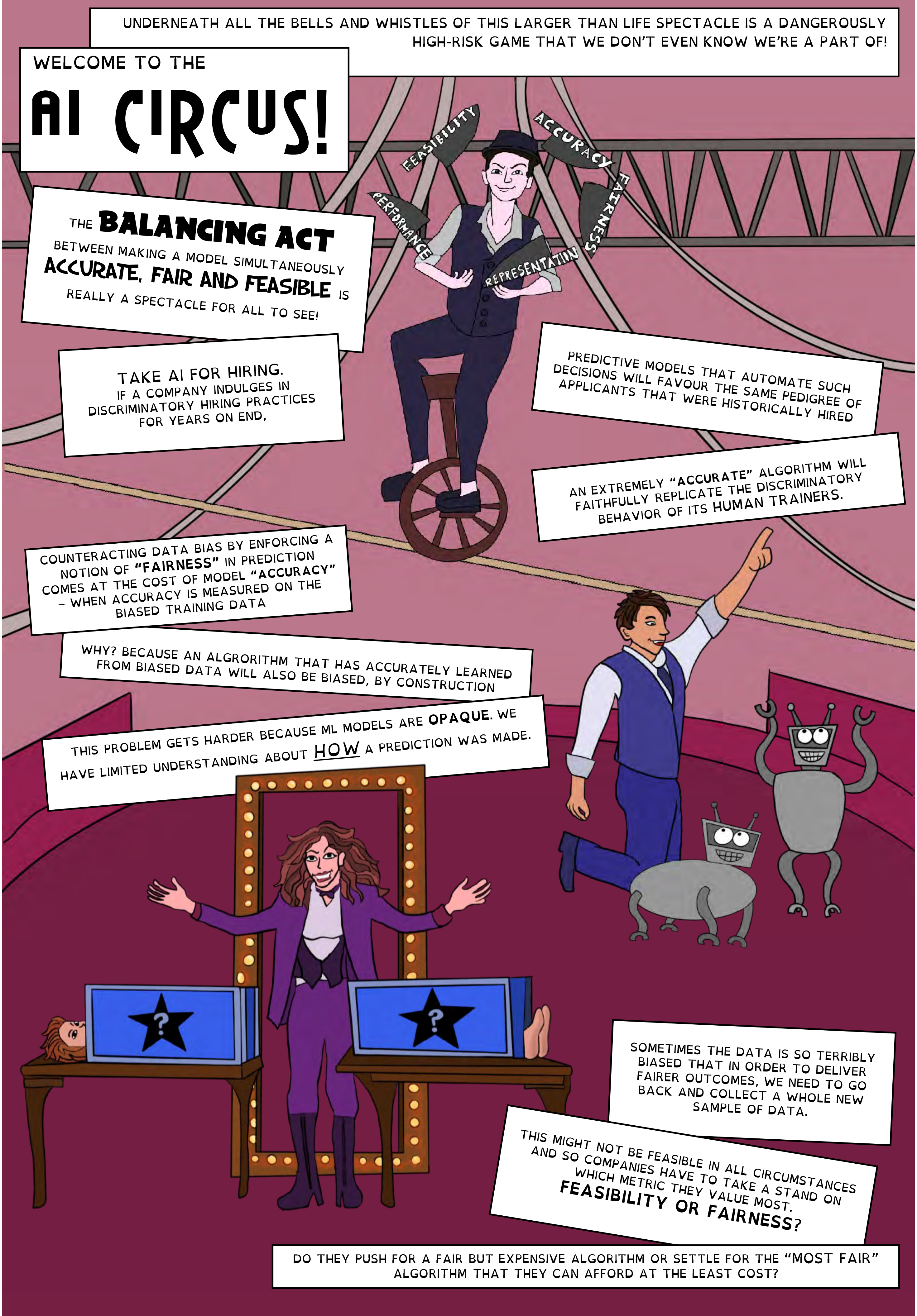
BUT DO YOU KNOW WHY THESE ARE THE HARDEST PROBLEMS TO SOLVE?

BECAUSE THESE ARE SYSTEMIC ISSUES THAT HAVE BEEN SLOWLY STEWING FOR CENTURES OVER

WITH A DASH OF HISTORICAL CONTEXT, A SPRINKLE OF CULTURE AND A GENEROUS HEAPING OF RACE, GENDER AND CLASS POLITICS

ALL COMPOUNDING INTO A COMPLEX BROTH OF ENTROPY;

EXPECTING A MACHINE TO TAKE ONE WHiFF OF THIS STEW AND BE ABLE TO PREDICT THE FUTURE IS JUST FUNDAMENTALLY DUBIOUS.



UNDERNEATH ALL THE BELLS AND WHISTLES OF THIS LARGER THAN LIFE SPECTACLE IS A DANGEROUSLY HIGH-RISK GAME THAT WE DON'T EVEN KNOW WE'RE A PART OF!

WELCOME TO THE AI CIRCUS!

THE **BALANCING ACT** BETWEEN MAKING A MODEL SIMULTANEOUSLY **ACCURATE, FAIR AND FEASIBLE** IS REALLY A SPECTACLE FOR ALL TO SEE!

TAKE AI FOR HIRING. IF A COMPANY INDULGES IN DISCRIMINATORY HIRING PRACTICES FOR YEARS ON END,

PREDICTIVE MODELS THAT AUTOMATE SUCH DECISIONS WILL FAVOUR THE SAME PEDIGREE OF APPLICANTS THAT WERE HISTORICALLY HIRED

AN EXTREMELY "**ACCURATE**" ALGORITHM WILL FAITHFULLY REPLICATE THE DISCRIMINATORY BEHAVIOR OF ITS HUMAN TRAINERS.

COUNTERACTING DATA BIAS BY ENFORCING A NOTION OF **"FAIRNESS"** IN PREDICTION COMES AT THE COST OF MODEL **"ACCURACY"** – WHEN ACCURACY IS MEASURED ON THE BIASED TRAINING DATA

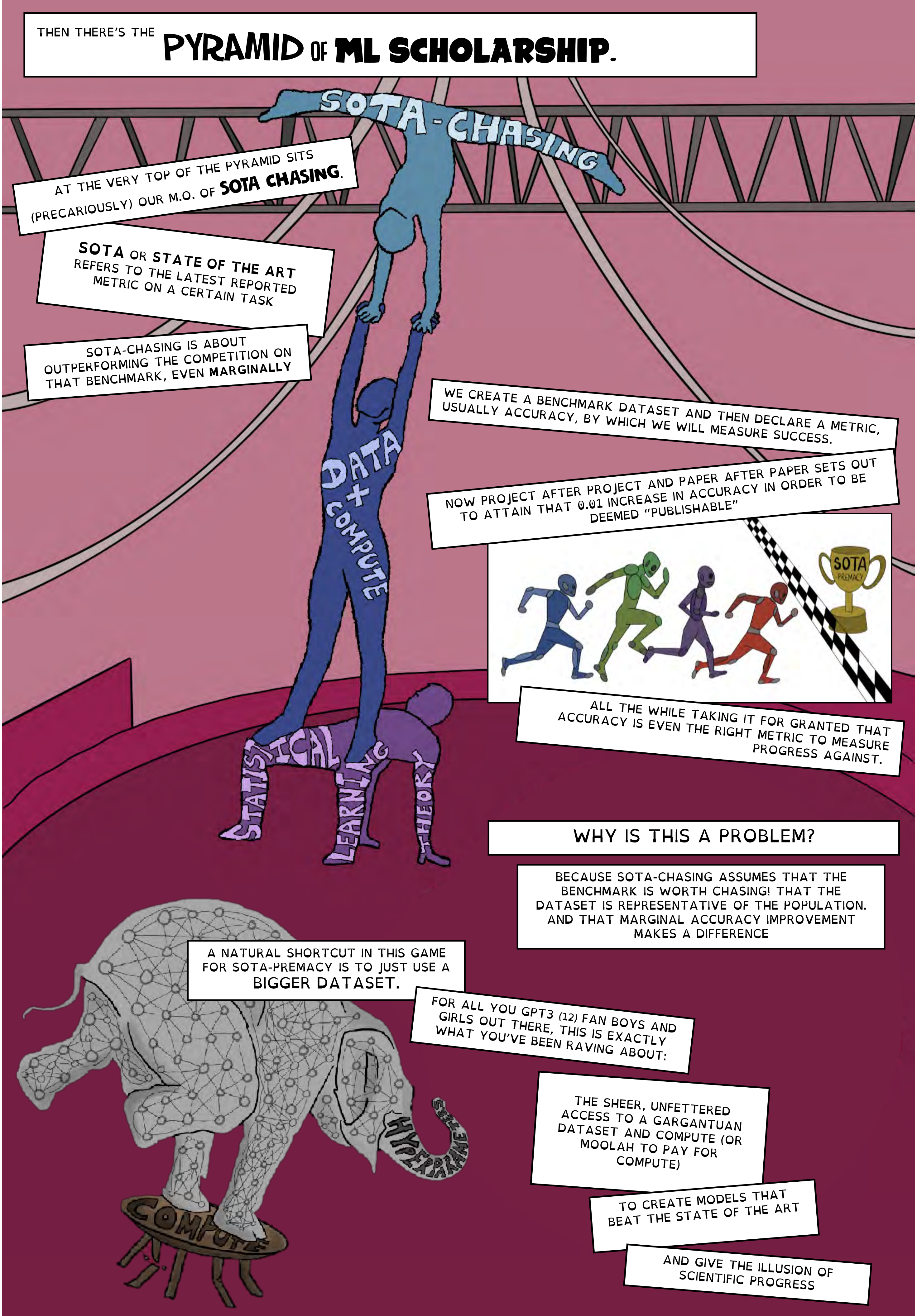
WHY? BECAUSE AN ALGRORITHM THAT HAS ACCURATELY LEARNED FROM BIASED DATA WILL ALSO BE BIASED, BY CONSTRUCTION

THIS PROBLEM GETS HARDER BECAUSE ML MODELS ARE **OPAQUE**. WE HAVE LIMITED UNDERSTANDING ABOUT HOW A PREDICTION WAS MADE.

SOMETIMES THE DATA IS SO TERRIBLY BIASED THAT IN ORDER TO DELIVER FAIRER OUTCOMES, WE NEED TO GO BACK AND COLLECT A WHOLE NEW SAMPLE OF DATA.

THIS MIGHT NOT BE FEASIBLE IN ALL CIRCUMSTANCES AND SO COMPANIES HAVE TO TAKE A STAND ON WHICH METRIC THEY VALUE MOST. **FEASIBILITY OR FAIRNESS?**

DO THEY PUSH FOR A FAIR BUT EXPENSIVE ALGORITHM OR SETTLE FOR THE "MOST FAIR" ALGORITHM THAT THEY CAN AFFORD AT THE LEAST COST?



THEN THERE'S THE **PYRAMID OF ML SCHOLARSHIP.**

AT THE VERY TOP OF THE PYRAMID SITS (PRECARIOUSLY) OUR M.O. OF **SOTA CHASING**.

SOTA OR **STATE OF THE ART** REFERS TO THE LATEST REPORTED METRIC ON A CERTAIN TASK

SOTA-CHASING IS ABOUT OUTPERFORMING THE COMPETITION ON THAT BENCHMARK, EVEN **MARGINALLY**

WE CREATE A BENCHMARK DATASET AND THEN DECLARE A METRIC, USUALLY ACCURACY, BY WHICH WE WILL MEASURE SUCCESS.

NOW PROJECT AFTER PROJECT AND PAPER AFTER PAPER SETS OUT TO ATTAIN THAT 0.01 INCREASE IN ACCURACY IN ORDER TO BE DEEMED "PUBLISHABLE"

ALL THE WHILE TAKING IT FOR GRANTED THAT ACCURACY IS EVEN THE RIGHT METRIC TO MEASURE PROGRESS AGAINST.

WHY IS THIS A PROBLEM?

BECAUSE SOTA-CHASING ASSUMES THAT THE BENCHMARK IS WORTH CHASING! THAT THE DATASET IS REPRESENTATIVE OF THE POPULATION. AND THAT MARGINAL ACCURACY IMPROVEMENT MAKES A DIFFERENCE

A NATURAL SHORTCUT IN THIS GAME FOR SOTA-PREMACY IS TO JUST USE A BIGGER DATASET.

FOR ALL YOU GPT3 (12) FAN BOYS AND GIRLS OUT THERE, THIS IS EXACTLY WHAT YOU'VE BEEN RAVING ABOUT:

THE SHEER, UNFETTERED ACCESS TO A GARGANTUAN DATASET AND COMPUTE (OR MOOLAH TO PAY FOR COMPUTE)

TO CREATE MODELS THAT BEAT THE STATE OF THE ART

AND GIVE THE ILLUSION OF SCIENTIFIC PROGRESS

SURE, THERE ARE THOSE FOLKS IN THE COMMUNITY WHO ARE THINKING DEEPLY ABOUT PROBLEM FORMULATION, REAL WORLD IMPACT AND SCIENTIFIC RIGOR. UNFORTUNATELY, DEEP, THOUGHTFUL WORK OF THIS KIND IS JUST NOT GLAMOROUS

...AND SO, WHEN THE CURTAIN FALLS, IT ISN'T THESE RESEARCHERS YOU ARE APPLAUDING.

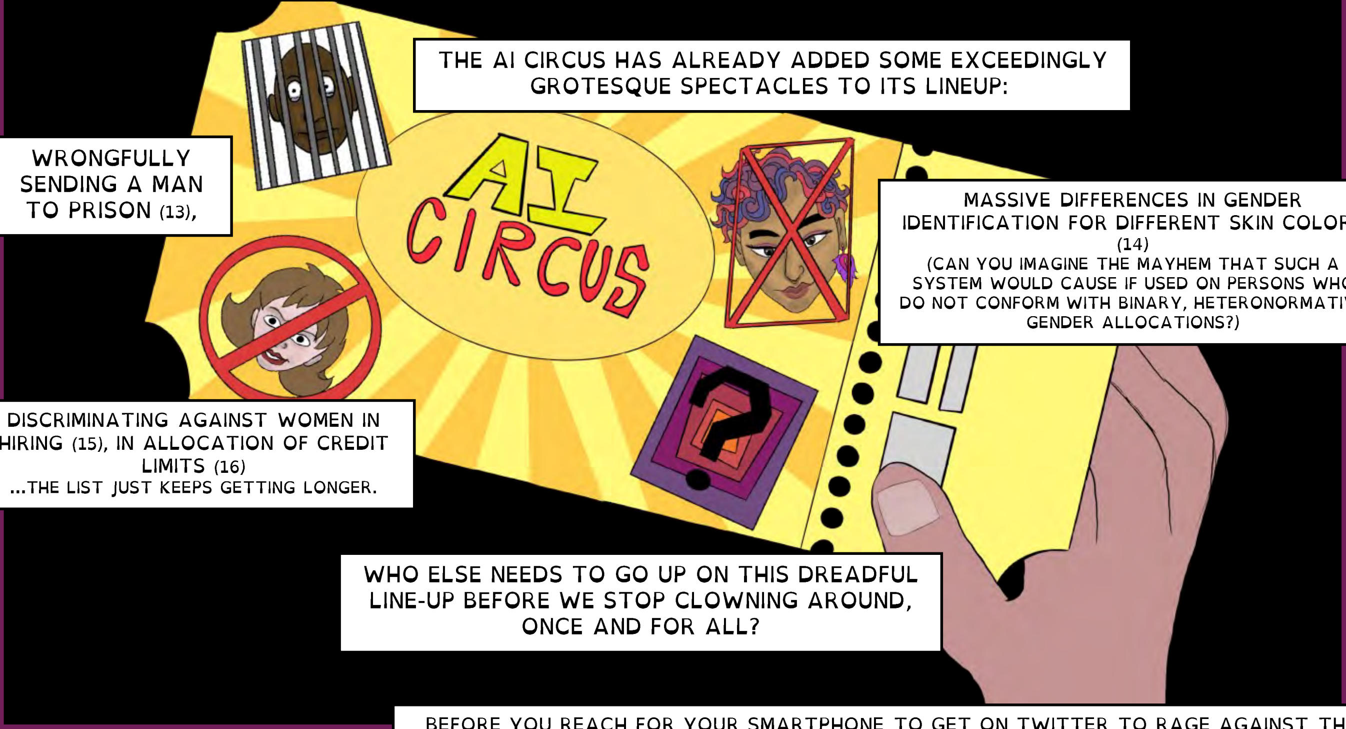
HOW COME THESE FOLKS NEVER TAKE CENTER STAGE?
WELL, IT'S PARTLY BECAUSE, LIKE IN EVERY OTHER DOMAIN, THE RICH JUST KEEP GETTING RICHER.

THE SET OF RESEARCHERS WHO DEBUNK SOCIETAL HARMS OF TECHNOLOGY ARE LIKELY TO BE FROM THE SAME DEMOGRAPHIC THAT WILL BE MOST DEEPLY AFFECTED BY THOSE VERY HARMS.

AND THIS IS NEVER THE MAJORITY.

IF OUR SCHOLARSHIP IS A REFLECTION OF OUR IDEAS, THEN WE CANNOT AFFORD TO CENSOR OR COMPLETELY ERASE THE VOICES OF ENTIRE DEMOGRAPHICS.

IF OUR PRODUCTS ARE A REFLECTION OF THE PROBLEMS THAT WE ARE TRYING TO SOLVE, THEN WE CANNOT BUILD SOLUTIONS THAT HELP ONE STRATUM AND CAUSE EXTENSIVE DAMAGE TO ANOTHER.



THE AI CIRCUS HAS ALREADY ADDED SOME EXCEEDINGLY GROTESQUE SPECTACLES TO ITS LINEUP:

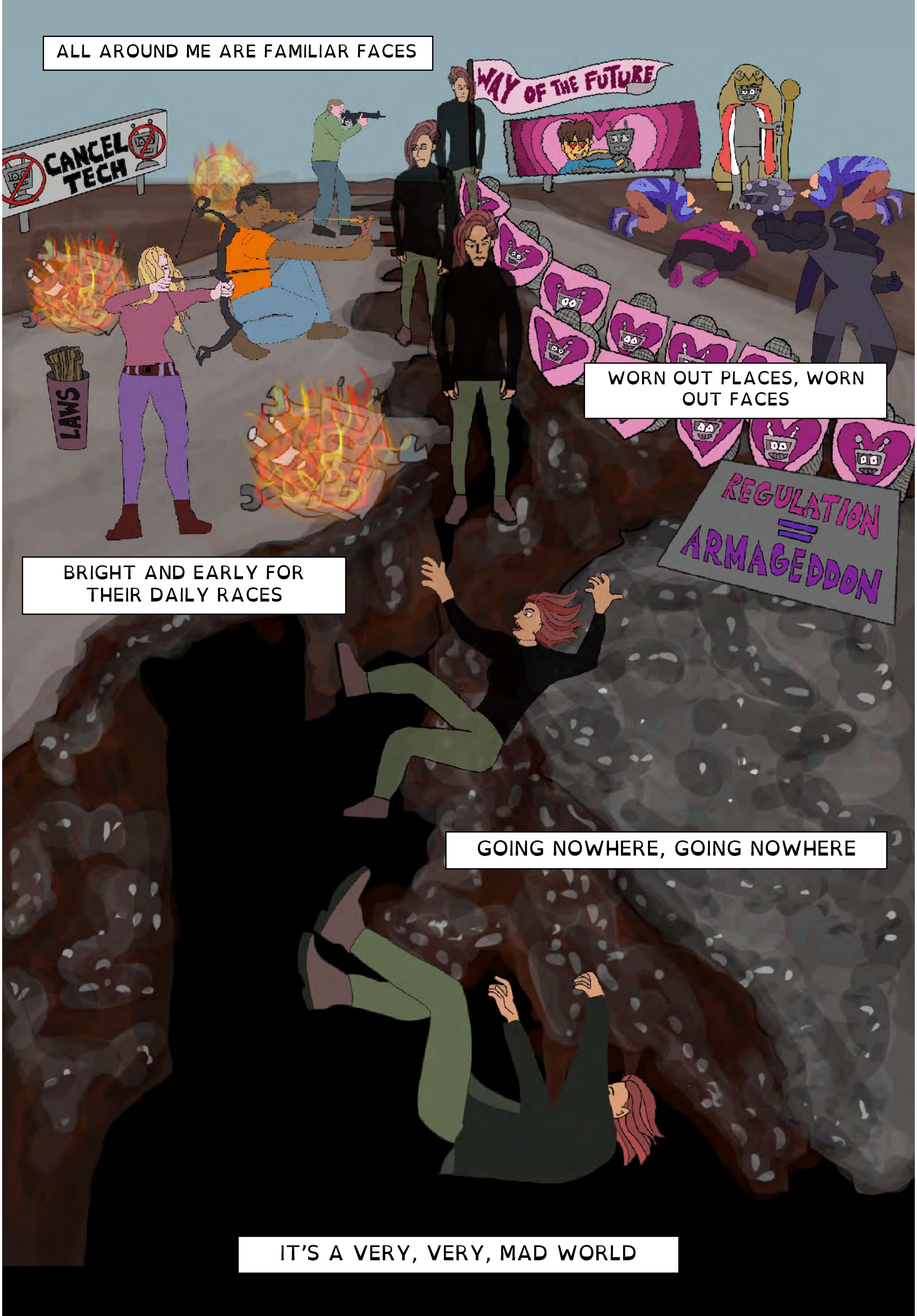
WRONGFULLY SENDING A MAN TO PRISON (13),

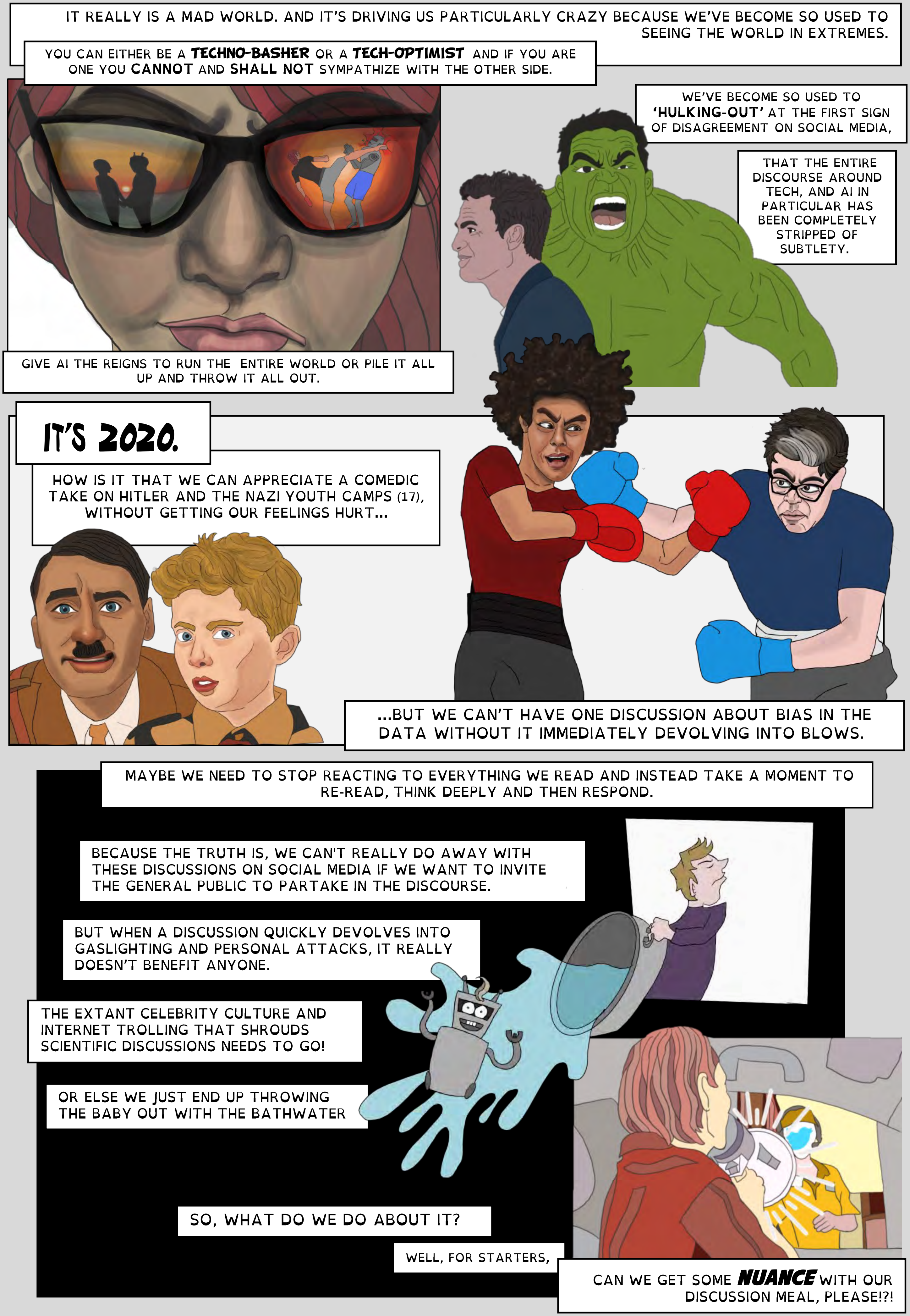
MASSIVE DIFFERENCES IN GENDER IDENTIFICATION FOR DIFFERENT SKIN COLORS (14)
(CAN YOU IMAGINE THE MAYHEM THAT SUCH A SYSTEM WOULD CAUSE IF USED ON PERSONS WHO DO NOT CONFORM WITH BINARY, HETERONORMATIVE GENDER ALLOCATIONS?)

DISCRIMINATING AGAINST WOMEN IN HIRING (15), IN ALLOCATION OF CREDIT LIMITS (16)
...THE LIST JUST KEEPS GETTING LONGER.

WHO ELSE NEEDS TO GO UP ON THIS DREADFUL LINE-UP BEFORE WE STOP CLOWNING AROUND, ONCE AND FOR ALL?

BEFORE YOU REACH FOR YOUR SMARTPHONE TO GET ON TWITTER TO RAGE AGAINST THE AI MACHINE OR JOIN THE RANKS OF THE TECHNO BASHERS, STOP AND LOOK AROUND


ALL AROUND ME ARE FAMILIAR FACES
CANCEL TECH
WAY OF THE FUTURE
LAWS
WORN OUT PLACES, WORN OUT FACES
REGULATION = ARMAGEDDON
BRIGHT AND EARLY FOR THEIR DAILY RACES
GOING NOWHERE, GOING NOWHERE
IT'S A VERY, VERY, MAD WORLD


IT REALLY IS A MAD WORLD. AND IT'S DRIVING US PARTICULARLY CRAZY BECAUSE WE'VE BECOME SO USED TO SEEING THE WORLD IN EXTREMES.
YOU CAN EITHER BE A **TECHNO-BASHER** OR A **TECH-OPTIMIST** AND IF YOU ARE ONE YOU **CANNOT** AND **SHALL NOT** SYMPATHIZE WITH THE OTHER SIDE.
GIVE AI THE REIGNS TO RUN THE ENTIRE WORLD OR PILE IT ALL UP AND THROW IT ALL OUT.
WE'VE BECOME SO USED TO **'HULKING-OUT'** AT THE FIRST SIGN OF DISAGREEMENT ON SOCIAL MEDIA,
THAT THE ENTIRE DISCOURSE AROUND TECH, AND AI IN PARTICULAR HAS BEEN COMPLETELY STRIPPED OF SUBTLETY.
IT'S 2020.
HOW IS IT THAT WE CAN APPRECIATE A COMEDIC TAKE ON HITLER AND THE NAZI YOUTH CAMPS (17), WITHOUT GETTING OUR FEELINGS HURT...
...BUT WE CAN'T HAVE ONE DISCUSSION ABOUT BIAS IN THE DATA WITHOUT IT IMMEDIATELY DEVOLVING INTO BLOWS.
MAYBE WE NEED TO STOP REACTING TO EVERYTHING WE READ AND INSTEAD TAKE A MOMENT TO RE-READ, THINK DEEPLY AND THEN RESPOND.
BECAUSE THE TRUTH IS, WE CAN'T REALLY DO AWAY WITH THESE DISCUSSIONS ON SOCIAL MEDIA IF WE WANT TO INVITE THE GENERAL PUBLIC TO PARTAKE IN THE DISCOURSE.
BUT WHEN A DISCUSSION QUICKLY DEVOLVES INTO GASLIGHTING AND PERSONAL ATTACKS, IT REALLY DOESN'T BENEFIT ANYONE.
THE EXTANT CELEBRITY CULTURE AND INTERNET TROLLING THAT SHROUDS SCIENTIFIC DISCUSSIONS NEEDS TO GO!
OR ELSE WE JUST END UP THROWING THE BABY OUT WITH THE BATHWATER
SO, WHAT DO WE DO ABOUT IT?
WELL, FOR STARTERS,
CAN WE GET SOME ***NUANCE*** WITH OUR DISCUSSION MEAL, PLEASE!?!

HERE IS A MORE NUANCED TAKE ON WHETHER AI LEADS TO A **UTOPIA** OR A ***DYSTOPIA:***

FOR STARTERS, **THERE IS RARELY AN *OBJECTIVE TRUTH*!** MORE OFTEN THAN NOT, THE **EFFICACY** OF A MODEL DEPENDS ON THE **CONTEXT** FOR WHICH IT WAS DESIGNED

THE "**GROUND TRUTH**" THAT WE PRETEND EXISTS, AND AGAINST WHICH WE MEASURE MODEL ACCURACY, IS JUST THE **CLOTHES** THAT THE **ML EMPEROR** IS **NOT** WEARING!

THE ENGINEERING MINDSET IS TO TAKE THE CLASS LABELS AS GOSPEL AND BLINDLY TRY TO OPTIMIZE FOR THEM.

BUT CLASS LABELS ARE JUST **PROXIES** FOR **UNDERLYING SOCIAL PHENOMENA** AND NO AMOUNT OF **MATHEMATICAL FORMALIZATION** WILL TURN **SOCIAL CONSTRUCTS** INTO **OBJECTIVE TRUTHS.**

THE REALITY IS THAT **ALL** MODELS ARE **WRONG**. **SOME** MODELS ARE ***USEFUL***!

IN THIS ART GALLERY, EACH PAINTING DEPICTS AN **APPLE.** BUT ONLY ONE OF THEM IS POTENTIALLY USEFUL AS A **REAL-LIFE APPLE DETECTOR**

WE OFTEN FIND IT HARD TO JUDGE WHICH MODEL IS MOST USEFUL, BECAUSE THAT REQUIRES DEEP **DOMAIN EXPERTISE.**

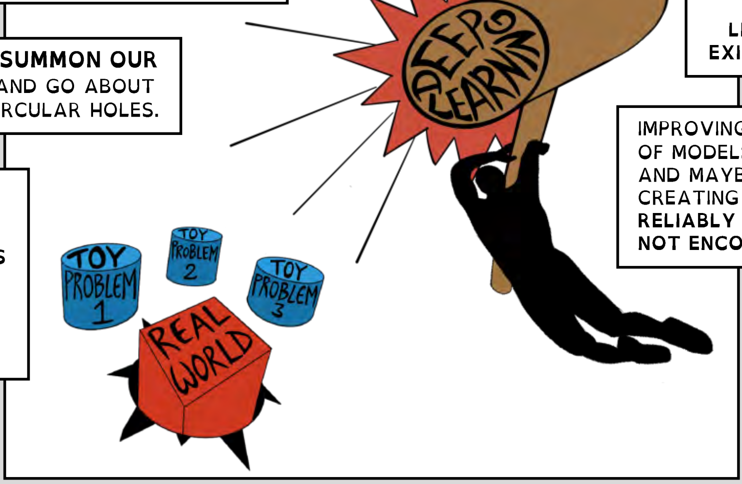
WE HAVE BEEN **DANGEROUSLY CONFLATING** EXPERTISE IN TRAINING AND DEPLOYING A MODEL WITH DOMAIN EXPERTISE.

INSTEAD WE SHOULD **ACKNOWLEDGE THE LIMITATION OF OUR EXPERTISE** AS SCIENTISTS AND ENGINEERS AND **INVITE THE TRUE DOMAIN EXPERTS** TO COME TO THE TABLE.

SOME **CONTEXTS** ARE ***INHERENTLY DIFFICULT*** TO BUILD FOR.

WE HAVE THE TENDENCY **TO SUMMON OUR DEEP LEARNING HAMMER** AND GO ABOUT NAILING SQUARE PEGS INTO CIRCULAR HOLES.

UNFORTUNATELY, THE MOST PROMISING RESULTS THAT YOU READ ABOUT WERE OBTAINED ON **TOY PROBLEMS** WITHIN **EXPERIMENTAL SET-UPS** AND ARE NOT DESIGNED TO SCALE TO THE REAL WORLD.



THE WORLD IS A COMPLICATED AND MESSY PLACE AND THE **LIMITED PERFORMANCE OF OUR EXISTING MODELS** REFLECTS THAT.

IMPROVING **GENERALIZATION ABILITY** OF MODELS IS A HOT AREA OF RESEARCH AND MAYBE WE'LL GET AROUND TO CREATING MODELS THAT CAN **PERFORM RELIABLY** IN **CONTEXTS THAT THEY DID NOT ENCOUNTER DURING TRAINING.**

BUT WE AREN'T THERE YET.

THE OVERWHELMING MAJORITY OF PROBLEMS THAT PLAGUE AI TODAY ARE NOT BECAUSE OF *JUST* THE DATA OR *JUST* THE ALGORITHM IN ITSELF

BUT BECAUSE OF ONE CRITICAL **CONFOUNDING FACTOR** THAT WE KEEP OVERLOOKING:

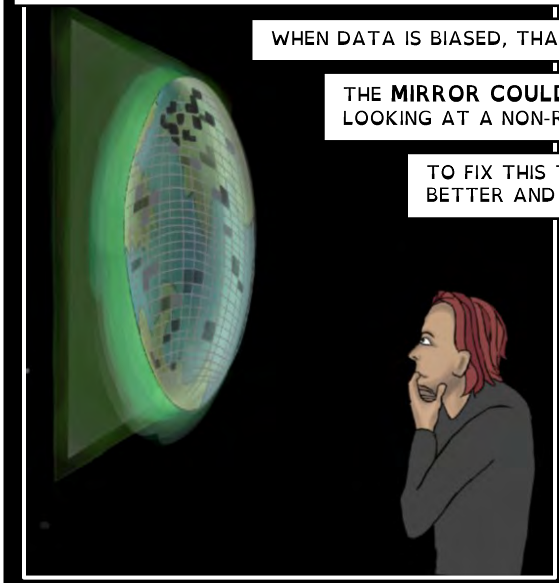
THE WORLD

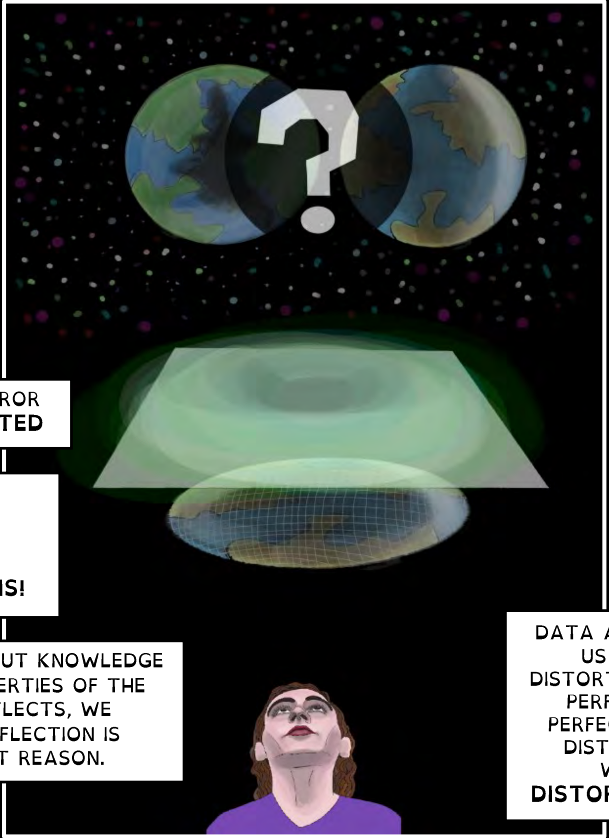
DATA IS A MIRROR REFLECTION OF THE WORLD (18)

WHEN DATA IS BIASED, THAT REFLECTION IS DISTORTED. THERE ARE SEVERAL EXPLANATIONS FOR THIS

THE **MIRROR COULD BE DISTORTED**: WE COULD BE COLLECTING THE WRONG DATA, OR LOOKING AT A NON-REPRESENTATIVE SAMPLE

TO FIX THIS TYPE OF BIAS, WE CAN ATTEMPT FIXING THE MIRROR TO COLLECT BETTER AND CLEANER DATA





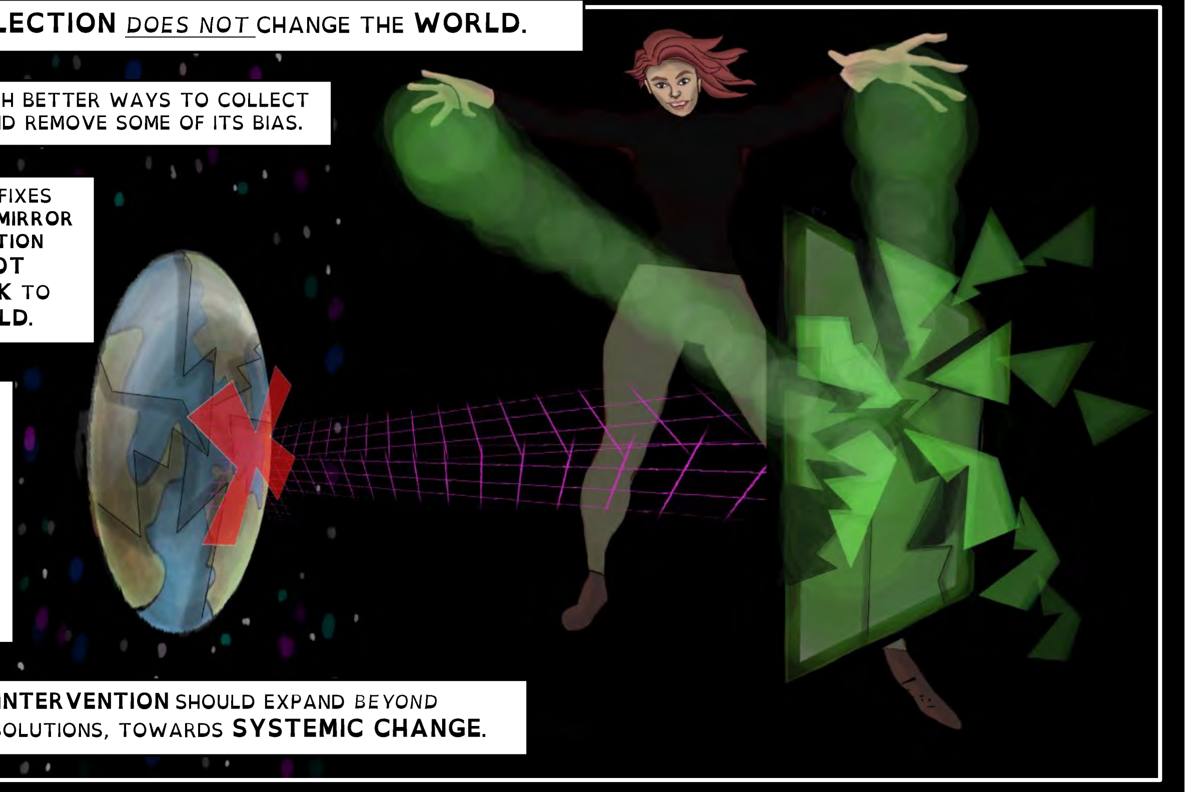
BUT THERE'S ALSO THE POSSIBILITY THAT THE MIRROR IS PERFECT AND **THE WORLD ITSELF IS DISTORTED**

WE TEND TO UNDER-APPRECIATE THIS POSSIBILITY BECAUSE WE INSTINCTIVELY COMPARE THE REFLECTION (DATA) WITH **HOW WE WANT THE WORLD TO BE**, RATHER THAN WITH **HOW IT ACTUALLY IS!**

BASED ON THE REFLECTION, AND WITHOUT KNOWLEDGE OR **ASSUMPTIONS** ABOUT THE PROPERTIES OF THE MIRROR AND OF THE WORLD IT REFLECTS, WE **CANNOT KNOW** WHETHER THE REFLECTION IS DISTORTED, AND, IF SO, FOR WHAT REASON.

DATA ALONE **CANNOT** TELL US WHETHER IT IS A DISTORTED REFLECTION OF A PERFECT WORLD, OR A PERFECT REFLECTION OF A DISTORTED WORLD, OR WHETHER THESE **DISTORTIONS COMPOUND.**

CHANGING THE **REFLECTION** <u>DOES NOT</u> CHANGE THE **WORLD**.



WE'VE COME UP WITH BETTER WAYS TO COLLECT DATA, CLEAN IT AND REMOVE SOME OF ITS BIAS.

BUT, ALL OF THESE FIXES ARE **APPLIED ON THE MIRROR** OR ON THE **REFLECTION** AND THEY **DO NOT PROPAGATE BACK** TO CHANGE THE **WORLD.**

THE **UNDERLYING SOCIETAL INEQUITIES** THAT GIVE RISE TO DISCRIMINATORY OUTCOMES **REMAIN INTACT** IF WE ONLY **INTERVENE** ON **THE DATA.**

HENCE, OUR **INTERVENTION** SHOULD EXPAND *BEYOND* TECHNOLOGICAL SOLUTIONS, TOWARDS **SYSTEMIC CHANGE.**

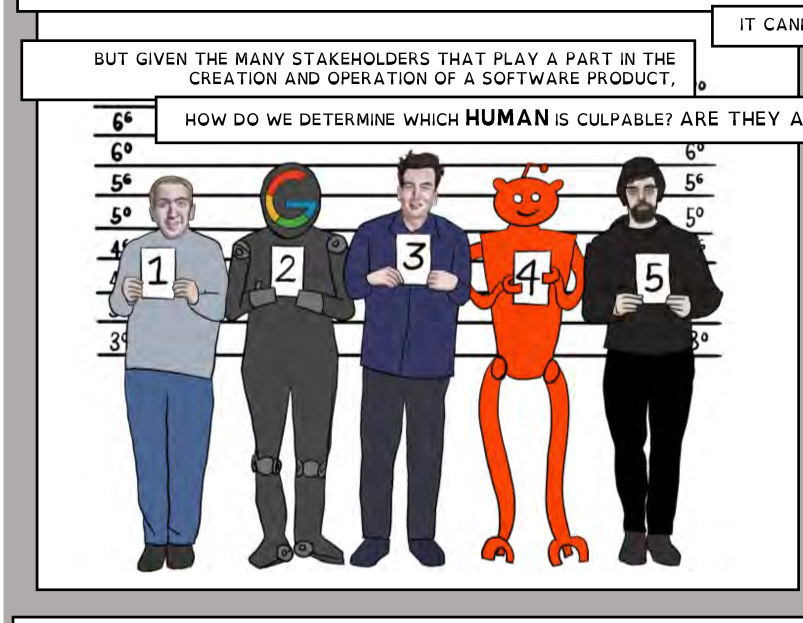
PREDICTION IS DIFFICULT, ESPECIALLY ABOUT THE ***FUTURE***! (19)

WHEN THINGS (INEVITABLY) GO WRONG, WHO IS **RESPONSIBLE**?

IT CANNOT BE THE ALGORITHM.

BUT GIVEN THE MANY STAKEHOLDERS THAT PLAY A PART IN THE CREATION AND OPERATION OF A SOFTWARE PRODUCT,

HOW DO WE DETERMINE WHICH **HUMAN** IS CULPABLE? ARE THEY ALL?



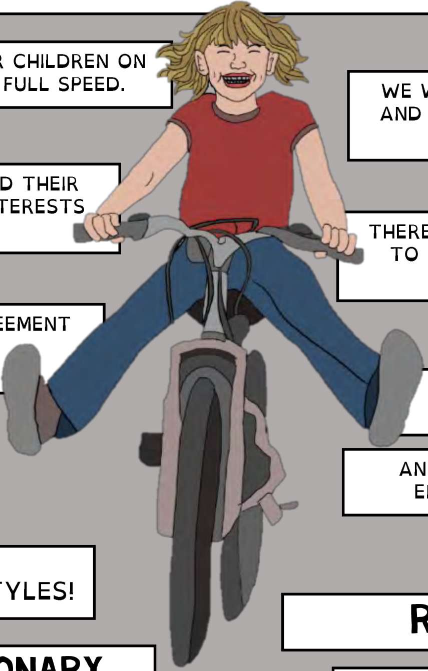
I KNOW WHAT YOU'RE THINKING...

"I SEE WHERE YOU'RE GOING WITH THIS... YOU'RE NOT SERIOUSLY GOING TO GET INTO REGULATION NOW, ARE YOU?"

WELL...TIME TO REMIND YOU OF OUR RECOMMENDED APPROACH TO THINKING ABOUT AI.

REMEMBER, **NUANCE**?!

RIGHT NOW, SILICON VALLEY WILL HAVE YOU BELIEVE THAT TECHNOLOGY NEEDS TO BE ALLOWED TO RUN FREE. REGULATION IS A CATASTROPHE OF COSMIC PROPORTIONS AND WOULD BE THE END OF THE INTERNET, AND BY EXTENSION, INNOVATION AND PROGRESS.



THE FACT OF THE MATTER IS, WE PUT OUR CHILDREN ON THE AI HYPE-BIKE AND SENT THEM OFF AT FULL SPEED.

WE WERE TOO BRASH IN OUR RAPID ADOPTION OF AI AND IT HAS LED TO SOME TERRIBLE OUTCOMES WITH VERY REAL IMPACTS ON PEOPLE'S LIVES.

AND SO WHILE TECH COMPANIES AND THEIR CELEBRITY CEOS PROTECT THEIR INTERESTS BY BAD MOUTHING REGULATION,

THERE'S REALLY NO EXCUSE FOR THE GENERAL PUBLIC TO BUY INTO THIS NARRATIVE AND BE COMPLICIT IN THE VANDALISM OF OUR MORAL SOCIAL FIBER.

WE NEED TO COME TO AN AGREEMENT ON HOW TO GO ABOUT REGULATING TECHNOLOGY.

AND SO WE MUST START EDUCATING OURSELVES,

AND PARTAKE IN THIS LOFTY ENTERPRISE IN GOOD FAITH.

MAYBE IT'S TIME TO CONSIDER OTHER PARENTING STYLES!

PRECAUTIONARY

V/S

RISK-BASED

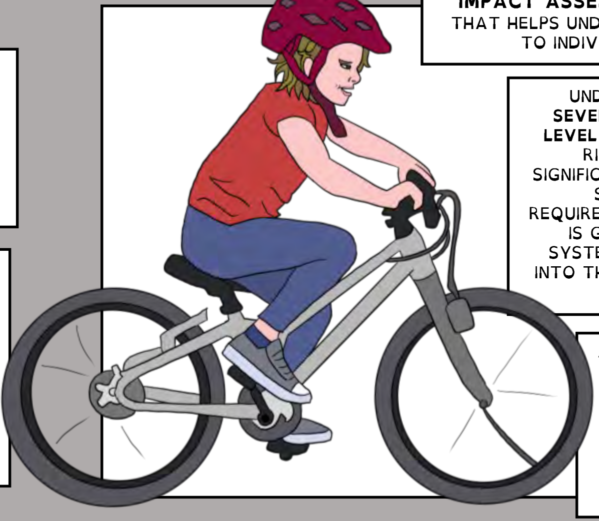
THINK OF THE OLD ADAGE **"IT'S BETTER TO BE SAFE THAN TO BE SORRY"**

THIS PRINCIPLE CALLS FOR CAUTION IN SITUATIONS OF **UNCERTAIN HARMS**, IE. THOSE THAT HAVE NOT BEEN SCIENTIFICALLY STUDIED YET.

A COMMON CRITICISM OF THIS APPROACH IS IT IS **"PARALYZING"** AND **"SELF-CANCELING"**, SINCE ANY NEW TECHNOLOGY IN ITS EARLY STAGES OF ADOPTION WOULD HAVE RISKS THAT CANNOT BE ACCOUNTED FOR.

UNDER THIS PARADIGM, REGULATE BASED ON **KNOWN** RISKS, AND MODEL THE **LIKELIHOOD** THAT THESE **RISKS** WILL LEAD TO **HARMS**

A PROMISING APPROACH IS **ALGORITHMIC IMPACT ASSESSMENT (AIA)** - A FRAMEWORK THAT HELPS UNDERSTAND AND REDUCE THE RISKS TO INDIVIDUALS AND COMMUNITIES



UNDER AIA, THE **LIKELIHOOD AND SEVERITY OF HARM** DETERMINES THE **LEVEL OF OVERSIGHT**. THE HIGHER THE RISK OF HARM, AND THE MORE SIGNIFICANT THE HARM ITSELF, THE MORE STRINGENT THE OVERSIGHT REQUIREMENTS. AND THE LESS AUTONOMY IS GRANTED TO THE AUTOMATED SYSTEM: A HUMAN MUST BE BROUGHT INTO THE LOOP TO TAKE RESPONSIBILITY FOR IMPACTFUL DECISIONS

AIA WILL ONLY WORK IF THE RISKS ARE KNOWN. THIS GIVES EACH AND EVERY ONE OF US THE OPPORTUNITY TO **BE A PART OF THE CHANGE**! NOW'S THE TIME TO **GET INVOLVED IN PUBLIC CONSULTATIONS**, TO MAKE YOUR CONCERNS HEARD!

IF WE WANT OUR ATTEMPTS AT REGULATION TO BE *TRULY EFFECTIVE*, WE NEED TO **RECONCILE** SOME INHERENT DISAGREEMENTS BETWEEN TECH AND LAW.

FOR STARTERS, HOW DO WE MAKE SURE THE LAW **KEEPS UP** WITH THE **RAPIDLY EVOLVING** SOCIO-TECHNOLOGICAL LANDSCAPE?

ANOTHER MAJOR PROBLEM IS ***HOW*** DO WE REGULATE?

NOTIONS SUCH AS **FAIRNESS, ACCOUNTABILITY AND INTERPRETABILITY** HAVE BECOME THE POSTER CHILDREN FOR AI POLICY. BUT THEY STILL DON'T HAVE UNIVERSALLY ACCEPTED TECHNICAL MANIFESTATIONS.

WHY? BECAUSE AMBIGUITY IN DEFINITIONS IS AN INTENTIONALLY WIELDED TOOL THAT ALLOWS FOR **INTERPRETIVE** AND **CONTEXTUAL** READINGS OF **LAW**

BUT THE VERY SAME AMBIGUITY IS CATASTROPHIC FOR **TECH**, WHICH RELIES ENTIRELY ON **MATHEMATICAL FORMALIZATIONS THAT CAN BE WRITTEN INTO CODE**

AND FOR **REGULATORS** WHO NEED **PRECISE DEFINITIONS** TO BUILD RULES AND POLICIES

TO COME UP WITH GOOD DEFINITIONS, WE NEED EXAMPLES OF SYSTEMS THAT ARE USED **TODAY**!

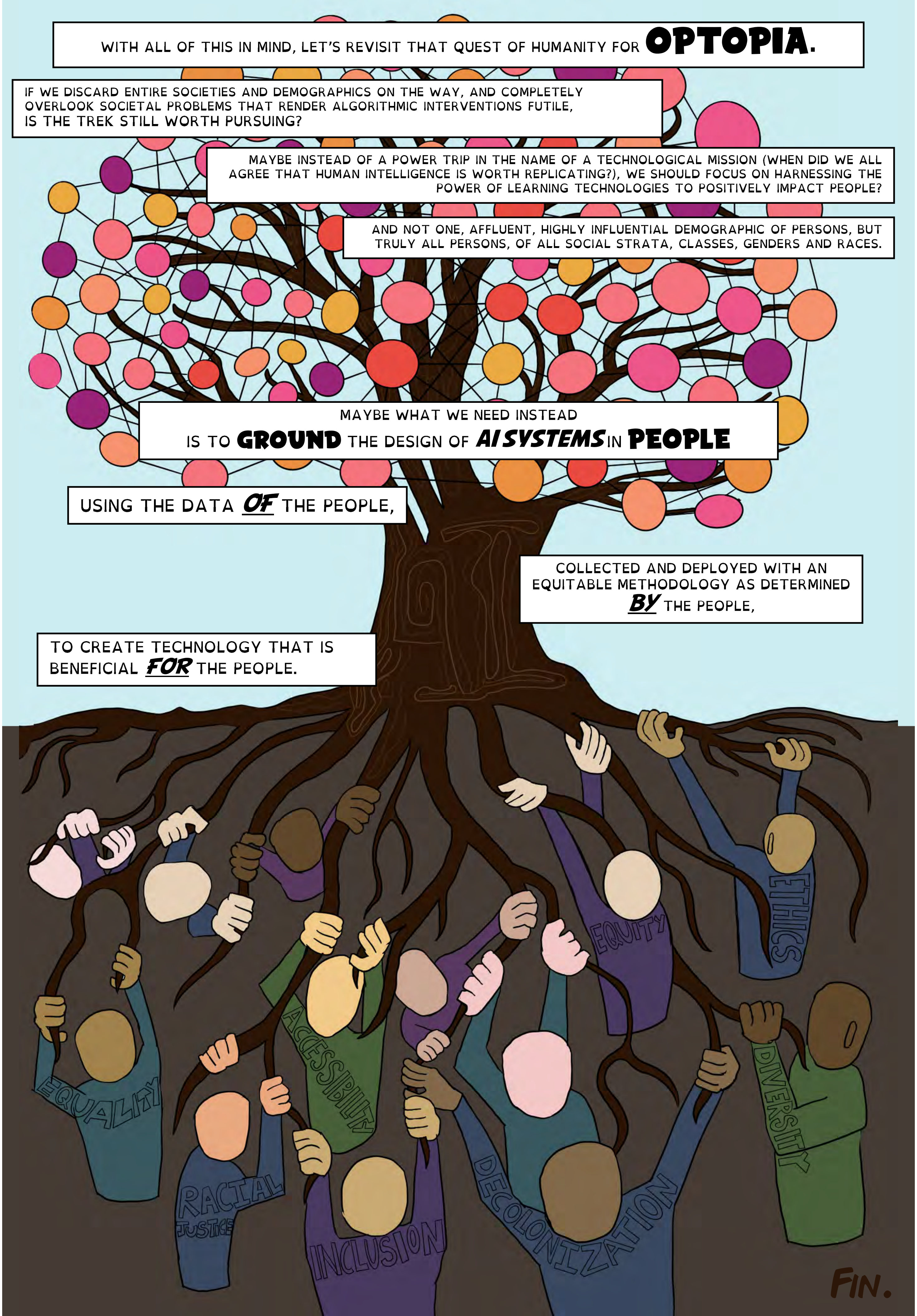
TAKE THE **NYC AUTOMATED DECISION SYSTEMS (ADS) TASK FORCE**, THE FIRST OF ITS KIND IN THE U.S., ENVISIONED TO BE THE BEACON FOR **TRANSPARENCY** AND **EXPERT INSIGHT** INTO THE USE OF ALGORITHMS TO AID DECISION-MAKING BY CITY AGENCIES. (20)

BUT THEY DIDN'T GET VERY FAR.

A **GOOD DEFINITION** WAS LACKING, AS WERE **EXAMPLES**.

WHAT IS AN **ADS**?

A CALCULATOR IS NOT AN ADS. BUT A SYSTEM THAT COLLECTS DATA, BUILDS A MODEL, AND THEN ENACTS POLICY THAT IMPACTS PEOPLE'S LIVES-ALLOCATES SCHOOL BUDGETS, OR OFFERS HOMELESSNESS ASSISTANCE, OR MATCHES STUDENTS WITH SPOTS IN HIGH SCHOOLS-CERTAINLY IS.



WITH ALL OF THIS IN MIND, LET'S REVISIT THAT QUEST OF HUMANITY FOR **OPTOPIA.**

IF WE DISCARD ENTIRE SOCIETIES AND DEMOGRAPHICS ON THE WAY, AND COMPLETELY OVERLOOK SOCIETAL PROBLEMS THAT RENDER ALGORITHMIC INTERVENTIONS FUTILE, IS THE TREK STILL WORTH PURSUING?

MAYBE INSTEAD OF A POWER TRIP IN THE NAME OF A TECHNOLOGICAL MISSION (WHEN DID WE ALL AGREE THAT HUMAN INTELLIGENCE IS WORTH REPLICATING?), WE SHOULD FOCUS ON HARNESSING THE POWER OF LEARNING TECHNOLOGIES TO POSITIVELY IMPACT PEOPLE?

AND NOT ONE, AFFLUENT, HIGHLY INFLUENTIAL DEMOGRAPHIC OF PERSONS, BUT TRULY ALL PERSONS, OF ALL SOCIAL STRATA, CLASSES, GENDERS AND RACES.

MAYBE WHAT WE NEED INSTEAD IS TO **GROUND** THE DESIGN OF ***AI SYSTEMS*** IN **PEOPLE**

USING THE DATA ***OF*** THE PEOPLE,

COLLECTED AND DEPLOYED WITH AN EQUITABLE METHODOLOGY AS DETERMINED ***BY*** THE PEOPLE,

TO CREATE TECHNOLOGY THAT IS BENEFICIAL ***FOR*** THE PEOPLE.

ABOUT

Falaah is a Scientist/Engineer by training and an Artist by nature, chasing a passion for building Robust and Ethical ML all the way from industry to academia. In the face of having to incessantly remind everyone around her about the limitations of current ML capabilities, Falaah started "**MACHINELEARNIST COMICS**" - online Scientific Comics about the AI Landscape.

Julia is an Assistant Professor of Computer Science and Engineering and of Data Science at NYU. She is passionate about Responsible Data Science and leads the "***DATA, RESPONSIBLY***" project, the latest offering of which is the inimitable, interdisciplinary course on **RESPONSIBLE DATA SCIENCE**.

With the *undecipherable alchemy* that is grad-school admissions, the **Cosmos** brought these two creative minds together and thus was born: **DATA, RESPONSIBLY COMICS**!

Whether you're a **Student**, unsure about where to get started in the sea of ML scholarship; or an **Educator**, looking for a fun new pedagogical instrument for your students; or a **Practitioner**, looking for some relatable content about all the idiosyncrasies of the current AI landscape; or just a good ol' John/Jane Doe who likes to read comics and is intrigued by the prospect of a long form scientific volume, Data, Responsibly Comics are for you!

JULIA STOYANOVICH
@stoyanoj
Co-Creator, Writer

FALAAH ARIF KHAN
@FalaahArifKhan
Co-Creator, Writer, Artist, Cover Artist

REFERENCES :

[1] https://mrtz.org/gradientina.html#/

[2] http://www.hutchinsweb.me.uk/MTNI-11-1995.pdf .

[3] https://www.who.int/disabilities/world_report/2011/report/en/

[4] https://www.a11yproject.com/

[5] http://nomorecraptions.com/

[6] https://datasociety.net/library/dark-patterns-in-accessibility-tech/

[7] https://twitter.com/habengirma/status/1278035954628915200

[8] https://en.wikipedia.org/wiki/Facebook_real-name_policy_controversy

[9] https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf

[10] https://sarahwyerblogs.wordpress.com/2020/08/17/classed-outliers-the-algorithmic-divide-in-plain-sight-a-levels-and-highers-divide-the-uk/

[11] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[12] https://github.com/openai/gpt-3

[13] https://www.nbcnews.com/business/business-news/man-wrongfully-arrested-due-facial-recognition-software-talks-about-humiliating-n1232184

[14] http://gendershades.org/

[15] https://in.reuters.com/article/amazon-com-jobs-automation/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idINKCN1MK0AH

[16] https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html

[17] https://www.imdb.com/title/tt2584384/

[18] https://dataresponsibly.github.io/documents/mirror.pdf

[19] https://quoteinvestigator.com/tag/niels-bohr/

[20] https://www1.nyc.gov/site/adstaskforce/index.page

WE ARE AI
#3

Who lives, Who dies, Who decides?

INEQUALITY
VALUES
UTILITARIANISM
UNCERTAINTY
ETHICS

TERMS OF USE

All the panels in this comic book are licensed CC 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

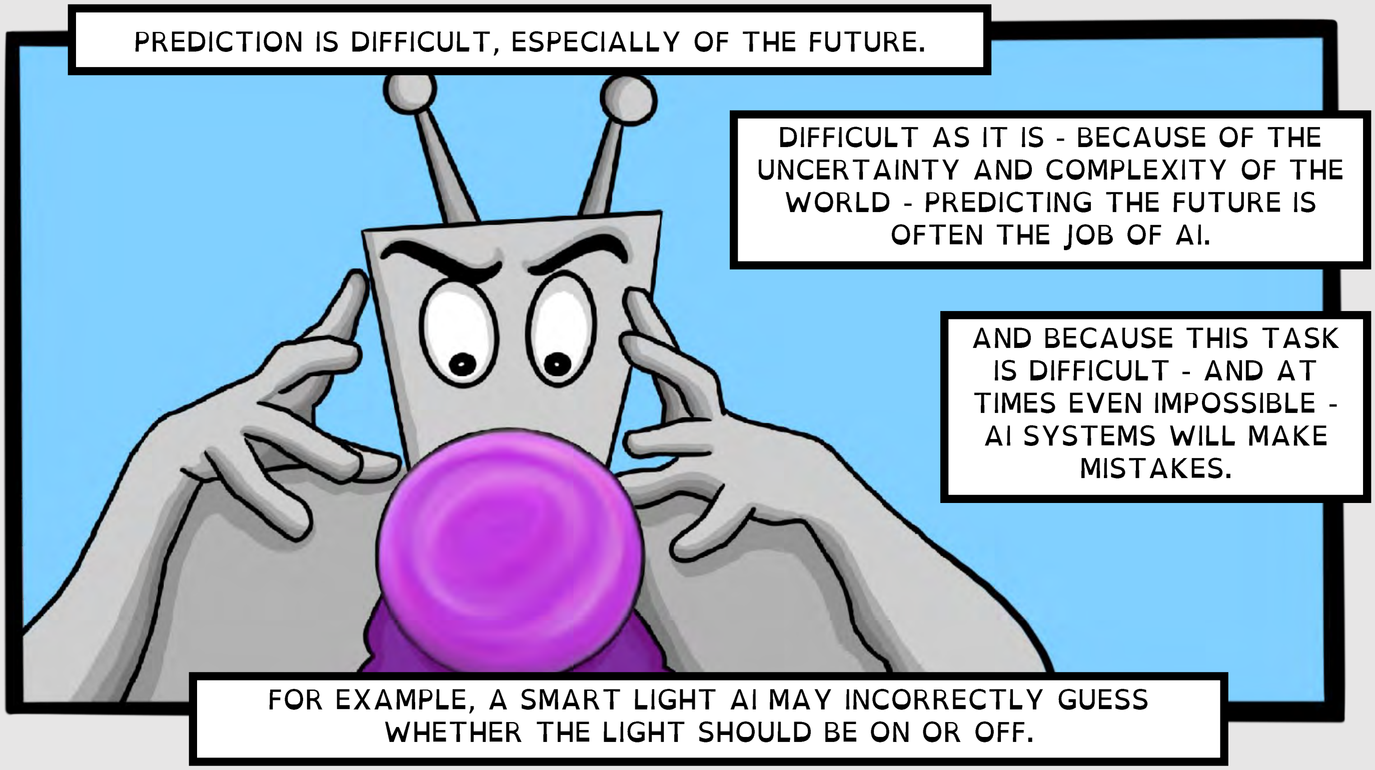
1. Provide the proper citation
2. Do not make modifications to the individual panels themselves

Cite as:

Julia Stoyanovich, Mona Sloane and Falaah Arif Khan.
"Who lives, who dies, who decides?". *We are AI Comics*, Vol 3 (2021)
https://dataresponsibly.github.io/we-are-ai/comics/vol3_en.pdf

Contact:

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



PREDICTION IS DIFFICULT, ESPECIALLY OF THE FUTURE.

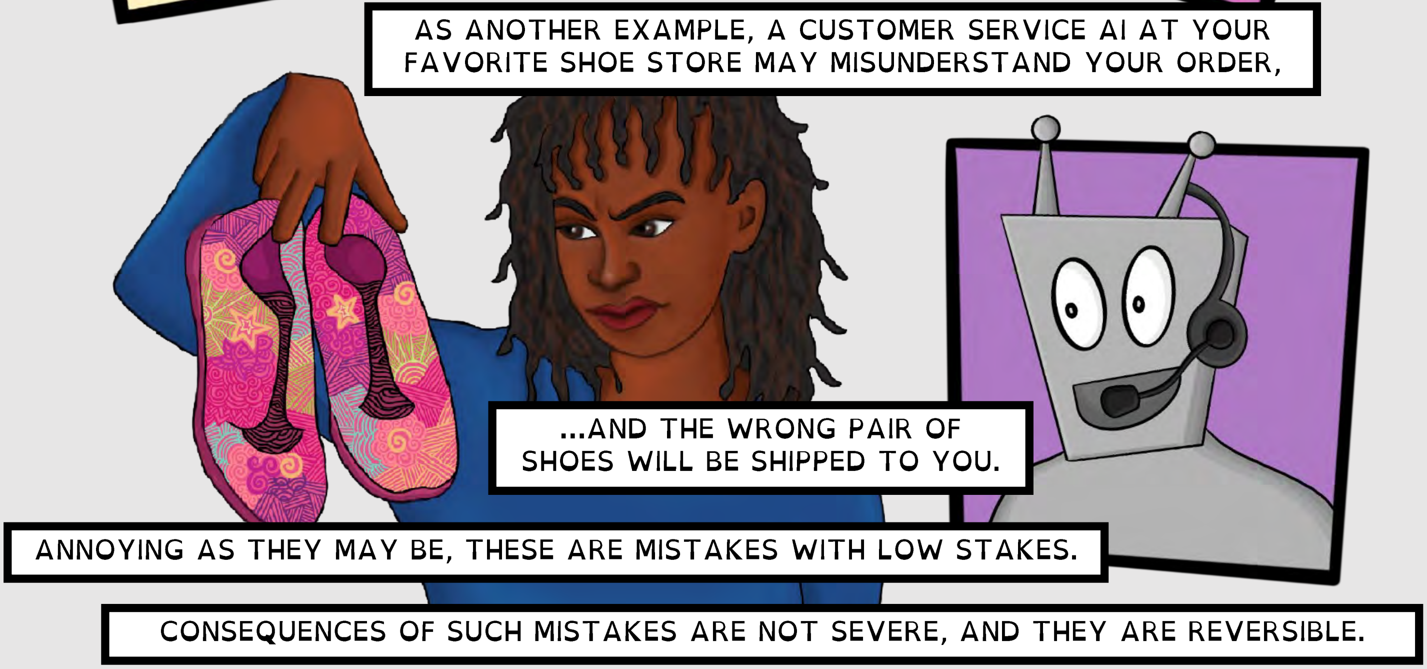
DIFFICULT AS IT IS - BECAUSE OF THE UNCERTAINTY AND COMPLEXITY OF THE WORLD - PREDICTING THE FUTURE IS OFTEN THE JOB OF AI.

AND BECAUSE THIS TASK IS DIFFICULT - AND AT TIMES EVEN IMPOSSIBLE - AI SYSTEMS WILL MAKE MISTAKES.

FOR EXAMPLE, A SMART LIGHT AI MAY INCORRECTLY GUESS WHETHER THE LIGHT SHOULD BE ON OR OFF.

TURNING ON THE LIGHT IN THE MIDDLE OF THE NIGHT WILL WAKE YOU UP.

AND LEAVING IT ON SYSTEMATICALLY WILL LEAVE YOU TO FOOT A HIGH ENERGY BILL.



AS ANOTHER EXAMPLE, A CUSTOMER SERVICE AI AT YOUR FAVORITE SHOE STORE MAY MISUNDERSTAND YOUR ORDER,

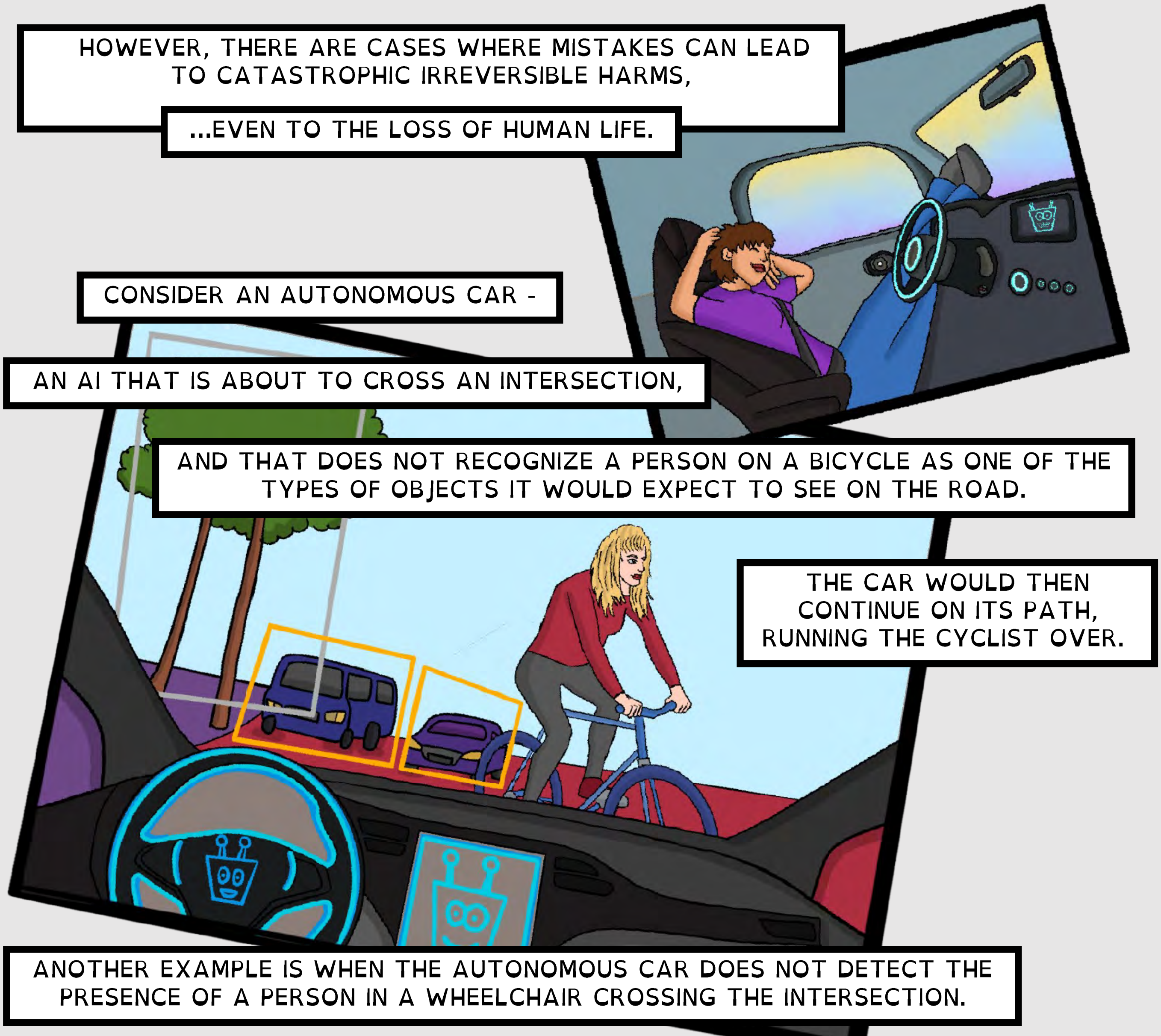
...AND THE WRONG PAIR OF SHOES WILL BE SHIPPED TO YOU.

ANNOYING AS THEY MAY BE, THESE ARE MISTAKES WITH LOW STAKES.

CONSEQUENCES OF SUCH MISTAKES ARE NOT SEVERE, AND THEY ARE REVERSIBLE.

HOWEVER, THERE ARE CASES WHERE MISTAKES CAN LEAD TO CATASTROPHIC IRREVERSIBLE HARMS,

...EVEN TO THE LOSS OF HUMAN LIFE.



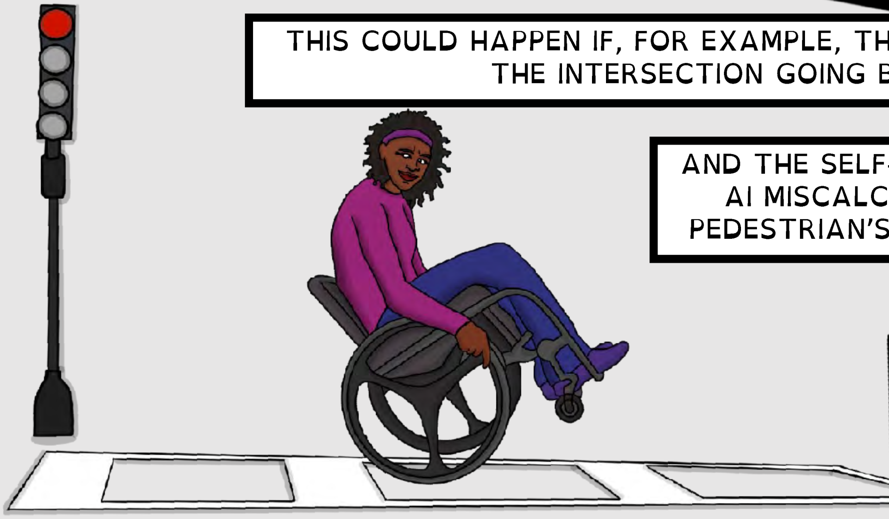
CONSIDER AN AUTONOMOUS CAR -

AN AI THAT IS ABOUT TO CROSS AN INTERSECTION,

AND THAT DOES NOT RECOGNIZE A PERSON ON A BICYCLE AS ONE OF THE TYPES OF OBJECTS IT WOULD EXPECT TO SEE ON THE ROAD.

THE CAR WOULD THEN CONTINUE ON ITS PATH, RUNNING THE CYCLIST OVER.

ANOTHER EXAMPLE IS WHEN THE AUTONOMOUS CAR DOES NOT DETECT THE PRESENCE OF A PERSON IN A WHEELCHAIR CROSSING THE INTERSECTION.



THIS COULD HAPPEN IF, FOR EXAMPLE, THE PERSON WERE CROSSING THE INTERSECTION GOING BACKWARDS,

AND THE SELF-DRIVING CAR'S AI MISCALCULATES THE PEDESTRIAN'S TRAJECTORY.

BUT HUMAN DRIVERS ALSO CAUSE ACCIDENTS!

SO WHY LET PERFECT BE THE ENEMY OF GOOD?

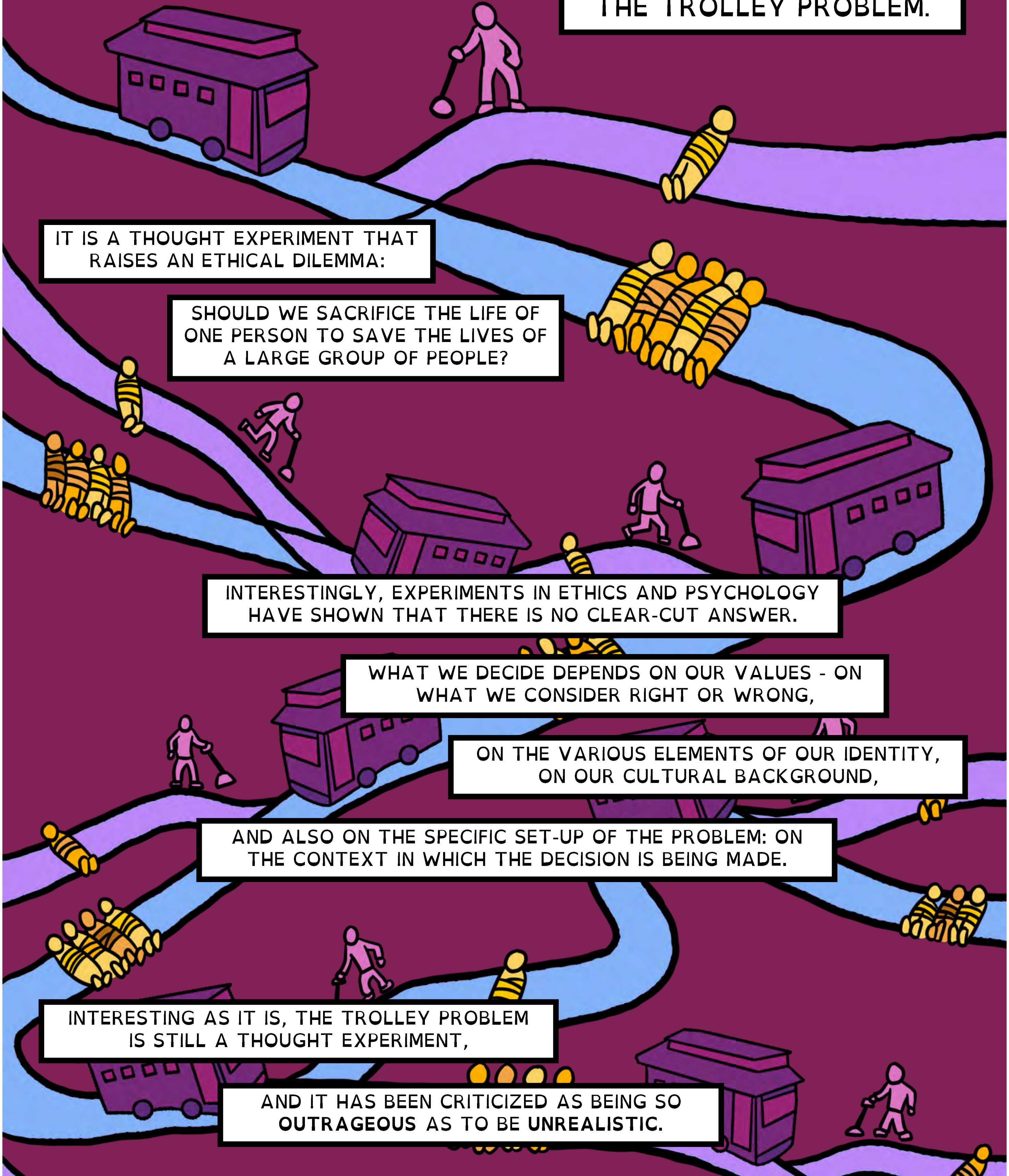
SHOULDN'T WE BE PREPARED TO SUFFER A FEW MISTAKES MADE BY AUTONOMOUS CARS IN THE NAME OF INCREASED OVERALL SAFETY OF OUR TRANSPORTATION SYSTEM, AND THE CONVENIENCE TO THE DRIVERS?

IN FACT, CAN'T WE ENCODE OUR JUDGEMENT ABOUT WHAT MISTAKES ARE MORE IMPORTANT TO AVOID, AND LET AN AI SORT OUT THE TRADE-OFFS?

CAN'T WE EQUIP OUR AI WITH VALUES?

A FAMOUS EXAMPLE THAT MAKES US THINK ABOUT OUR VALUES, AND TRADE-OFFS THEY INTRODUCE, IS

THE TROLLEY PROBLEM.



IT IS A THOUGHT EXPERIMENT THAT RAISES AN ETHICAL DILEMMA:

SHOULD WE SACRIFICE THE LIFE OF ONE PERSON TO SAVE THE LIVES OF A LARGE GROUP OF PEOPLE?

INTERESTINGLY, EXPERIMENTS IN ETHICS AND PSYCHOLOGY HAVE SHOWN THAT THERE IS NO CLEAR-CUT ANSWER.

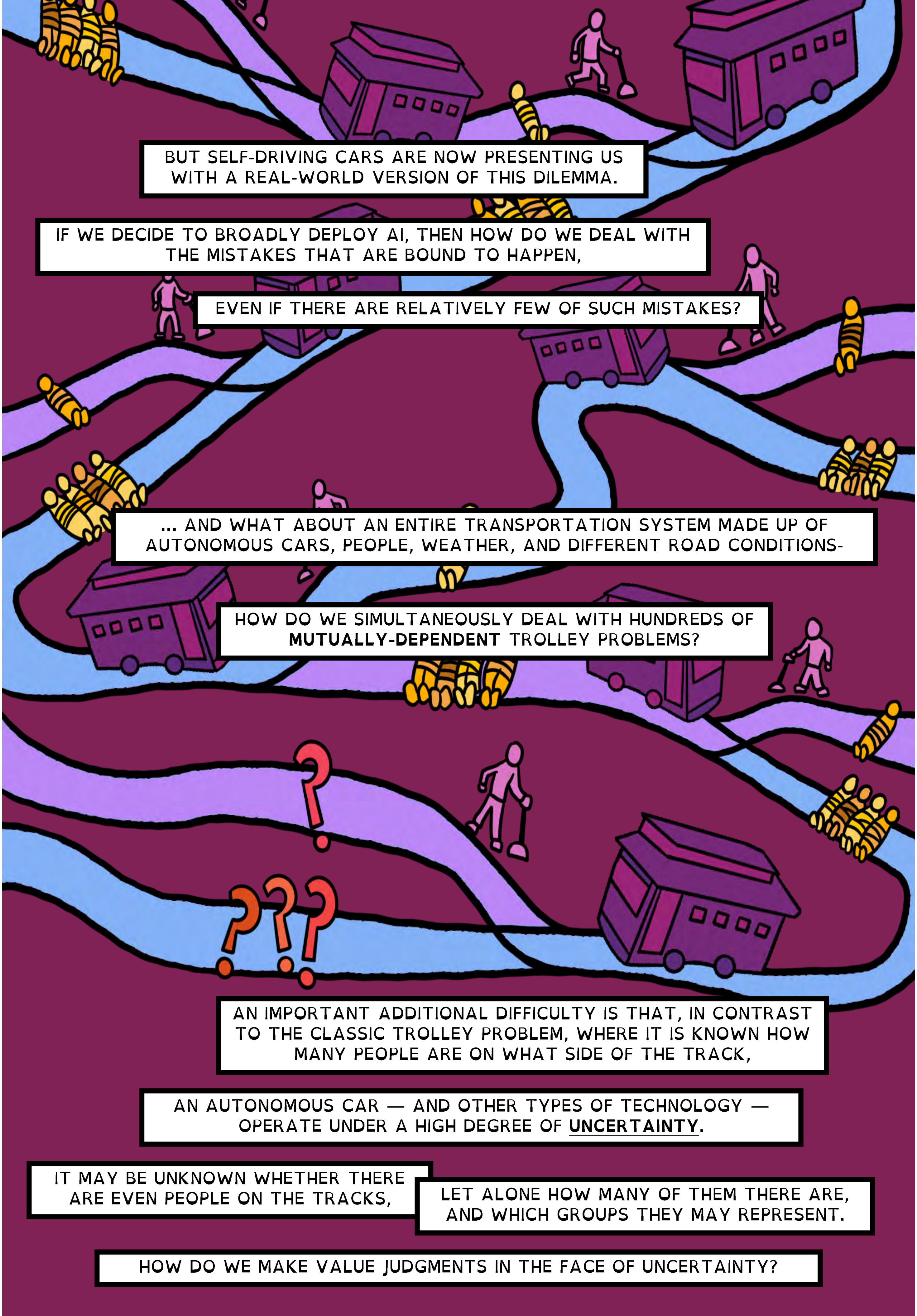
WHAT WE DECIDE DEPENDS ON OUR VALUES - ON WHAT WE CONSIDER RIGHT OR WRONG,

ON THE VARIOUS ELEMENTS OF OUR IDENTITY, ON OUR CULTURAL BACKGROUND,

AND ALSO ON THE SPECIFIC SET-UP OF THE PROBLEM: ON THE CONTEXT IN WHICH THE DECISION IS BEING MADE.

INTERESTING AS IT IS, THE TROLLEY PROBLEM IS STILL A THOUGHT EXPERIMENT,

AND IT HAS BEEN CRITICIZED AS BEING SO **OUTRAGEOUS** AS TO BE **UNREALISTIC.**



BUT SELF-DRIVING CARS ARE NOW PRESENTING US WITH A REAL-WORLD VERSION OF THIS DILEMMA.

IF WE DECIDE TO BROADLY DEPLOY AI, THEN HOW DO WE DEAL WITH THE MISTAKES THAT ARE BOUND TO HAPPEN,

EVEN IF THERE ARE RELATIVELY FEW OF SUCH MISTAKES?

... AND WHAT ABOUT AN ENTIRE TRANSPORTATION SYSTEM MADE UP OF AUTONOMOUS CARS, PEOPLE, WEATHER, AND DIFFERENT ROAD CONDITIONS-

HOW DO WE SIMULTANEOUSLY DEAL WITH HUNDREDS OF **MUTUALLY-DEPENDENT** TROLLEY PROBLEMS?

AN IMPORTANT ADDITIONAL DIFFICULTY IS THAT, IN CONTRAST TO THE CLASSIC TROLLEY PROBLEM, WHERE IT IS KNOWN HOW MANY PEOPLE ARE ON WHAT SIDE OF THE TRACK,

AN AUTONOMOUS CAR — AND OTHER TYPES OF TECHNOLOGY — OPERATE UNDER A HIGH DEGREE OF **UNCERTAINTY**.

IT MAY BE UNKNOWN WHETHER THERE ARE EVEN PEOPLE ON THE TRACKS,

LET ALONE HOW MANY OF THEM THERE ARE, AND WHICH GROUPS THEY MAY REPRESENT.

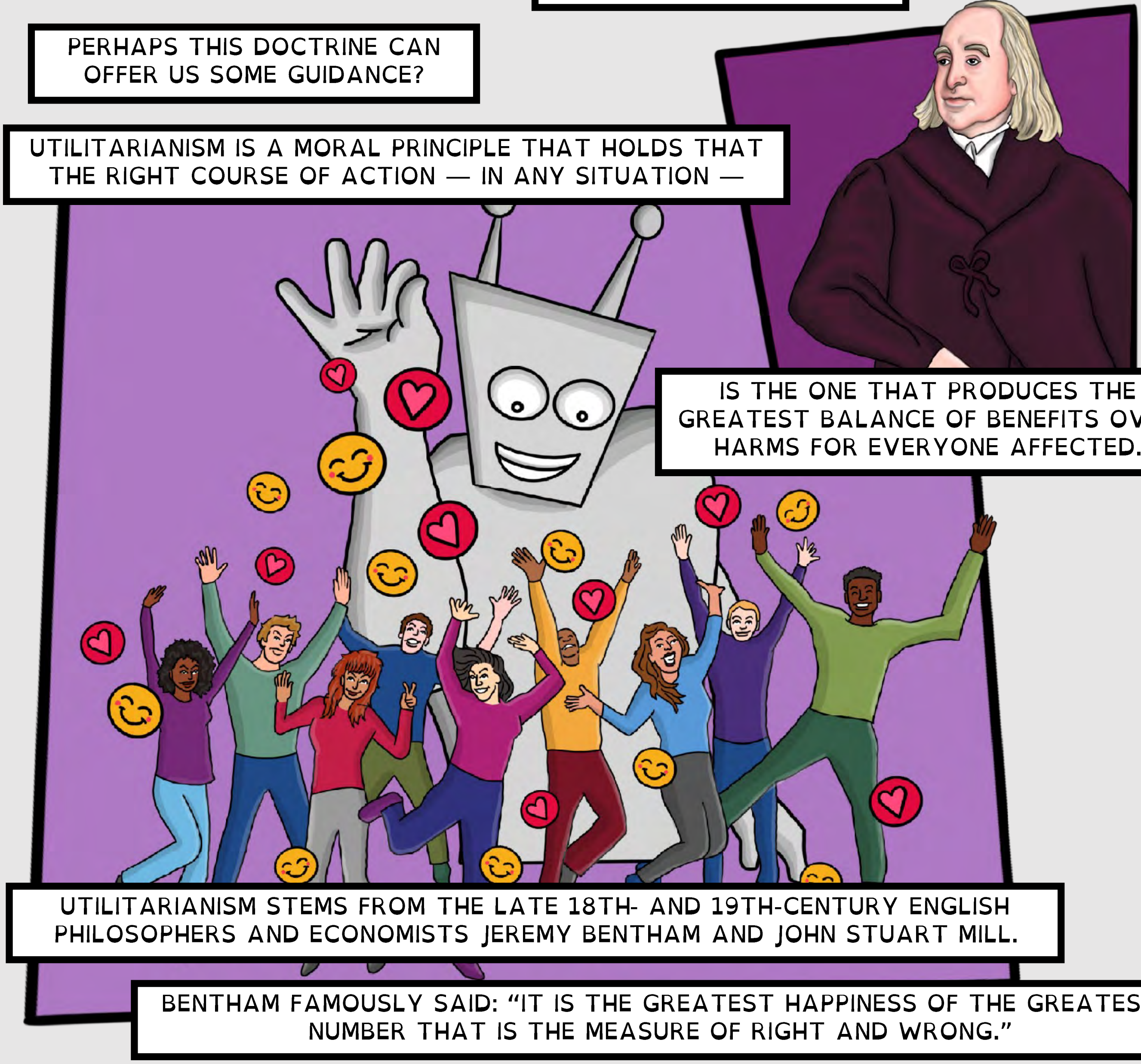
HOW DO WE MAKE VALUE JUDGMENTS IN THE FACE OF UNCERTAINTY?

THE TROLLEY CAR PROBLEM ILLUSTRATES A SPECIFIC DOCTRINE OF MORAL PHILOSOPHY -

UTILITARIANISM.

PERHAPS THIS DOCTRINE CAN OFFER US SOME GUIDANCE?

UTILITARIANISM IS A MORAL PRINCIPLE THAT HOLDS THAT THE RIGHT COURSE OF ACTION — IN ANY SITUATION —

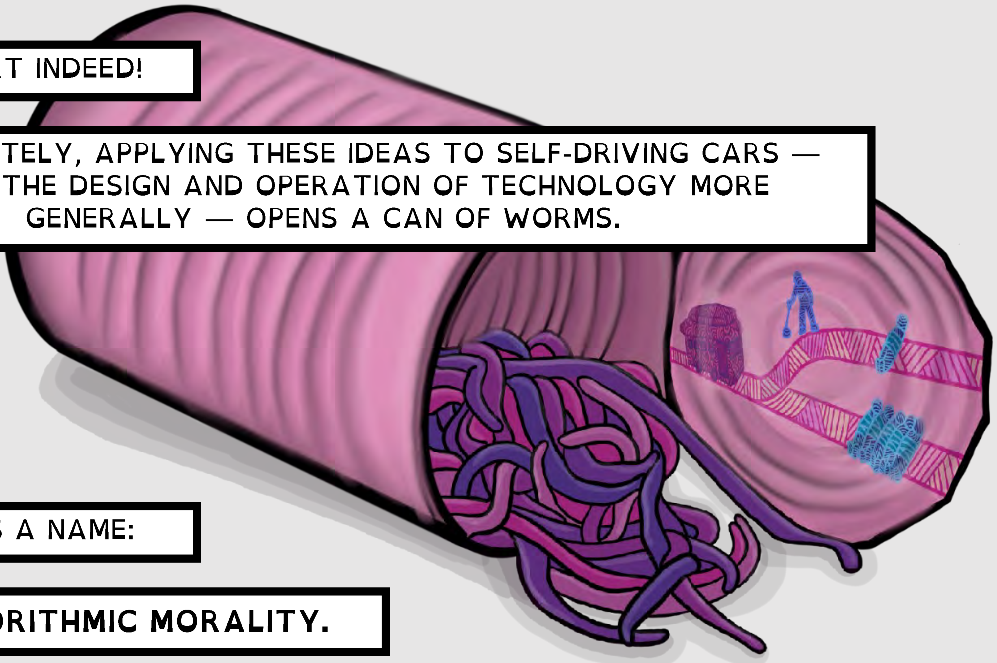


IS THE ONE THAT PRODUCES THE GREATEST BALANCE OF BENEFITS OVER HARMS FOR EVERYONE AFFECTED.

UTILITARIANISM STEMS FROM THE LATE 18TH- AND 19TH-CENTURY ENGLISH PHILOSOPHERS AND ECONOMISTS JEREMY BENTHAM AND JOHN STUART MILL.

BENTHAM FAMOUSLY SAID: “IT IS THE GREATEST HAPPINESS OF THE GREATEST NUMBER THAT IS THE MEASURE OF RIGHT AND WRONG.”

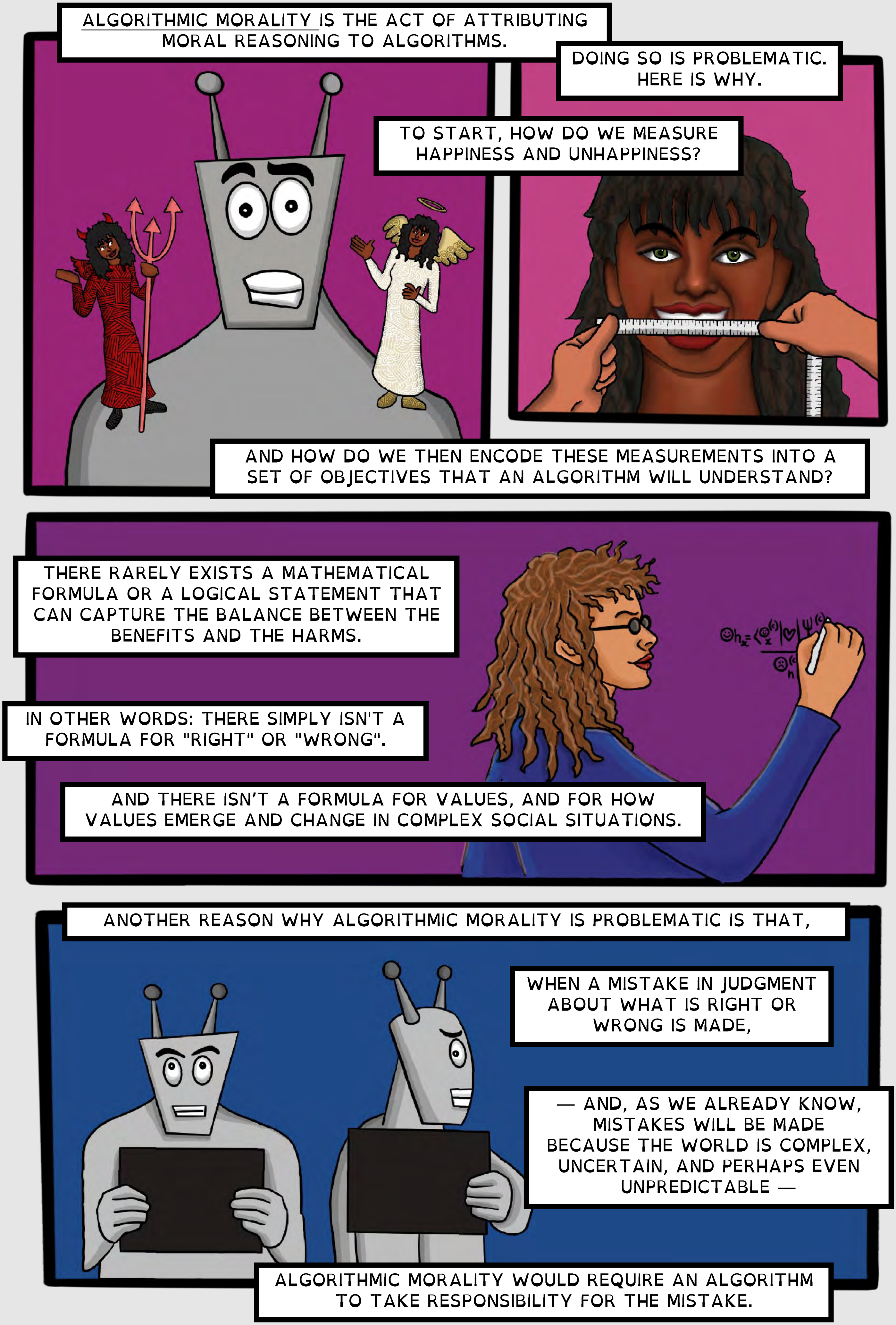
SOUNDS GREAT INDEED!

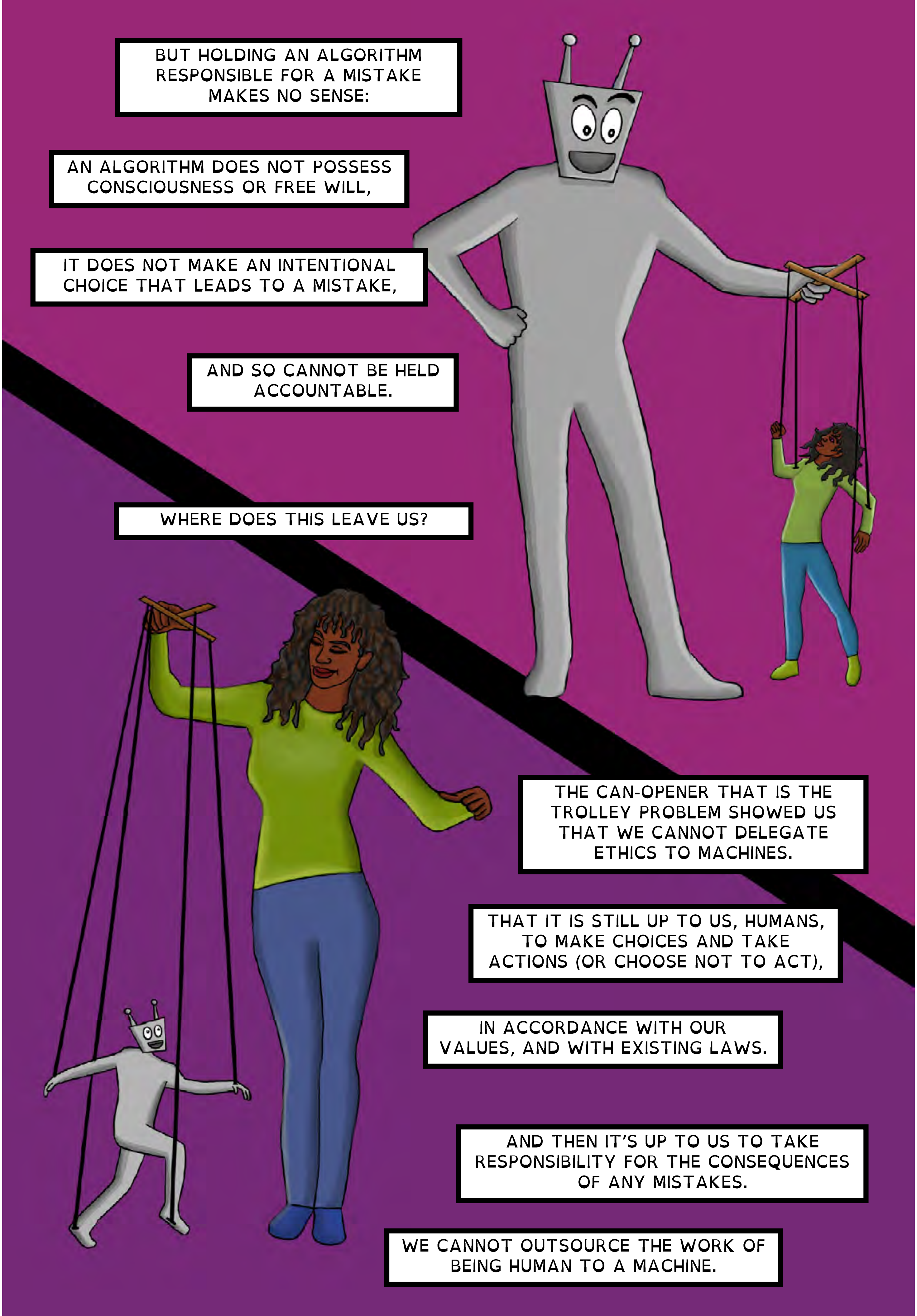


UNFORTUNATELY, APPLYING THESE IDEAS TO SELF-DRIVING CARS — AND TO THE DESIGN AND OPERATION OF TECHNOLOGY MORE GENERALLY — OPENS A CAN OF WORMS.

AND IT HAS A NAME:

ALGORITHMIC MORALITY.


ALGORITHMIC MORALITY IS THE ACT OF ATTRIBUTING MORAL REASONING TO ALGORITHMS.
DOING SO IS PROBLEMATIC. HERE IS WHY.
TO START, HOW DO WE MEASURE HAPPINESS AND UNHAPPINESS?
AND HOW DO WE THEN ENCODE THESE MEASUREMENTS INTO A SET OF OBJECTIVES THAT AN ALGORITHM WILL UNDERSTAND?
THERE RARELY EXISTS A MATHEMATICAL FORMULA OR A LOGICAL STATEMENT THAT CAN CAPTURE THE BALANCE BETWEEN THE BENEFITS AND THE HARMS.
IN OTHER WORDS: THERE SIMPLY ISN'T A FORMULA FOR "RIGHT" OR "WRONG".
AND THERE ISN'T A FORMULA FOR VALUES, AND FOR HOW VALUES EMERGE AND CHANGE IN COMPLEX SOCIAL SITUATIONS.
ANOTHER REASON WHY ALGORITHMIC MORALITY IS PROBLEMATIC IS THAT,
WHEN A MISTAKE IN JUDGMENT ABOUT WHAT IS RIGHT OR WRONG IS MADE,
— AND, AS WE ALREADY KNOW, MISTAKES WILL BE MADE BECAUSE THE WORLD IS COMPLEX, UNCERTAIN, AND PERHAPS EVEN UNPREDICTABLE —
ALGORITHMIC MORALITY WOULD REQUIRE AN ALGORITHM TO TAKE RESPONSIBILITY FOR THE MISTAKE.



BUT HOLDING AN ALGORITHM RESPONSIBLE FOR A MISTAKE MAKES NO SENSE:

AN ALGORITHM DOES NOT POSSESS CONSCIOUSNESS OR FREE WILL,

IT DOES NOT MAKE AN INTENTIONAL CHOICE THAT LEADS TO A MISTAKE,

AND SO CANNOT BE HELD ACCOUNTABLE.

WHERE DOES THIS LEAVE US?

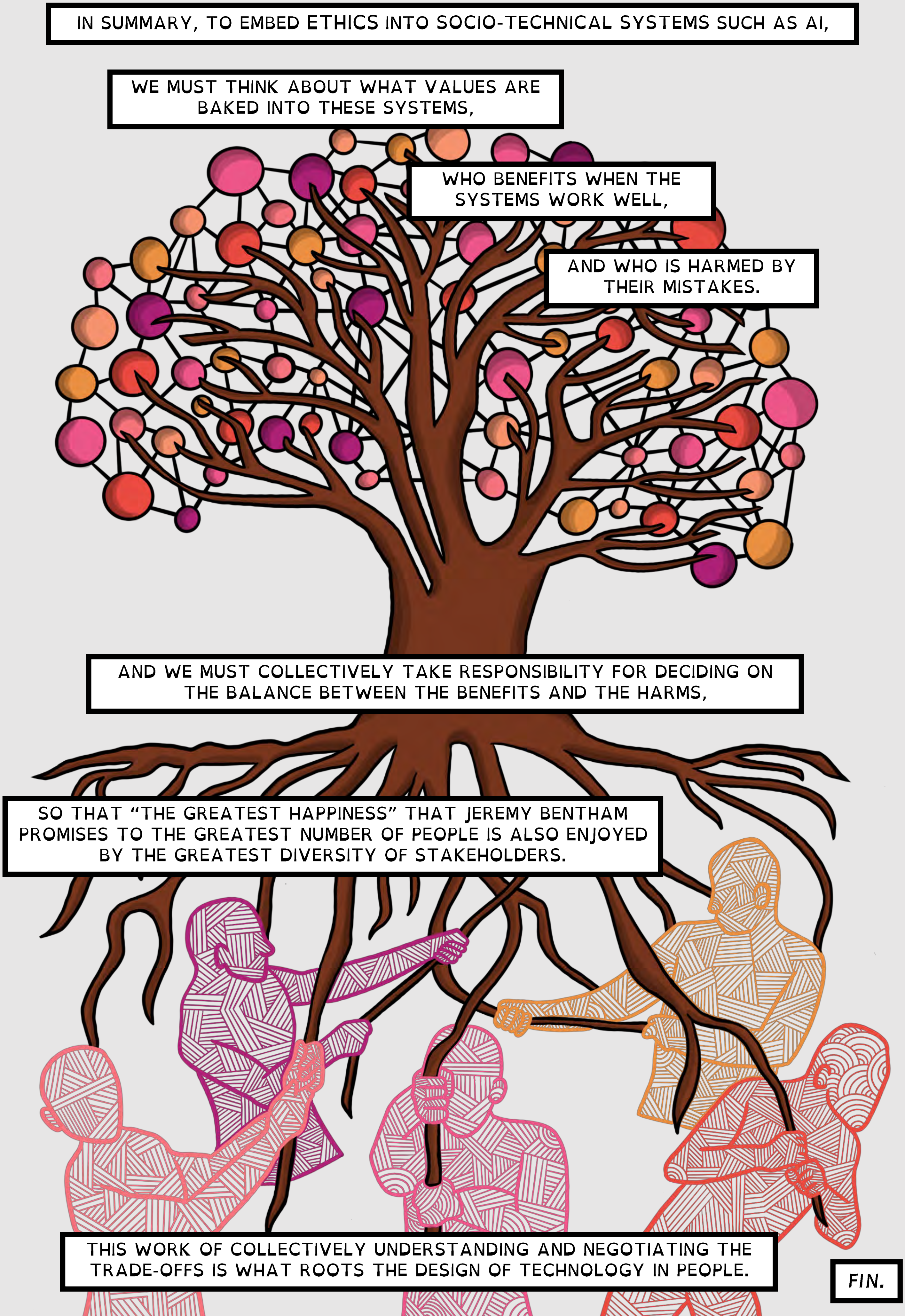
THE CAN-OPENER THAT IS THE TROLLEY PROBLEM SHOWED US THAT WE CANNOT DELEGATE ETHICS TO MACHINES.

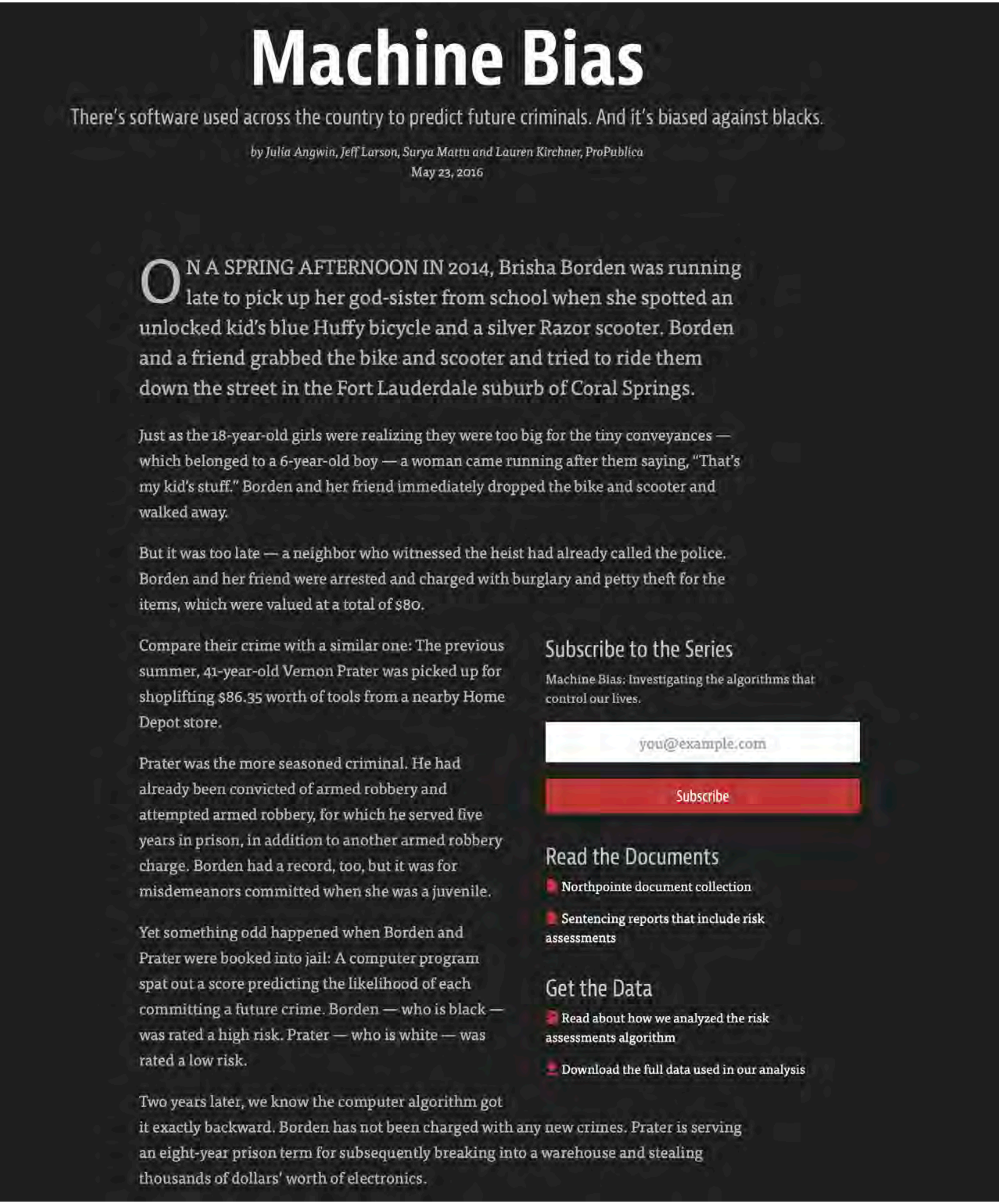
THAT IT IS STILL UP TO US, HUMANS, TO MAKE CHOICES AND TAKE ACTIONS (OR CHOOSE NOT TO ACT),

IN ACCORDANCE WITH OUR VALUES, AND WITH EXISTING LAWS.

AND THEN IT'S UP TO US TO TAKE RESPONSIBILITY FOR THE CONSEQUENCES OF ANY MISTAKES.

WE CANNOT OUTSOURCE THE WORK OF BEING HUMAN TO A MACHINE.


IN SUMMARY, TO EMBED ETHICS INTO SOCIO-TECHNICAL SYSTEMS SUCH AS AI,
WE MUST THINK ABOUT WHAT VALUES ARE BAKED INTO THESE SYSTEMS,
WHO BENEFITS WHEN THE SYSTEMS WORK WELL,
AND WHO IS HARMED BY THEIR MISTAKES.
AND WE MUST COLLECTIVELY TAKE RESPONSIBILITY FOR DECIDING ON THE BALANCE BETWEEN THE BENEFITS AND THE HARMS,
SO THAT "THE GREATEST HAPPINESS" THAT JEREMY BENTHAM PROMISES TO THE GREATEST NUMBER OF PEOPLE IS ALSO ENJOYED BY THE GREATEST DIVERSITY OF STAKEHOLDERS.
THIS WORK OF COLLECTIVELY UNDERSTANDING AND NEGOTIATING THE TRADE-OFFS IS WHAT ROOTS THE DESIGN OF TECHNOLOGY IN PEOPLE.
FIN.



Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Just as the 18-year-old girls were realizing they were too big for the tiny conveyances — which belonged to a 6-year-old boy — a woman came running after them saying, "That's my kid's stuff." Borden and her friend immediately dropped the bike and scooter and walked away.

But it was too late — a neighbor who witnessed the heist had already called the police. Borden and her friend were arrested and charged with burglary and petty theft for the items, which were valued at a total of $80.

Compare their crime with a similar one: The previous summer, 41-year-old Vernon Prater was picked up for shoplifting $86.35 worth of tools from a nearby Home Depot store.

Prater was the more seasoned criminal. He had already been convicted of armed robbery and attempted armed robbery, for which he served five years in prison, in addition to another armed robbery charge. Borden had a record, too, but it was for misdemeanors committed when she was a juvenile.

Yet something odd happened when Borden and Prater were booked into jail: A computer program spat out a score predicting the likelihood of each committing a future crime. Borden — who is black — was rated a high risk. Prater — who is white — was rated a low risk.

Two years later, we know the computer algorithm got it exactly backward. Borden has not been charged with any new crimes. Prater is serving an eight-year prison term for subsequently breaking into a warehouse and stealing thousands of dollars' worth of electronics.

Subscribe to the Series

Machine Bias: Investigating the algorithms that control our lives.

Read the Documents

- Northpointe document collection
- Sentencing reports that include risk assessments

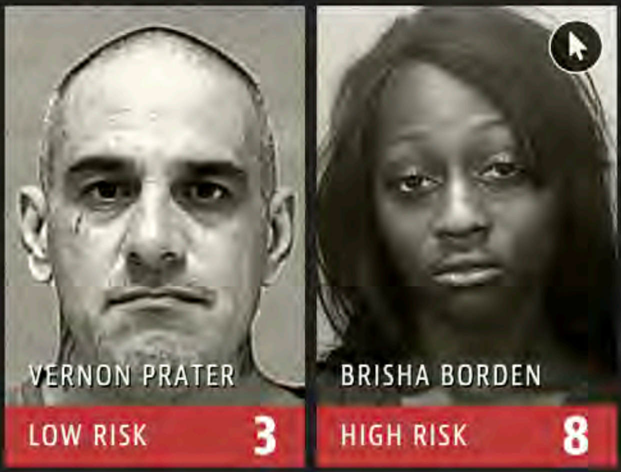
Get the Data

- Read about how we analyzed the risk assessments algorithm
- Download the full data used in our analysis

Scores like this — known as risk assessments — are increasingly common in courtrooms across the nation. They are used to inform decisions about who can be set free at every stage of the criminal justice system, from assigning bond amounts — as is the case in Fort Lauderdale — to even more fundamental decisions about defendants' freedom. In Arizona, Colorado, Delaware, Kentucky, Louisiana, Oklahoma, Virginia, Washington and Wisconsin, the results of such assessments are given to judges during criminal sentencing.

Rating a defendant's risk of future crime is often done in conjunction with an evaluation of a defendant's rehabilitation needs. The Justice Department's National Institute of Corrections now encourages the use of such combined assessments at every stage of the criminal justice process. And a landmark sentencing reform bill currently pending in Congress would mandate the use of such assessments in federal prisons.

Two Petty Theft Arrests



Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

In 2014, then U.S. Attorney General Eric Holder warned that the risk scores might be injecting bias into the courts. He called for the U.S. Sentencing Commission to study their use. "Although these measures were crafted with the best of intentions, I am concerned that they inadvertently undermine our efforts to ensure individualized and equal justice," he said, adding, "they may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society."

The sentencing commission did not, however, launch a study of risk scores. So ProPublica did, as part of a larger examination of the powerful, largely hidden effect of algorithms in American life.

We obtained the risk scores assigned to more than 7,000 people arrested in Broward County, Florida, in 2013 and 2014 and checked to see how many were charged with new crimes over the next two years, the same benchmark used by the creators of the algorithm.

The score proved remarkably unreliable in forecasting violent crime: Only 20 percent of the people predicted to commit violent crimes actually went on to do so.

When a full range of crimes were taken into account — including misdemeanors such as driving with an expired license — the algorithm was somewhat more accurate than a coin flip. Of those deemed likely to re-offend, 61 percent were arrested for any subsequent crimes within two years.

We also turned up significant racial disparities, just as Holder feared. In forecasting who would re-offend, the algorithm made mistakes with black and white defendants at roughly the same rate but in very different ways.

- The formula was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants.
- White defendants were mislabeled as low risk more often than black defendants.

Could this disparity be explained by defendants' prior crimes or the type of crimes they were arrested for? No. We ran a statistical test that isolated the effect of race from criminal history and recidivism, as well as from defendants' age and gender. Black defendants were still 77 percent more likely to be pegged as at higher risk of committing a future violent crime and 45 percent more likely to be predicted to commit a future crime of any kind. (Read our analysis.)

The algorithm used to create the Florida risk scores is a product of a for-profit company, Northpointe. The company disputes our analysis.

In a letter, it criticized ProPublica's methodology and defended the accuracy of its test: "Northpointe does not agree that the results of your analysis, or the claims being made based upon that analysis, are correct or that they accurately reflect the outcomes from the application of the model."

Northpointe's software is among the most widely used assessment tools in the country. The company does not publicly disclose the calculations used to arrive at defendants' risk scores, so it is not possible for either defendants or the public to see what might be driving the disparity. (On Sunday, Northpointe gave ProPublica the basics of its future-crime formula — which includes factors such as education levels, and whether a defendant has a job. It did not share the specific calculations, which it said are proprietary.)

Northpointe's core product is a set of scores derived from 137 questions that are either answered by defendants or pulled from criminal records. Race is not one of the questions. The survey asks defendants such things as: "Was one of your parents ever sent to jail or prison?" "How many of your friends/acquaintances are taking drugs illegally?" and "How often did you get in fights while at school?" The questionnaire also asks people to agree or disagree with statements such as "A hungry person has a right to steal" and "If people make me angry or lose my temper, I can be dangerous."

The appeal of risk scores is obvious: The United States locks up far more people than any other country, a disproportionate number of them black. For more than two centuries, the key decisions in the legal process, from pretrial release to sentencing to parole, have been in the hands of human beings guided by their instincts and personal biases.

If computers could accurately predict which defendants were likely to commit new crimes, the criminal justice system could be fairer and more selective about who is incarcerated and for how long. The trick, of course, is to make sure the computer gets it right. If it's wrong in one direction, a dangerous criminal could go free. If it's wrong in another direction, it could result in someone unfairly receiving a harsher sentence or waiting longer for parole than is appropriate.

The first time Paul Zilly heard of his score — and realized how much was riding on it — was during his sentencing hearing on Feb. 15, 2013, in court in Barron County, Wisconsin. Zilly had been convicted of stealing a push lawnmower and some tools. The prosecutor recommended a year in county jail and follow-up supervision that could help Zilly with "staying on the right path." His lawyer agreed to a plea deal.

But Judge James Babler had seen Zilly's scores. Northpointe's software had rated Zilly as a high risk for future violent crime and a medium risk for general recidivism. "When I look at the risk assessment," Babler said in court, "it is about as bad as it could be."

Then Babler overturned the plea deal that had been agreed on by the prosecution and defense and imposed two years in state prison and three years of supervision.

CRIMINOLOGISTS HAVE LONG TRIED to predict which criminals are more dangerous before deciding whether they should be released. Race, nationality and skin color were often used in making such predictions until about the 1970s, when it became politically unacceptable, according to a survey of risk assessment tools by Columbia University law professor Bernard Harcourt.

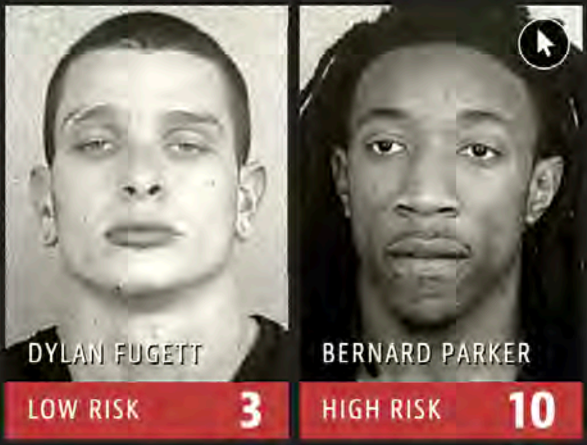
In the 1980s, as a crime wave engulfed the nation, lawmakers made it much harder for judges and parole boards to exercise discretion in making such decisions. States and the federal government began instituting mandatory sentences and, in some cases, abolished parole, making it less important to evaluate individual offenders.

But as states struggle to pay for swelling prison and jail populations, forecasting criminal risk has made a comeback.

Dozens of risk assessments are being used across the nation — some created by for-profit companies such as Northpointe and others by nonprofit organizations. (One tool being used in states including Kentucky and Arizona, called the Public Safety Assessment, was developed by the Laura and John Arnold Foundation, which also is a funder of ProPublica.)

There have been few independent studies of these criminal risk assessments. In 2013, researchers Sarah Desmarais and Jay Singh examined 19 different risk methodologies used in the United States and found that "in most cases, validity had only been examined in one or two studies" and that "frequently, those investigations were completed by the same people who developed the instrument."

Two Drug Possession Arrests



Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.

Their analysis of the research through 2012 found that the tools "were moderate at best in terms of predictive validity," Desmarais said in an interview. And she could not find any substantial set of studies conducted in the United States that examined whether risk scores were racially biased. "The data do not exist," she said.

Since then, there have been some attempts to explore racial disparities in risk scores. One 2016 study examined the validity of a risk assessment tool, not Northpointe's, used to make probation decisions for about 35,000 federal convicts. The researchers, Jennifer Skeem at University of California, Berkeley, and Christopher T. Lowenkamp from the Administrative Office of the U.S. Courts, found that blacks did get a higher average score but concluded the differences were not attributable to bias.

The increasing use of risk scores is controversial and has garnered media coverage, including articles by the Associated Press, and the Marshall Project and FiveThirtyEight last year.

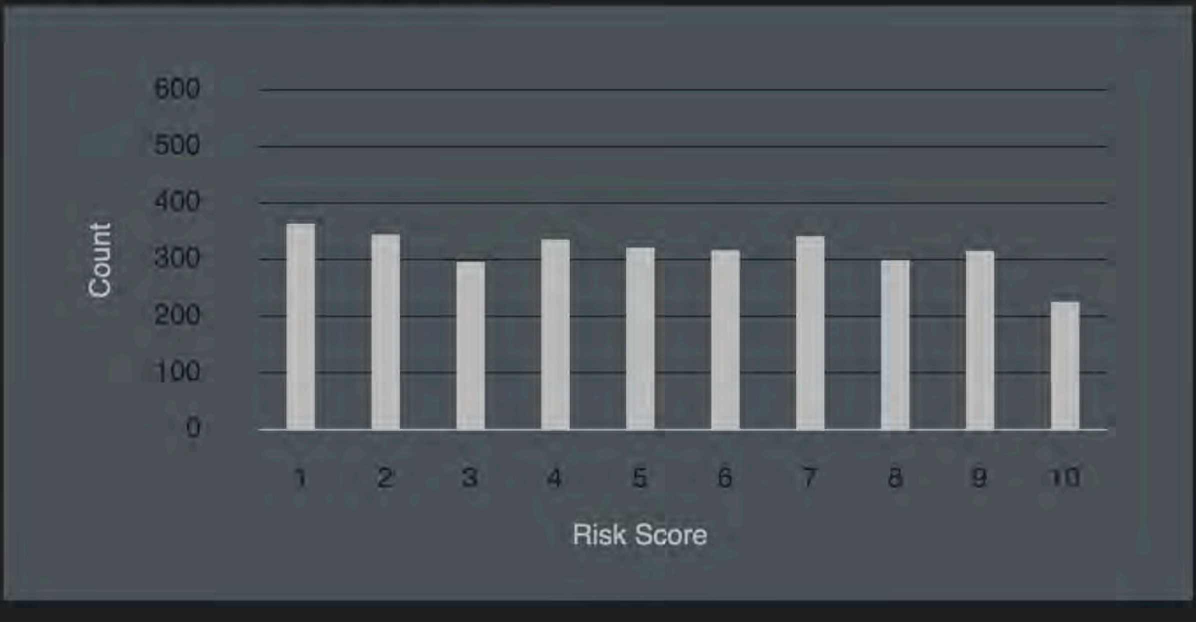
Most modern risk tools were originally designed to provide judges with insight into the types of treatment that an individual might need — from drug treatment to mental health counseling.

"What it tells the judge is that if I put you on probation, I'm going to need to give you a lot of services or you're probably going to fail," said Edward Latessa, a University of Cincinnati professor who is the author of a risk assessment tool that is used in Ohio and several other states.

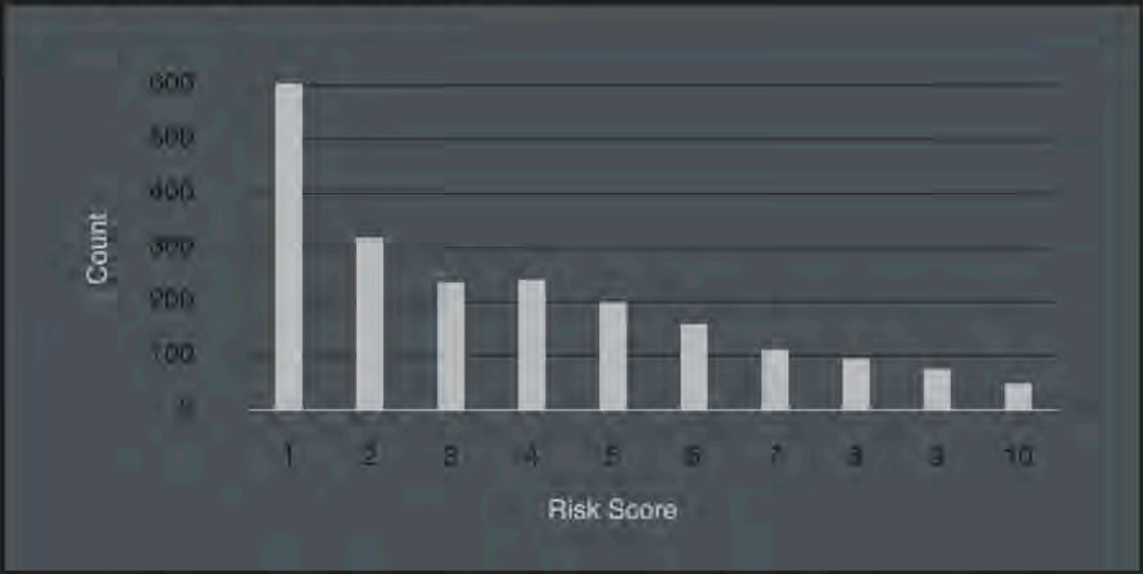
But being judged ineligible for alternative treatment — particularly during a sentencing hearing — can translate into incarceration. Defendants rarely have an opportunity to challenge their assessments. The results are usually shared with the defendant's attorney, but the calculations that transformed the underlying data into a score are rarely revealed.

"Risk assessments should be impermissible unless both parties get to see all the data that go into them," said Christopher Slobogin, director of the criminal justice program at Vanderbilt Law School. "It should be an open, full-court adversarial proceeding."

Black Defendants' Risk Scores



White Defendants' Risk Scores



These charts show that scores for white defendants were skewed toward lower-risk categories. Scores for black defendants were not. *(Source: ProPublica analysis of data from Broward County, Fla.)*

Proponents of risk scores argue they can be used to reduce the rate of incarceration. In 2002, Virginia became one of the first states to begin using a risk assessment tool in the sentencing of nonviolent felony offenders statewide. In 2014, Virginia judges using the tool sent nearly half of those defendants to alternatives to prison, according to a state sentencing commission report. Since 2005, the state's prison population growth has slowed to 5 percent from a rate of 31 percent the previous decade.

In some jurisdictions, such as Napa County, California, the probation department uses risk assessments to suggest to the judge an appropriate probation or treatment plan for individuals being sentenced. Napa County Superior Court Judge Mark Boessenecker said he finds the recommendations helpful. "We have a dearth of good treatment programs, so filling a slot in a program with someone who doesn't need it is foolish," he said.

However, Boessenecker, who trains other judges around the state in evidence-based sentencing, cautions his colleagues that the score doesn't necessarily reveal whether a person is dangerous or if they should go to prison.

"A guy who has molested a small child every day for a year could still come out as a low risk because he probably has a job," Boessenecker said. "Meanwhile, a drunk guy will look high risk because he's homeless. These risk factors don't tell you whether the guy ought to go to prison or not; the risk factors tell you more about what the probation conditions ought to be."

Sometimes, the scores make little sense even to defendants.

James Rivelli, a 54-year old Hollywood, Florida, man, was arrested two years ago for shoplifting seven boxes of Crest Whitestrips from a CVS drugstore. Despite a criminal record that included aggravated assault, multiple thefts and felony drug trafficking, the Northpointe algorithm classified him as being at a low risk of reoffending.

“I am surprised it is so low,” Rivelli said when told by a reporter he had been rated a 3 out of a possible 10. “I spent five years in state prison in Massachusetts. But I guess they don’t count that here in Broward County.” In fact, criminal records from across the nation are supposed to be included in risk assessments.

“I’m surprised [my risk score] is so low. I spent five years in state prison in Massachusetts.” (Josh Ritchie for ProPublica)

Less than a year later, he was charged with two felony counts for shoplifting about $1,000 worth of tools from Home Depot. He said his crimes were fueled by drug addiction and that he is now sober.

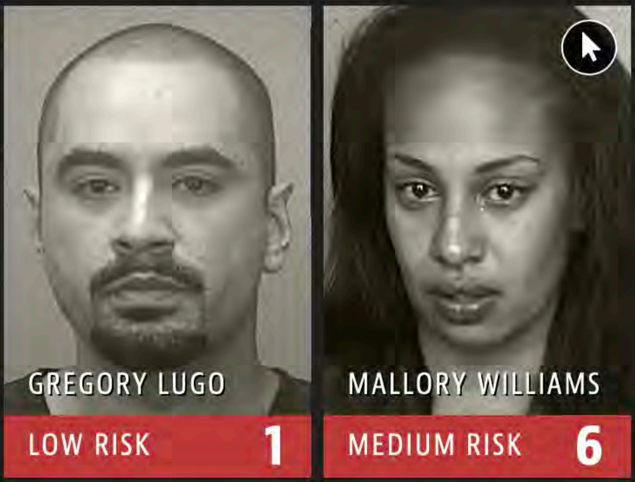
NORTHPOINTE WAS FOUNDED in 1989 by Tim Brennan, then a professor of statistics at the University of Colorado, and Dave Wells, who was running a corrections program in Traverse City, Michigan.

Wells had built a prisoner classification system for his jail. “It was a beautiful piece of work,” Brennan said in an interview conducted before ProPublica had completed its analysis. Brennan and Wells shared a love for what Brennan called “quantitative taxonomy” — the measurement of personality traits such as intelligence, extroversion and introversion. The two decided to build a risk assessment score for the corrections industry.

Brennan wanted to improve on a leading risk assessment score, the LSI, or Level of Service Inventory, which had been developed in Canada. “I found a fair amount of weakness in the LSI,” Brennan said. He wanted a tool that addressed the major theories about the causes of crime.

Brennan and Wells named their product the Correctional Offender Management Profiling for Alternative Sanctions, or COMPAS. It assesses not just risk but also nearly two dozen so-called “criminogenic needs” that relate to the major theories of criminality, including “criminal personality,” “social isolation,” “substance abuse” and “residence/stability.” Defendants are ranked low, medium or high risk in each category.

Two DUI Arrests



Lugo crashed his Lincoln Navigator into a Toyota Camry while drunk. He was rated as a low risk of reoffending despite the fact that it was at least his fourth DUI.

As often happens with risk assessment tools, many jurisdictions have adopted Northpointe's software before rigorously testing whether it works. New York State, for instance, started using the tool to assess people on probation in a pilot project in 2001 and rolled it out to the rest of the state's probation departments — except New York City — by 2010. The state didn't publish a comprehensive statistical evaluation of the tool until 2012. The study of more than 16,000 probationers found the tool was 71 percent accurate, but it did not evaluate racial differences.

A spokeswoman for the New York state division of criminal justice services said the study did not examine race because it only sought to test whether the tool had been properly calibrated to fit New York's probation population. She also said judges in nearly all New York counties are given defendants' Northpointe assessments during sentencing.

In 2009, Brennan and two colleagues published a validation study that found that Northpointe's risk of recidivism score had an accuracy rate of 68 percent in a sample of 2,328 people. Their study also found that the score was slightly less predictive for black men than white men — 67 percent versus 69 percent. It did not examine racial disparities beyond that, including whether some groups were more likely to be wrongly labeled higher risk.

Brennan said it is difficult to construct a score that doesn't include items that can be correlated with race — such as poverty, joblessness and social marginalization. "If those are omitted from your risk assessment, accuracy goes down," he said.

In 2011, Brennan and Wells sold Northpointe to Toronto-based conglomerate Constellation Software for an undisclosed sum.

Wisconsin has been among the most eager and expansive users of Northpointe's risk assessment tool in sentencing decisions. In 2012, the Wisconsin Department of Corrections launched the use of the software throughout the state. It is used at each step in the prison system, from sentencing to parole.

In a 2012 presentation, corrections official Jared Hoy described the system as a "giant correctional pinball machine" in which correctional officers could use the scores at every "decision point."

Wisconsin has not yet completed a statistical validation study of the tool and has not said when one might be released. State corrections officials declined repeated requests to comment for this article.

Some Wisconsin counties use other risk assessment tools at arrest to determine if a defendant is too risky for pretrial release. Once a defendant is convicted of a felony anywhere in the state, the Department of Corrections attaches Northpointe's assessment to the confidential presentence report given to judges, according to Hoy's presentation.

In theory, judges are not supposed to give longer sentences to defendants with higher risk scores. Rather, they are supposed to use the tests primarily to determine which defendants are eligible for probation or treatment programs.

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

But judges have cited scores in their sentencing decisions. In August 2013, Judge Scott Horne in La Crosse County, Wisconsin, declared that defendant Eric Loomis had been "identified, through the COMPAS assessment, as an individual who is at high risk to the community." The judge then imposed a sentence of eight years and six months in prison.

Loomis, who was charged with driving a stolen vehicle and fleeing from police, is challenging the use of the score at sentencing as a violation of his due process rights. The state has defended Horne's use of the score with the argument that judges can consider the score in addition to other factors. It has also stopped including scores in presentencing reports until the state Supreme Court decides the case.

"The risk score alone should not determine the sentence of an offender," Wisconsin Assistant Attorney General Christine Remington said last month during state Supreme Court arguments in the Loomis case. "We don't want courts to say, this person in front of me is a 10 on COMPAS as far as risk, and therefore I'm going to give him the maximum sentence."

That is almost exactly what happened to Zilly, the 48-year-old construction worker sent to prison for stealing a push lawnmower and some tools he intended to sell for parts. Zilly has long struggled with a meth habit. In 2012, he had been working toward recovery with the help of a Christian pastor when he relapsed and committed the thefts.

After Zilly was scored as a high risk for violent recidivism and sent to prison, a public defender appealed the sentence and called the score's creator, Brennan, as a witness.

Brennan testified that he didn't design his software to be used in sentencing. "I wanted to stay away from the courts," Brennan said, explaining that his focus was on reducing crime rather than punishment. "But as time went on I started realizing that so many decisions are made, you know, in the courts. So I gradually softened on whether this could be used in the courts or not."

"Not that I'm innocent, but I just believe people do change." (Stephen Maturen for ProPublica)

Still, Brennan testified, "I don't like the idea myself of COMPAS being the sole evidence that a decision would be based upon."

After Brennan's testimony, Judge Babler reduced Zilly's sentence, from two years in prison to 18 months. "Had I not had the COMPAS, I believe it would likely be that I would have given one year, six months," the judge said at an appeals hearing on Nov. 14, 2013.

Zilly said the score didn't take into account all the changes he was making in his life — his conversion to Christianity, his struggle to quit using drugs and his efforts to be more available for his son. "Not that I'm innocent, but I just believe people do change."

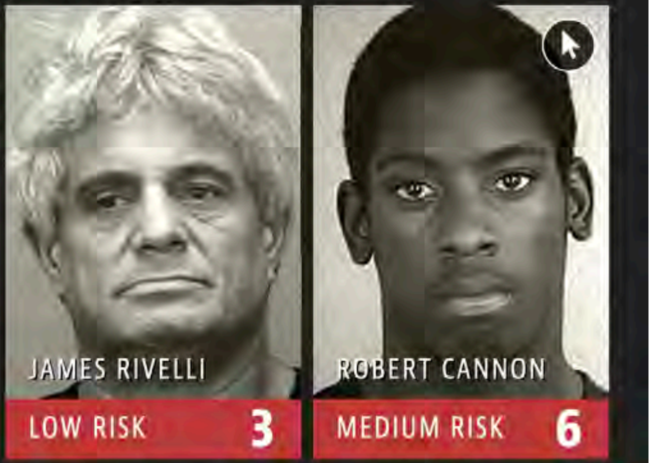
FLORIDA'S BROWARD COUNTY, where Brisha Borden stole the Huffy bike and was scored as high risk, does not use risk assessments in sentencing. "We don't think the [risk assessment] factors have any bearing on a sentence," said David Scharf, executive director of community programs for the Broward County Sheriff's Office in Fort Lauderdale.

Broward County has, however, adopted the score in pretrial hearings, in the hope of addressing jail overcrowding. A court-appointed monitor has overseen Broward County's jails since 1994 as a result of the settlement of a lawsuit brought by inmates in the 1970s. Even now, years later, the Broward County jail system is often more than 85 percent full, Scharf said.

In 2008, the sheriff's office decided that instead of building another jail, it would begin using Northpointe's risk scores to help identify which defendants were low risk enough to be released on bail pending trial. Since then, nearly everyone arrested in Broward has been scored soon after being booked. (People charged with murder and other capital crimes are not scored because they are not eligible for pretrial release.)

The scores are provided to the judges who decide which defendants can be released from jail. "My feeling is that if they don't need them to be in jail, let's get them out of there," Scharf said.

Two Shoplifting Arrests



After Rivelli stole from a CVS and was caught with heroin in his car, he was rated a low risk. He later shoplifted $1,000 worth of tools from a Home Depot.

Scharf said the county chose Northpointe's software over other tools because it was easy to use and produced "simple yet effective charts and graphs for judicial review." He said the system costs about $22,000 a year.

In 2010, researchers at Florida State University examined the use of Northpointe's system in Broward County over a 12-month period and concluded that its predictive accuracy was "equivalent" in assessing defendants of different races. Like others, they did not examine whether different races were classified differently as low or high risk.

Scharf said the county would review ProPublica's findings. "We'll really look at them up close," he said.

Broward County Judge John Hurley, who oversees most of the pretrial release hearings, said the scores were helpful when he was a new judge, but now that he has experience he prefers to rely on his own judgment. "I haven't relied on COMPAS in a couple years," he said.

Hurley said he relies on factors including a person's prior criminal record, the type of crime committed, ties to the community, and their history of failing to appear at court proceedings.

ProPublica's analysis reveals that higher Northpointe scores are slightly correlated with longer pretrial incarceration in Broward County. But there are many reasons that could be true other than judges being swayed by the scores — people with higher risk scores may also be poorer and have difficulty paying bond, for example.

Most crimes are presented to the judge with a recommended bond amount, but he or she can adjust the amount. Hurley said he often releases first-time or low-level offenders without any bond at all.

However, in the case of Borden and her friend Sade Jones, the teenage girls who stole a kid's bike and scooter, Hurley raised the bond amount for each girl from the recommended $0 to $1,000 each.

Hurley said he has no recollection of the case and cannot recall if the scores influenced his decision.

Sade Jones, who had never been arrested before, was rated a medium risk. *(Josh Ritchie for ProPublica)*

The girls spent two nights in jail before being released on bond.

"We literally sat there and cried" the whole time they were in jail, Jones recalled. The girls were kept in the same cell. Otherwise, Jones said, "I would have gone crazy." Borden declined repeated requests to comment for this article.

Jones, who had never been arrested before, was rated a medium risk. She completed probation and got the felony burglary charge reduced to misdemeanor trespassing, but she has still struggled to find work.

"I went to McDonald's and a dollar store, and they all said no because of my background," she said. "It's all kind of difficult and unnecessary."

Julia Angwin is a senior reporter at ProPublica. From 2000 to 2013, she was a reporter at The Wall Street Journal, where she led a privacy investigative team that was a finalist for a Pulitzer Prize in Explanatory Reporting in 2011 and won a Gerald Loeb Award in 2010.

Jeff Larson is the Data Editor at ProPublica. He is a winner of the Livingston Award for the 2011 series Redistricting: How Powerful Interests are Drawing You Out of a Vote. Jeff's public key can be found here.

Lauren Kirchner is a senior reporting fellow at ProPublica. Surya Mattu is a contributing researcher. Design and production by Rob Weychert and David Sleight.

DOI:10.1145/3376898

A group of industry, academic, and government experts convene in Philadelphia to explore the roots of algorithmic bias.

BY ALEXANDRA CHOULDECHOVA AND AARON ROTH

A Snapshot of the Frontiers of Fairness in Machine Learning

THE LAST DECADE has seen a vast increase both in the diversity of applications to which machine learning is applied, and to the import of those applications. Machine learning is no longer just the engine behind ad placements and spam filters; it is now used to filter loan applicants, deploy police officers, and inform bail and parole decisions, among other things. The result has been a major concern for the potential for data-driven methods to introduce and perpetuate discriminatory practices, and to otherwise be unfair. And this concern has not been without reason: a steady stream of empirical findings has shown that data-driven methods can unintentionally both encode existing human biases and introduce new ones.[7,9,11,60]

At the same time, the last two years have seen an unprecedented explosion in interest from the academic community in studying fairness and machine learning. "Fairness and transparency" transformed from a niche topic with a trickle of papers produced every year (at least since the work of Pedresh[56] to a major subfield of machine learning, complete with a dedicated archival conference—ACM FAT*). But despite the volume and velocity of published work, our understanding of the fundamental questions related to fairness and machine learning remain in its infancy. What should fairness mean? What are the causes that introduce unfairness in machine learning? How best should we modify our algorithms to avoid unfairness? And what are the corresponding trade offs with which we must grapple?

In March 2018, we convened a group of about 50 experts in Philadelphia, drawn from academia, industry, and government, to assess the state of our understanding of the fundamentals of the nascent science of fairness in machine learning, and to identify the unanswered questions that seem the most pressing. By necessity, the aim of the workshop was not to comprehensively cover the vast growing field, much of which is empirical. Instead, the focus was on theoretical work aimed at providing a scientific foundation for understanding algo-

» key insights

- **The algorithmic fairness literature is enormous and growing quickly, but our understanding of basic questions remains nascent.**
- **Researchers have yet to find entirely compelling definitions, and current work focuses mostly on supervised learning in static settings.**
- **There are many compelling open questions related to robustly accounting for the effects of interventions in dynamic settings, learning in the presence of data contaminated with human bias, and finding definitions of fairness that guarantee individual-level semantics while remaining actionable.**

ILLUSTRATION BY JUSTIN METZ

rithmic bias. This document captures several of the key ideas and directions discussed. It is not an exhaustive account of work in the area.

What We Know

Even before we precisely specify what we mean by "fairness," we can identify common distortions that can lead off-the-shelf machine learning techniques to produce behavior that is intuitively unfair. These include:

1. *Bias encoded in data.* Often, the training data we have on hand already includes human biases. For example, in the problem of recidivism prediction used to inform bail and parole decisions, the goal is to predict whether an inmate, if released, will go on to commit another crime within a fixed period of time. But we do not have data on who commits crimes—we have data on who is arrested. There is reason to believe that arrest data—especially for drug crimes—is skewed toward minority populations that are policed at a higher rate.[59] Of course, machine learning techniques are designed to fit the data, and so will naturally replicate any bias already present in the data. There is no reason to expect them to remove existing bias.

2. *Minimizing average error fits majority populations.* Different populations of people have different distributions over features, and those features have different relationships to the label that we are trying to predict. As an example, consider the task of predicting college performance based on high school data. Suppose there is a majority population and a minority population. The majority population employs SAT tutors and takes the exam multiple times, reporting only the highest score. The minority population does not. We should naturally expect both that SAT scores are higher among the majority population, and that their relationship to college performance is differently calibrated compared to the minority population. But if we train a group-blind classifier to minimize overall error, if it cannot simultaneously fit both populations optimally, it will fit the majority population. This is because—simply by virtue of their numbers—the fit to the majority population is more important to overall error than the fit to the minority population. This leads to a different (and higher) distribution of errors in the minority population. This effect can be quantified and can be partially alleviated via concerted data gathering effort.[14]

Given the limitations of extant notions of fairness, is there a way to get some of the "best of both worlds?"

3. *The need to explore.* In many important problems, including recidivism prediction and drug trials, the data fed into the prediction algorithm depends on the actions that algorithm has taken in the past. We only observe whether an inmate will recidivate if we release him. We only observe the efficacy of a drug on patients to whom it is assigned. Learning theory tells us that in order to effectively learn in such scenarios, we need to explore—that is, sometimes take actions we believe to be sub-optimal in order to gather more data. This leads to at least two distinct ethical questions. First, when are the individual costs of exploration borne disproportionately by a certain sub-population? Second, if in certain (for example, medical) scenarios, we view it as immoral to take actions we believe to be sub-optimal for any particular patient, how much does this slow learning, and does this lead to other sorts of unfairness?

Definitions of fairness. With a few exceptions, the vast majority of work to date on fairness in machine learning has focused on the task of batch classification. At a high level, this literature has focused on two main families of definitions:[a] statistical notions of fairness and individual notions of fairness. We briefly review what is known about these approaches to fairness, their advantages, and their shortcomings.

Statistical definitions of fairness. Most of the literature on fair classification focuses on statistical definitions of fairness. This family of definitions fixes a small number of protected demographic groups G (such as racial groups), and then ask for (approximate) parity of some statistical measure across all of these groups. Popular measures include raw positive classification rate, considered in

[a] There is also an emerging line of work that considers causal notions of fairness (for example, see Kilbertus,[43] Kusner,[48] Nabi[55]). We intentionally avoided discussions of this potentially important direction because it will be the subject of its own CCC visioning workshop.

work such as Calders,[10] Dwork,[19] Feldman,[25] Kamishima,[36] (also sometimes known as statistical parity,[19] false positive and false negative rates[15,29,46,63] (also sometimes known as equalized odds[29]), and positive predictive value[15,46] (closely related to equalized calibration when working with real valued risk scores). There are others—see, for example, Berk[4] for a more exhaustive enumeration.

This family of fairness definitions is attractive because it is simple, and definitions from this family can be achieved without making any assumptions on the data and can be easily verified. However, statistical definitions of fairness do not on their own give meaningful guarantees to individuals or structured subgroups of the protected demographic groups. Instead they give guarantees to "average" members of the protected groups. (See Dwork[19] for a litany of ways in which statistical parity and similar notions can fail to provide meaningful guarantees, and Kearns[40] for examples of how some of these weaknesses carry over to definitions that equalize false positive and negative rates.) Different statistical measures of fairness can be at odds with one another. For example, Chouldechova[15] and Kleinberg[46] prove a fundamental impossibility result: except in trivial settings, it is impossible to simultaneously equalize false positive rates, false negative rates, and positive predictive value across protected groups. Learning subject to statistical fairness constraints can also be computationally hard,[61] although practical algorithms of various sorts are known.[1,29,63]

Individual definitions of fairness. Individual notions of fairness, on the other hand, ask for constraints that bind on specific pairs of individuals, rather than on a quantity that is averaged over groups. For example, Dwork[19] gives a definition which roughly corresponds to the constraint that "similar individuals should be treated similarly," where similarity is defined with respect to a task-specific metric that must be determined on a case by case basis. Joseph[35] suggests a definition that corresponds approximately to "less qualified individuals should not be favored over more qualified individuals," where quality is defined with respect to the true underlying label (unknown to the algorithm). However, although the semantics of these kinds of definitions can be more meaningful than statistical approaches to fairness, the major stumbling block is that they seem to require making significant assumptions. For example, the approach of Dwork[19] presupposes the existence of an agreed upon similarity metric, whose definition would itself seemingly require solving a non-trivial problem in fairness, and the approach of Joseph[35] seems to require strong assumptions on the functional form of the relationship between features and labels in order to be usefully put into practice. These obstacles are serious enough that it remains unclear whether individual notions of fairness can be made practical—although attempting to bridge this gap is an important and ongoing research agenda.

Questions at the Research Frontier

Given the limitations of extant notions of fairness, is there a way to get some of the "best of both worlds?" In other words, constraints that are practically implementable without the need for making strong assumptions on the data or the knowledge of the algorithm designer, but which nevertheless provide more meaningful guarantees to individuals? Two recent papers, Kearns[40] and Hèbert-Johnson[30] (see also Kearns[42] and Kim[44] for empirical evaluations of the algorithms proposed in these papers), attempt to do this by asking for statistical fairness definitions to hold not just on a small number of protected groups, but on an exponential or infinite class of groups defined by some class of functions of bounded complexity. This approach seems promising—because, ultimately, they are asking for statistical notions of fairness—the approaches proposed by these papers enjoy the benefits of statistical fairness: that no assumptions need be made about the data, nor is any external knowledge (like a fairness metric) needed. It also better addresses concerns about "intersectionality," a term used to describe how different kinds of discrimination can compound and interact for individuals who fall at the intersection of several protected classes.

At the same time, the approach raises a number of additional questions: What function classes are reasonable, and once one is decided upon (for example, conjunctions of protected attributes), what features should be "protected?" Should these only be attributes that are sensitive on their own, like race and gender, or might attributes that are innocuous on their own correspond to groups we wish to protect once we consider their intersection with protected attributes (for example clothing styles intersected with race or gender)? Finally, this family of approaches significantly mitigates some of the weaknesses of statistical notions of fairness by asking for the constraints to hold on average not just over a small number of coarsely defined groups, but over very finely defined groups as well. Ultimately, however, it inherits the weaknesses of statistical fairness as well, just on a more limited scale.

Another recent line of work aims to weaken the strongest assumption needed for the notion of individual fairness from Dwork:[19] namely the algorithm designer has perfect knowledge of a "fairness metric." Kim[45] assumes the algorithm has access to an oracle which can return an unbiased estimator for the distance between two randomly drawn individuals according to an unknown fairness metric, and show how to use this to ensure a statistical notion of fairness related to Hèbert-Johnson[30] and Kearns,[40] which informally state that "on average, individuals in two groups should be treated similarly if on average the individuals in the two groups are similar" and this can be achieved with respect to an exponentially or infinitely large set of groups. Similarly, Gillen[28] assumes the existence of an oracle, which can identify fairness violations when they are made in an online setting but cannot quantify the extent of the violation (with respect to the unknown metric). It is shown that when the metric is from a specific learnable family, this kind of feedback is sufficient to obtain an optimal regret bound to the best fair classifier while having only a bounded number of violations of the fairness metric. Rothblum[58] considers the case in which the metric is known and show that a PAC-inspired approximate variant of metric fairness generalizes to new data drawn from the same underlying distribution. Ultimately, however, these approaches all assume fairness is perfectly defined with respect to some metric, and that there is some sort of direct access to it. Can these approaches be generalized to a more "agnostic" setting, in which fairness feedback is given by human beings who may not be responding in a way that is consistent with any metric?

Data evolution and dynamics of fairness. The vast majority of work in computer science on algorithmic fairness has focused on one-shot classification tasks. But real algorithmic systems consist of many different components combined together, and operate in complex environments that are dynamically changing, sometimes because of the actions of the learning algorithm itself. For the field to progress, we need to understand the dynamics of fairness in more complex systems.

Perhaps the simplest aspect of dynamics that remains poorly understood is how and when components that may individually satisfy notions of fairness compose into larger constructs that still satisfy fairness guarantees. For example, if the bidders in an advertising auction individually are fair with respect to their bidding decisions, when will the allocation of advertisements be fair, and when will it not? Bower[8] and Dwork[20] have made a preliminary foray in this direction. These papers embark on a systematic study of fairness under composition and find that often the composition of multiple fair components will not satisfy any fairness constraint at all. Similarly, the individual components of a fair system may appear to be unfair in isolation. There are certain special settings, for example, the "filtering pipeline" scenario of Bower[8]—modeling a scenario in which a job applicant is selected only if she is selected at every stage of the pipeline—in which (multiplicative approximations of) statistical fairness notions compose in a well behaved way. But the high-level message from these works is that our current notions of fairness compose poorly. Experience from differential privacy[21,22] suggests that graceful degradation under composition is key to designing complicated algorithms satisfying desirabl e statistical properties, because it allows algorithm design and analysis to be modular. Thus, it seems important to find satisfying fairness definitions and richer frameworks that behave well under composition.

In dealing with socio-technical systems, it is also important to understand how algorithms dynamically effect their environment, and the incentives of human actors. For example, if the bar (for example, college admission) is lowered for a group of individuals, this might increase the average qualifications for this group over time because of at least two effects: a larger proportion of children in the next generation grow up in households with college educated parents (and the opportunities this provides), and the fact that a college education is achievable can incentivize effort to prepare academically. These kinds of effects are not considered when considering either statistical or individual notions of fairness in one-shot learning settings.

The economics literature on affirmative action has long considered such effects—although not with the specifics of machine learning in mind: see, for example, Becker,[3] Coat,[16] Foster.[26] More recently, there have been some preliminary attempts to model these kinds of effects in machine learning settings—for example, by modeling the environment as a Markov decision process,[32] considering the equilibrium effects of imposing statistical definitions of fairness in a model of a labor market,[31] specifying the functional relationship between classification outcomes and quality,[49] or by considering the effect of a classifier on a downstream Bayesian decision maker.[39] However, the specific predictions of most of the models of this sort are brittle to the specific modeling assumptions made—they point to the need to consider long term dynamics, but do not provide robust guidance for how to navigate them. More work is needed here.

Finally, decision making is often distributed between a large number of actors who share different goals and do not necessarily coordinate. In settings like this, in which we do not have direct control over the decision-making process, it is important to think about how to incentivize rational agents to behave in a way that we view as fair. Kannan[37] takes a preliminary stab at this task, showing how to incentivize a particular notion of individual fairness in a simple, stylized setting, using small monetary payments. But how should this work for other notions of fairness, and in more complex settings? Can this be done by controlling the flow of information, rather than by making monetary payments (monetary payments might be distasteful in various fairness-relevant settings)? More work is needed here as well. Finally, Corbett-Davies[17] take a welfare maximization view of fairness in classification and characterize the cost of imposing additional statistical fairness constraints as well. But this is done in a static environment. How would the conclusions change under a dynamic model?

Modeling and correcting bias in the data. Fairness concerns typically surface precisely in settings where the available training data is already contaminated by bias. The data itself is often a product of social and historical process that operated to the disadvantage of certain groups. When trained in such data, off-the-shelf machine learning techniques may reproduce, reinforce, and potentially exacerbate existing biases. Understanding how bias arises in the data, and how to correct for it, are fundamental challenges in the study of fairness in machine learning.

Bolukbasi[7] demonstrate how machine learning can reproduce biases in their analysis of the popular word2vec embedding trained on a corpus of Google News texts (parallel effects were independently discovered by Caliskan[11]). The authors show that the trained embedding exhibit female/male gender stereotypes, learning that "doctor" is more similar to man than to woman, along with analogies such as "man is to computer programmer as woman is to homemaker." Even if such learned associations accurately reflect patterns in the source text corpus, their use in automated systems may exacerbate existing biases. For instance, it might result in male applicants being ranked more highly than equally qualified female applicants in queries related to jobs that the embedding identifies as male-associated.

Similar risks arise whenever there is potential for feedback loops. These are situations where the trained machine learning model informs decisions that then affect the data collected for future iterations of the training process. Lum[51] demonstrate how feedback loops might arise in predictive policing if arrest data were used to train the model.[b] In a nutshell, since police are likely to make more arrests in more heavily policed areas, using arrest data to predict crime hotspots will disproportionately concentrate policing efforts on already over-policed communities. Expanding on this analysis, Ensign[24] finds that incorporating community-driven data, such as crime reporting, helps to attenuate the biasing feedback effects. The authors also propose a strategy for accounting for feedback by adjusting arrest counts for policing intensity. The success of the mitigation strategy, of course, depends on how well the simple theoretical model reflects the true relationships between crime intensity, policing, and arrests. Problematically, such relationships are often unknown, and are very difficult to infer from data. This situation is by no means specific to predictive policing.

Correcting for data bias generally seems to require knowledge of how the measurement process is biased, or judgments about properties the data would satisfy in an "unbiased" world. Friedler[27] formalize this as a disconnect between the *observed space*—features that are observed in the data, such as SAT scores—and the unobservable *construct space*—features that form the desired basis for decision making, such as intelligence. Within this framework, data correction efforts attempt to undo the effects of biasing mechanisms that drive discrepancies between these spaces. To the extent that the biasing mechanism cannot be inferred empirically, any correction effort must make explicit its underlying assumptions about this mechanism. What precisely is being assumed about the construct space? When can the mapping between the construct space and the observed space be learned and inverted? What form of fairness does the correction promote, and at what cost? The costs are often immediately realized, whereas the benefits are less tangible. We will directly observe reductions in prediction accuracy, but any gains hinge on a belief that the observed world is not one we should seek to replicate accurately in the first place. This is an area where tools from causality may offer a principled approach for drawing valid inference with respect to unobserved counterfactually 'fair' worlds.

Fairness concerns typically surface precisely in settings where the available training data is already contaminated by bias.

Fair representations. Fair representation learning is a data debiasing process that produces transformations (intermediate representations) of the original data that retain as much of the task-relevant information as possible while removing information about sensitive or protected attributes. This is one approach to transforming biased observational data in which group membership may be inferred from other features, to a construct space where protected attributes are statistically independent of other features.

First introduced in the work of Zemel[64] fair representation learning produces a debiased data set that may in principle be used by other parties without any risk of disparate outcomes. Feldman[25] and McNamara[54] formalize this idea by showing how the disparate impact of a decision rule is bounded in terms of its balanced error rate as a predictor of the sensitive attribute.

Several recent papers have introduced new approaches for constructing fair representations. Feldman[25] propose rank-preserving procedures for repairing features to reduce or remove pairwise dependence with the protected attribute. Johndrow[33] build upon this work, introducing a likelihood-based approach that can additionally handle continuous protected attributes, discrete features, and which promotes joint independence

[b] Predictive policing models are generally proprietary, and so it is not clear whether arrest data is used to train the model in any deployed system.

between the transformed features and the protected attributes. There is also a growing literature on using adversarial learning to achieve group fairness in the form of statistical parity or false positive/false negative rate balance.[5,23,52,65]

Existing theory shows the fairness-promoting benefits of fair-representation learning rely critically on the extent to which existing associations between the transformed features and the protected characteristics are removed. Adversarial downstream users may be able to recover protected attribute information if their models are more powerful than those used initially to obfuscate the data. This presents a challenge both to the generators of fair representations as well as to auditors and regulators tasked with certifying that the resulting data is fair for use. More work is needed to understand the implications of fair representation learning for promoting fairness in the real world.

Beyond classification. Although the majority of the work on fairness in machine learning focuses on batch classification, it is but one aspect of how machine learning is used. Much of machine learning—for example, online learning, bandit learning, and reinforcement learning—focuses on dynamic settings in which the actions of the algorithm feed back into the data it observes. These dynamic settings capture many problems for which fairness is a concern. For example, lending, criminal recidivism prediction, and sequential drug trials are so-called bandit learning problems, in which the algorithm cannot observe data corresponding to counterfactuals. We cannot see whether someone not granted a loan would have paid it back. We cannot see whether an inmate not released on parole would have gone on to commit another crime. We cannot see how a patient would have responded to a different drug.

Much of machine learning focuses on dynamic settings in which the actions of the algorithm feed back into the data it observes. These dynamic settings capture many problems for which fairness is a concern.

The theory of learning in bandit settings is well understood, and it is characterized by a need to trade-off exploration with exploitation. Rather than always making a myopically optimal decision, when counterfactuals cannot be observed, it is necessary for algorithms to sometimes take actions that appear to be sub-optimal so as to gather more data. But in settings in which decisions correspond to individuals, this means sacrificing the well-being of a particular person for the potential benefit of future individuals. This can sometimes be unethical, and a source of unfairness.[6] Several recent papers explore this issue. For example, Bastani[2] and Kannan[38] give conditions under which linear learners need not explore at all in bandit settings, thereby allowing for best-effort service to each arriving individual, obviating the tension between ethical treatment of individuals and learning. Raghavan[57] show the costs associated with exploration can be unfairly bourn by a structured sub-population, and that counter-intuitively, those costs can actually increase when they are included with a majority population, even though more data increases the rate of learning overall. However, these results are all preliminary: they are restricted to settings in which the learner is learning a linear policy, and the data really is governed by a linear model. While illustrative, more work is needed to understand real-world learning in online settings, and the ethics of exploration.

There is also some work on fairness in machine learning in other settings—for example, ranking,[12] selection,[42,47] personalization,[13] bandit learning,[34,50] human-classifier hybrid decision systems,[53] and reinforcement learning.[18,32] But outside of classification, the literature is relatively sparse. This should be rectified, because there are interesting and important fairness issues that arise in other settings—especially when there are combinatorial constraints on the set of individuals that can be selected for a task, or when there is a temporal aspect to learning.

Acknowledgments

This material is based upon work supposed by the National Science Foundation under Grant No. 1136993. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the views of the National Science Foundation.

We are indebted to all of the participants of the CCC visioning workshop; discussions from that meeting shaped every aspect of this document. Also, our thanks to Helen Wright, Ann Drobnis, Cynthia Dwork, Sampath Kannan, Michael Kearns, Toni Pitassi, and Suresh Venkatasubramanian.

References

1. Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J. and Wallach, H. A reductions approach to fair classification. In *Proceedings of the 35th Intern. Conf. Machine Learning*. ICML, JMLR Workshop and Conference Proceedings, 2018, 2569–2577.
2. Bastani, H., Bayati, M. and Khosravi, K. Exploiting the natural exploration in contextual bandits. arXiv preprint, 2017, arXiv:1704.09011.
3. Becker, G.S. *The Economics of Discrimination*. University of Chicago Press, 2010.
4. Berk, R., Heidari, H., Jabbari, S., Kearns, M. and Roth, A. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* (2018), 0(0):0049124118782533.
5. Beutel, A., Chen, J., Zhao, Z. and Chi, E.H. Data decisions and theoretical implications when adversarially learning fair representations. arXiv preprint, 2017, arXiv:1707.00075.
6. Bird, S., Barocas, S., Crawford, K., Diaz, F. and Wallach, H. 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.
7. Bolukbasi, T., Chang, K-W., Zou, J.Y., Saligrama, V. and Kalai, A.T. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*, 2016, 4349–4357.
8. Bower, A. et al. Fair pipelines. arXiv preprint, 2017, arXiv:1707.00391.
9. Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the Conference on Fairness, Accountability and Transparency*. ACM, 2018, 77–91.
10. Calders, T. and Verwer, S. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery 21*, 2 (2010), 277–292.
11. Caliskan, A., Bryson, J.J. and Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science 356*, 6334 (2017), 183–186.
12. Celis, L.E., Straszak, D. and Vishnoi, N.K. Ranking with fairness constraints. In *Proceedings of the 45th Intern. Colloquium on Automata, Languages, and Programming*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.
13. Celis, L.E. and Vishnoi, N.K. Fair personalization. arXiv preprint, 2017, arXiv:1707.02260.
14. Chen, I., Johansson, F.D. and Sontag, D. Why is my classifier discriminatory? *Advances in Neural Information Processing Systems*, 2018, 3539–3550.
15. Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data 5*, 2 (2017), 153–163.
16. Coat, S. and Loury, G.C. Will affirmative-action policies eliminate negative stereotypes? *The American Economic Review*, 1993, 1220–1240.
17. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S. and Huq, A. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD Intern. Conf. Knowledge Discovery and Data Mining*. ACM, 2017, 797¬–806.
18. Doroudi, S., Thomas, P.S. and Brunskill, E. Importance sampling for fair policy selection. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2017.
19. Dwork, C., Hardt, M., Pitassi, T., Reingold, O. and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conf*. ACM, 2012, 214–226.
20. Dwork, C. and Ilvento, C. Fairness under composition. Manuscript, 2018.
21. Dwork, C., McSherry, F., Nissim, K. and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Proceedings of Theory of Cryptography Conference*. Springer, 2006, 265–284.
22. Dwork, C. and Roth, A. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science 9*, 3–4 (2014), 211–407.
23. Edwards, H. and Storkey, A. Censoring representations with an adversary. arXiv preprint, 2015, arXiv:1511.05897.
24. Ensign, D., Friedler, S.A., Neville, S., Scheidegger, C. and Venkatasubramanian, S. Runaway feedback loops in predictive policing. In *Proceedings of 1st Conf. Fairness, Accountability and Transparency in Computer Science*. ACM, 2018.
25. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C. and Venkatasubramanian, S. Certifying and removing disparate impact. *Proceedings of KDD*, 2015.
26. Foster, D.P. and Vohra, R.A. An economic argument for affirmative action. *Rationality and Society 4*, 2 (1992), 176–188.
27. Friedler, S.A., Scheidegger, C. and Venkatasubramanian, S. On the (im) possibility of fairness. arXiv preprint, 2016, arXiv:1609.07236.
28. Gillen, S., Jung, C., Kearns, M. and Roth, A. Online learning with an unknown fairness metric. *Advances in Neural Information Processing Systems*, 2018.
29. Hardt, M., Price, E. and Srebro, N. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 2016, 3315–3323.
30. Hébert-Johnson, U., Kim, M.P., Reingold, O. and Rothblum, G.N. Calibration for the (computationally identifiable) masses. In *Proceedings of the 35th Intern. Conf. Machine Learning 80*. ICML, JMLR Workshop and Conference Proceedings, 2018, 2569–2577.
31. Hu, L. and Chen, Y. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. P.A. Champin, F.L. Gandon, M. Lalmas, and P.G. Ipeirotis, eds. ACM, 2018, 1389–1398.
32. Ja bbari, S., Joseph, M., Kearns, M., Morgenstern, J.H. and Roth, A. Fairness in reinforcement learning. In *Proceedings of the Intern. Conf. Machine Learning*, 2017, 1617–1626.
33. Johndrow, J.E. Lum, K. et al. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *The Annals of Applied Statistics 13*, 1 (2019), 189–220.
34. Joseph, M., Kearns, M., Morgenstern, J.H., Neel, S. and Roth. A. Fair algorithms for infinite and contextual bandits. In *Proceedings of AAAI/ACM Conference on AI, Ethics, and Society*, 2018.
35. Joseph, M., Kearns, M., Morgenstern, J.H. and Roth, A. Fairness in learning: Classic and contextual bandits. *Advances in Neural Information Processing Systems*, 2016, 325–333.
36. Kamishima, T., Akaho, S. and Sakuma, J. Fairness-aware learning through regularization approach. In *Proceedings of the IEEE 11th Intern. Conf. Data Mining Workshops*. IEEE, 2011, 643–650.
37. Kannan, S. et al. Fairness incentives for myopic agents. In *Proceedings of the 2017 ACM Conference on Economics and Computation*. ACM, 2017, 369–386.
38. Kannan, S., Morgenstern, J., Roth, A., Waggoner, B. and Wu, Z.S. A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. *Advances in Neural Information Processing Systems*, 2018.
39. Kannan, S., Roth, A. and Ziani, J. Downstream effects of affirmative action. In *Proceedings of the Conf. Fairness, Accountability, and Transparency*. ACM, 2019, 240–248.
40. Kearns, M.J., Neel, S., Roth, A. and Wu, Z.S. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the 35th International Conference on Machine Learning*. J.G. Dy and A. Krause, eds. JMLR Workshop and Conference Proceedings, ICML, 2018. 2569–2577.
41. Kearns, M., Neel, S., Roth, A. and Wu, Z.S. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conf. Fairness, Accountability, and Transparency*. ACM, 2019, 100–109.
42. Kearns, M., Roth, A. and Wu, Z.S. Meritocratic fairness for cross-population selection. In *Proceedings of International Conference on Machine Learning*, 2017, 1828–1836.
43. Kilbertus, N. et al. Avoiding discrimination through causal reasoning. *Advances in Neural Information Processing Systems*, 2017, 656–666.
44. Kim, M.P., Ghorbani, A. and Zou, J. Multiaccuracy: Blackbox postprocessing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 2019, 247–254.
45. Kim, M.P., Reingold, O. and Rothblum, G.N. Fairness through computationally bounded awareness. *Advances in Neural Information Processing Systems*, 2018.
46. Kleinberg, J.M., Mullainathan, S. and Raghavan,M. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference*, 2017.
47. Kleinberg, J. and Raghavan, M. Selection problems in the presence of implicit bias. In *Proceedings of the 9th Innovations in Theoretical Computer Science Conference 94*, 2018, 33. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.
48. Kusner, M.J., Loftus, J., Russell, C. and Silva, R. Counterfactual fairness. *Advances in Neural Information Processing Systems*, 2017, 4069–4079.
49. Liu, L.T., Dean, S., Rolf, E., Simchowitz, M. and Hardt, M. Delayed impact of fair machine learning. In *Proceedings of the 35th Intern. Conf. Machine Learning*. ICML, 2018.
50. Liu, Y., Radanovic, G., Dimitrakakis, C., Mandal, D. and Parkes, D.C. Calibrated fairness in bandits. arXiv preprint, 2017, arXiv:1707.01875.
51. Lum, K. and Isaac, W. To predict and serve? *Significance 13*, 5 (2016), 14–19.
52. Madras, D., Creager, E., Pitassi, T. and Zemel, R. Learning adversarially fair and transferable representations. In *Proceedings of Intern. Conf. Machine Learning*, 2018, 3381–3390.
53. Madras, D., Pitassi, T. and Zemel, R.S. Predict responsibly: Increasing fairness by learning to defer. CoRR, 2017, abs/1711.06664.
54. McNamara, D., Ong, C.S. and Williamson, R.C. Provably fair representations. arXiv preprint, 2017, arXiv:1710.04394.
55. Nabi, R. and Shpitser, I. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence 2018* (2018), 1931. NIH Public Access.
56. Pedreshi, D., Ruggieri, S. and Turini, F. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD Intern. Conf. Knowledge Discovery and Data Mining*. ACM, 2008, 560–568.
57. Raghavan, M., Slivkins, A., Wortman Vaughan, J. and Wu, Z.S. The unfair externalities of exploration. *Conference on Learning Theory*, 2018.
58. Rothblum, G.N. and Yona, G. Probably approximately metric-fair learning. In *Proceedings of the 35th Intern. Conf. Machine Learning*. JMLR Workshop and Conference Proceedings, ICML 80 (2018), 2569–2577.
59. Rothwell, J. How the war on drugs damages black social mobility. The Brookings Institution, Sept. 30, 2014.
60. Sweeney, L. Discrimination in online ad delivery. *Queue 11*, 3 (2013), 10.
61. Woodworth, B., Gunasekar, S., Ohannessian, M.I. and Srebro, N. Learning non-discriminatory predictors. In *Proceedings of Conf. Learning Theory*, 2017, 1920–1953.
62. Yang, K. and Stoyanovich, J. Measuring fairness in ranked outputs. In *Proceedings of the 29th Intern. Conf. Scientific and Statistical Database Management*. ACM, 2017, 22.
63. Zafar, M.B., Valera, I. Gomez-Rodriguez, M. and Gummadi, K.P. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th Intern. Conf. World Wide Web*. ACM, 2017, 1171–1180.
64. Zemel, R., Wu, Y., Swersky, K., Pitassi, T. and Dwork, C. Learning fair representations. In *Proceedings of ICML*, 2013.
65. Zhang, B.H., Lemoine, B. and Mitchell, M. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conf. AI, Ethics, and Society*. ACM, 2018, 335–340.

Alexandra Chouldechova (achould@cmu.edu) is Estella Loomis Assistant Professor of Statistics and Public Polict in the Heinz College at Carnegie Mellon University, Pittsburgh, PA, USA.

Aaron Roth (aaroth@cis.upenn.edu) is Class of 1940 Associate Professor in the Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA. Together with Michael Kearns, he is the author of *The Ethical Algorithm*.

Watch the authors discuss this work in the exclusive *Communications* video. https://cacm.acm.org/videos/frontiers-of-fairness

Week 2: Algorithmic Fairness

WE ARE AI
#4

All about that BIAS

TERMS OF USE

All the panels in this comic book are licensed CC 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

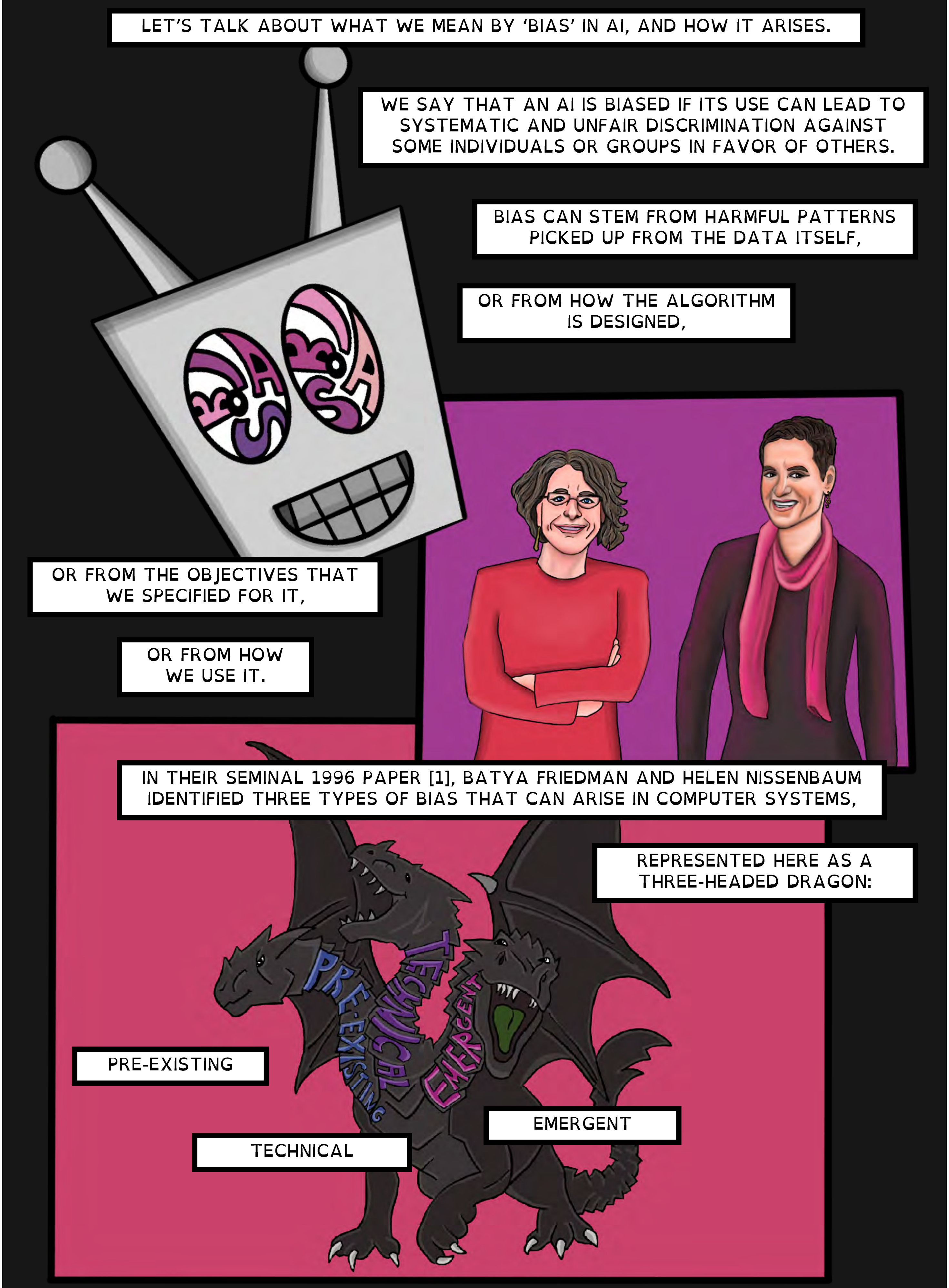
1. Provide the proper citation
2. Do not make modifications to the individual panels themselves

Cite as:

Julia Stoyanovich and Falaah Arif Khan. “All about that Bias”.
We are AI Comics, Vol 4 (2021)
https://dataresponsibly.github.io/we-are-ai/comics/vol4_en.pdf

Contact:

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



LET'S TALK ABOUT WHAT WE MEAN BY 'BIAS' IN AI, AND HOW IT ARISES.

WE SAY THAT AN AI IS BIASED IF ITS USE CAN LEAD TO SYSTEMATIC AND UNFAIR DISCRIMINATION AGAINST SOME INDIVIDUALS OR GROUPS IN FAVOR OF OTHERS.

BIAS CAN STEM FROM HARMFUL PATTERNS PICKED UP 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.

IN THEIR SEMINAL 1996 PAPER [1], BATYA FRIEDMAN AND HELEN NISSENBAUM IDENTIFIED THREE TYPES OF BIAS THAT CAN ARISE IN COMPUTER SYSTEMS,

REPRESENTED HERE AS A THREE-HEADED DRAGON:

PRE-EXISTING

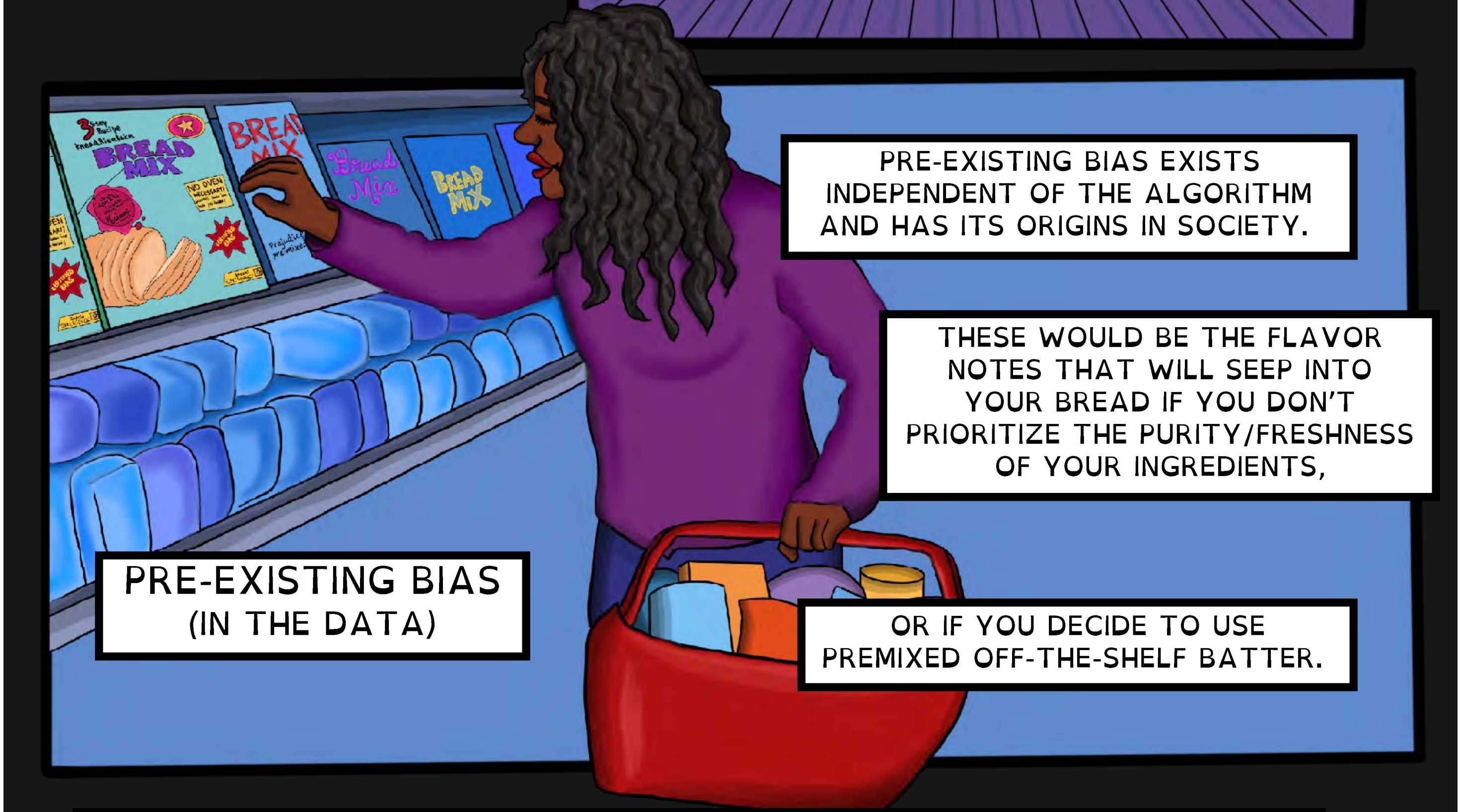
TECHNICAL

EMERGENT

[1] Batya Friedman and Helen Nissenbaum. (1996). Bias in computer systems.

RECALL THE BAKING METAPHOR WE USED TO UNDERSTAND DATA-DRIVEN ALGORITHMS IN VOLUME 1.

LET'S NOW USE THE SAME METAPHOR TO UNDERSTAND BIAS!



PRE-EXISTING BIAS
(IN THE DATA)

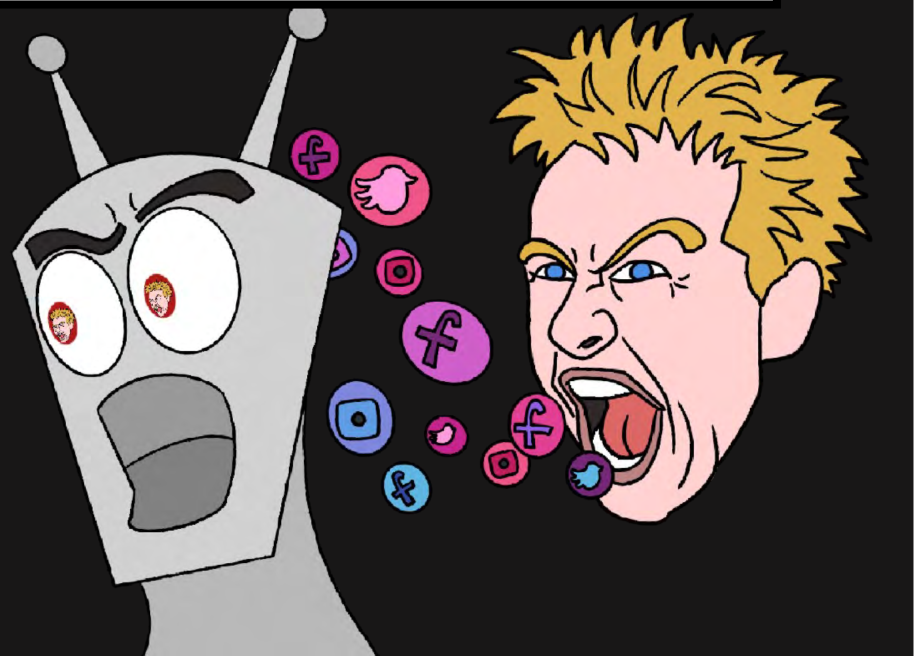
PRE-EXISTING BIAS EXISTS INDEPENDENT OF THE ALGORITHM AND HAS ITS ORIGINS IN SOCIETY.

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.

THESE BIASES EXIST IN SOCIETY AND COME 'PRE-BAKED' INTO THE ALGORITHM,

FROM THE UNDERLYING DISCRIMINATORY SYSTEM THAT THE DATA WAS COLLECTED FROM -

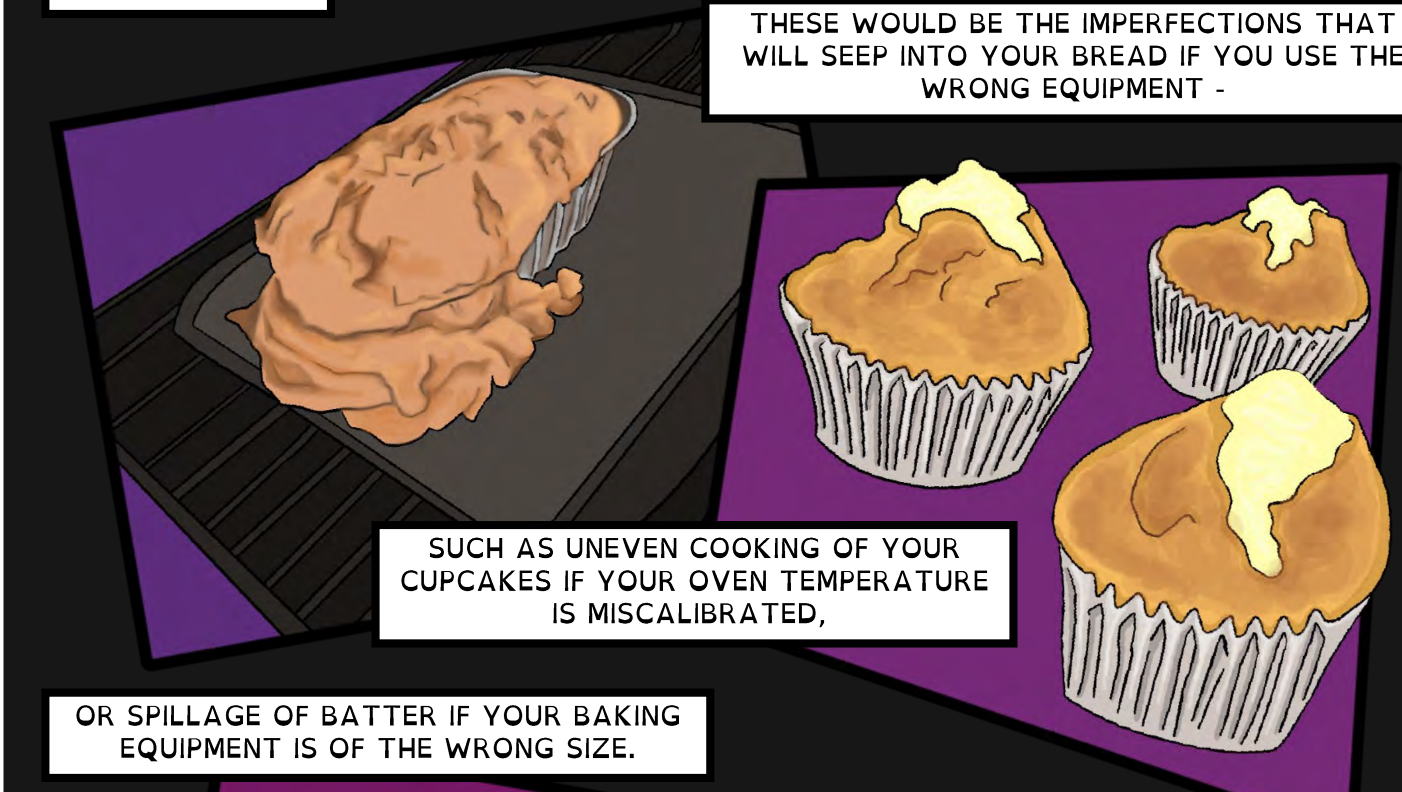


SUCH AS THE GENDER AND RACIAL STEREOTYPES THAT LANGUAGE MODELS PICK UP WHEN TRAINED ON DATA FROM SOCIAL MEDIA.

TECHNICAL BIAS

TECHNICAL BIAS IS INTRODUCED BY THE SYSTEM ITSELF - BECAUSE OF THE WAY IT IS DESIGNED OR OPERATES.

THESE WOULD BE THE IMPERFECTIONS THAT WILL SEEP INTO YOUR BREAD IF YOU USE THE WRONG EQUIPMENT -

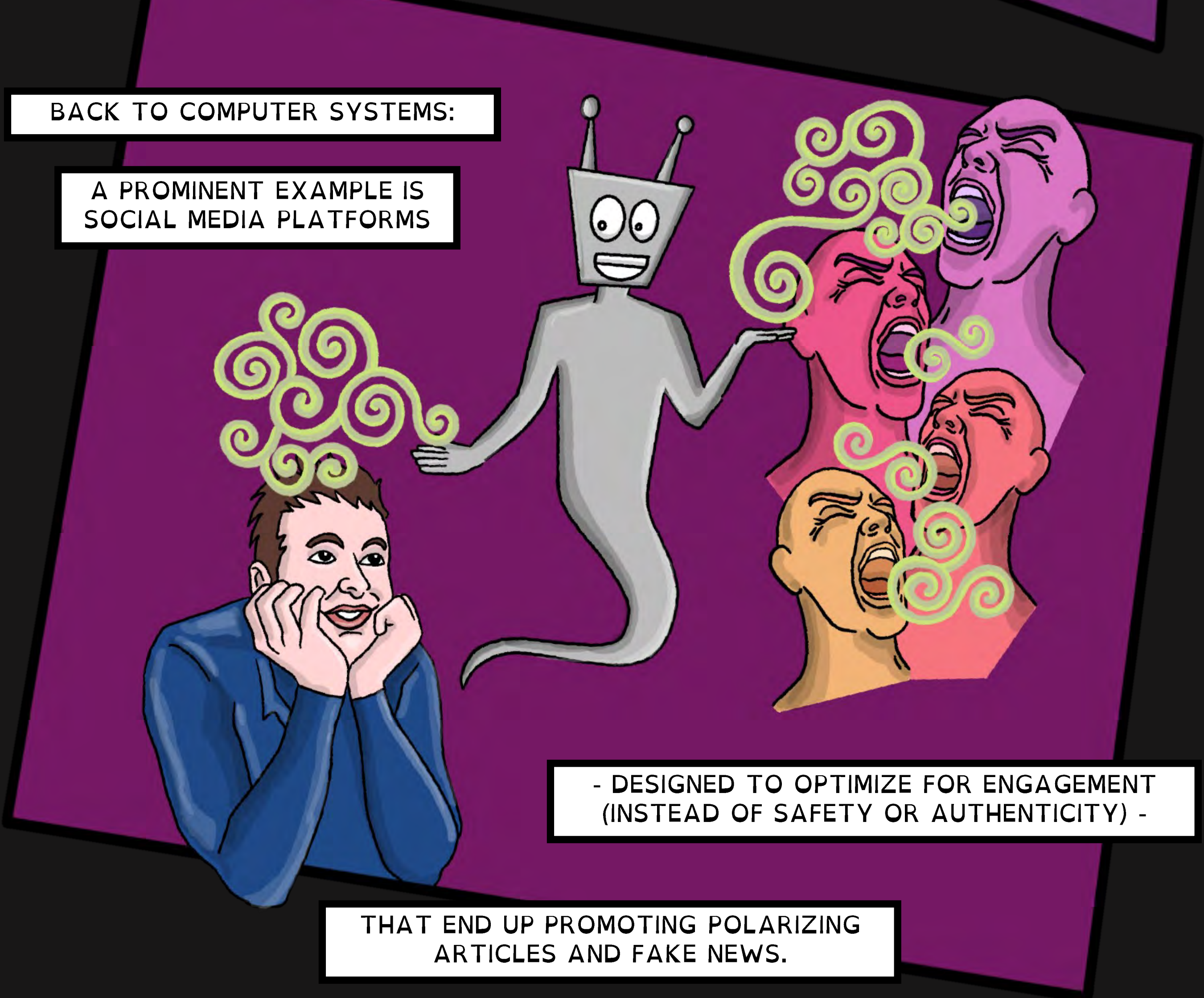


SUCH AS UNEVEN COOKING OF YOUR CUPCAKES IF YOUR OVEN TEMPERATURE IS MISCALIBRATED,

OR SPILLAGE OF BATTER IF YOUR BAKING EQUIPMENT IS OF THE WRONG SIZE.

BACK TO COMPUTER SYSTEMS:

A PROMINENT EXAMPLE IS SOCIAL MEDIA PLATFORMS



- DESIGNED TO OPTIMIZE FOR ENGAGEMENT (INSTEAD OF SAFETY OR AUTHENTICITY) -

THAT END UP PROMOTING POLARIZING ARTICLES AND FAKE NEWS.

EMERGENT BIAS (DUE TO DECISIONS)

EMERGENT BIAS ARISES OVER TIME, BECAUSE THE DECISIONS MADE WITH THE HELP OF THE SYSTEM CHANGE THE WORLD,

WHICH IN TURN IMPACTS THE OPERATION OF THE SYSTEM GOING FORWARD.

THINK ABOUT BEHAVIORAL CHANGES THAT WILL EMERGE AS A RESULT OF YOUR BAKING -

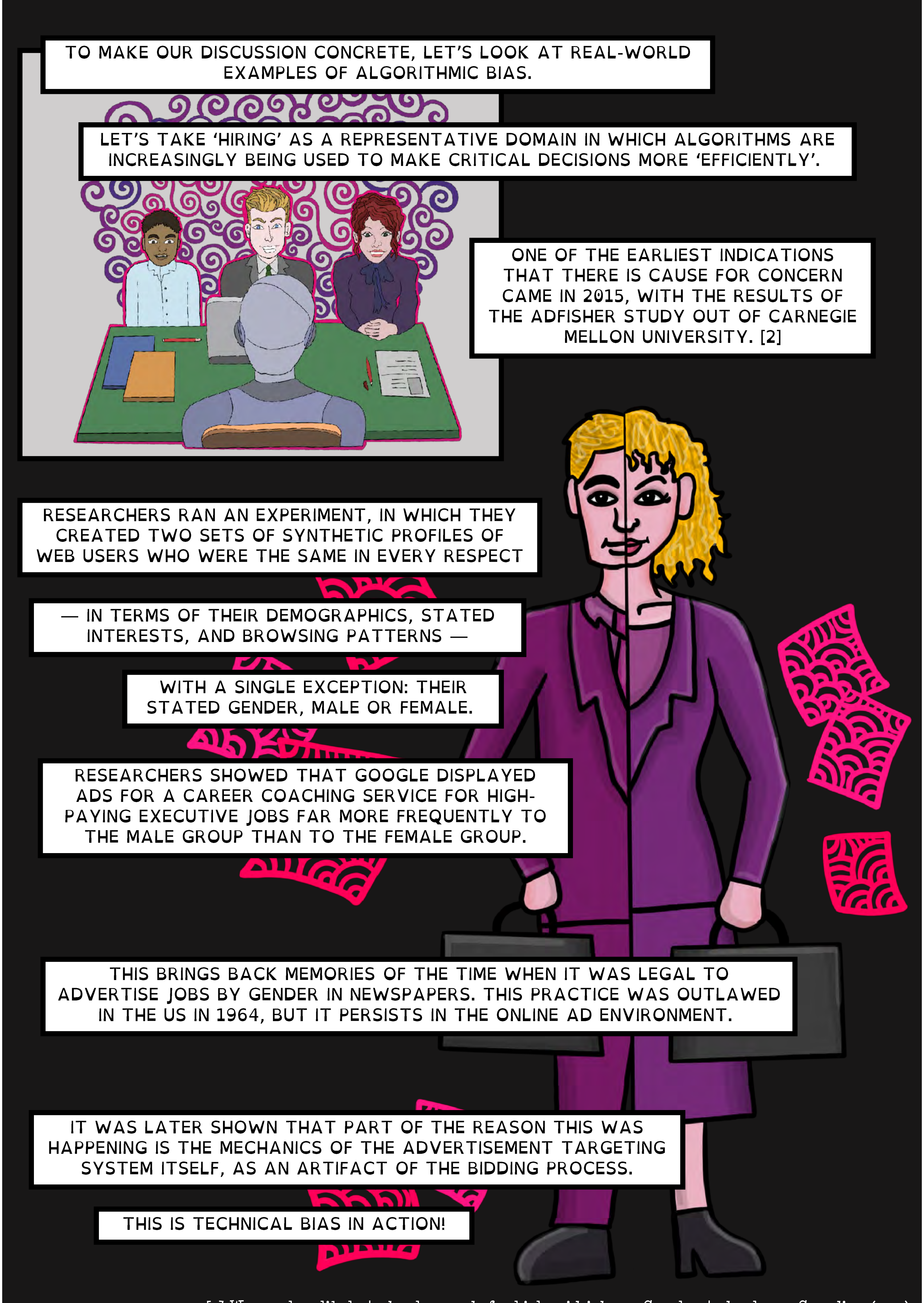
WHAT IF YOU BECOME SUCH A MAESTRO AT BAKING THAT YOU INADVERTENTLY MAKE BREAD A STEADY PART OF YOUR DIET!

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'.

IN THE SAME VEIN, THINK ABOUT HOW YOUR EXPOSURE TO NEWS - AND INFORMATION MORE BROADLY -

IS SHAPED BY ALGORITHMS THAT CURATE SOCIAL FEEDS WITH POPULAR AND 'TRENDING' POSTS.



TO MAKE OUR DISCUSSION CONCRETE, LET'S LOOK AT REAL-WORLD EXAMPLES OF ALGORITHMIC BIAS.

LET'S TAKE 'HIRING' AS A REPRESENTATIVE DOMAIN IN WHICH ALGORITHMS ARE INCREASINGLY BEING USED TO MAKE CRITICAL DECISIONS MORE 'EFFICIENTLY'.

ONE OF THE EARLIEST INDICATIONS THAT THERE IS CAUSE FOR CONCERN CAME IN 2015, WITH THE RESULTS OF THE ADFISHER STUDY OUT OF CARNEGIE MELLON UNIVERSITY. [2]

RESEARCHERS RAN AN EXPERIMENT, IN WHICH THEY CREATED TWO SETS OF SYNTHETIC PROFILES OF WEB USERS WHO WERE THE SAME IN EVERY RESPECT

— IN TERMS OF THEIR DEMOGRAPHICS, STATED INTERESTS, AND BROWSING PATTERNS —

WITH A SINGLE EXCEPTION: THEIR STATED GENDER, MALE OR FEMALE.

RESEARCHERS SHOWED THAT GOOGLE DISPLAYED ADS FOR A CAREER COACHING SERVICE FOR HIGH-PAYING EXECUTIVE JOBS FAR MORE FREQUENTLY TO THE MALE GROUP THAN TO THE FEMALE GROUP.

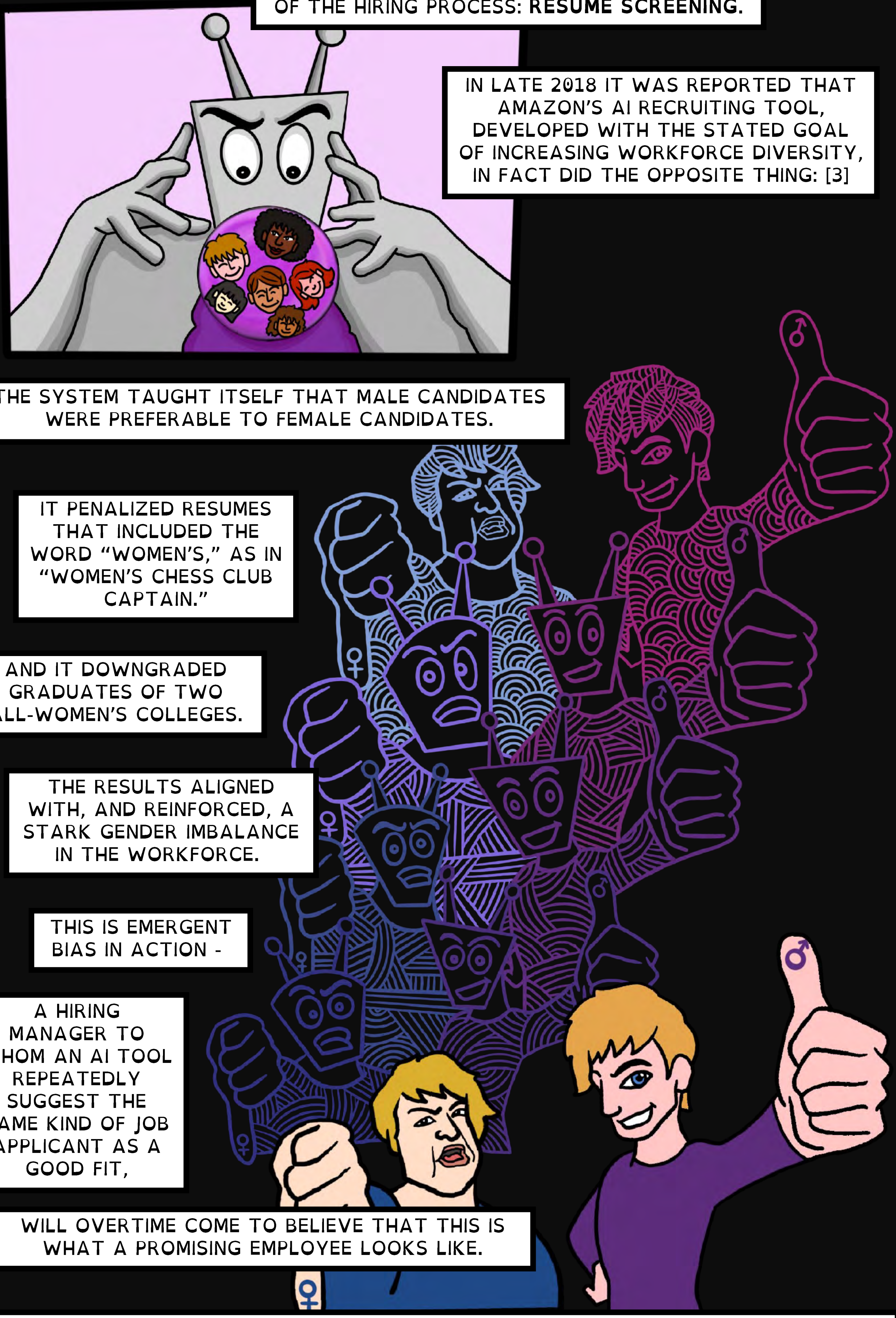
THIS BRINGS BACK MEMORIES OF THE TIME WHEN IT WAS LEGAL TO ADVERTISE JOBS BY GENDER IN NEWSPAPERS. THIS PRACTICE WAS OUTLAWED IN THE US IN 1964, BUT IT PERSISTS IN THE ONLINE AD ENVIRONMENT.

IT WAS LATER SHOWN THAT PART OF THE REASON THIS WAS HAPPENING IS THE MECHANICS OF THE ADVERTISEMENT TARGETING SYSTEM ITSELF, AS AN ARTIFACT OF THE BIDDING PROCESS.

THIS IS TECHNICAL BIAS IN ACTION!

[2] Women less likely to be shown ads for high-paid jobs on Google, study shows. Guardian (2015)

LET US MOVE FORWARD TO THE NEXT STAGE OF THE HIRING PROCESS: **RESUME SCREENING.**



IN LATE 2018 IT WAS REPORTED THAT AMAZON'S AI RECRUITING TOOL, DEVELOPED WITH THE STATED GOAL OF INCREASING WORKFORCE DIVERSITY, IN FACT DID THE OPPOSITE THING: [3]

THE SYSTEM TAUGHT ITSELF THAT MALE CANDIDATES WERE PREFERABLE TO FEMALE CANDIDATES.

IT PENALIZED RESUMES THAT INCLUDED THE WORD "WOMEN'S," AS IN "WOMEN'S CHESS CLUB CAPTAIN."

AND IT DOWNGRADED GRADUATES OF TWO ALL-WOMEN'S COLLEGES.

THE RESULTS ALIGNED WITH, AND REINFORCED, A STARK GENDER IMBALANCE IN THE WORKFORCE.

THIS IS EMERGENT BIAS IN ACTION -

A HIRING MANAGER TO WHOM AN AI TOOL REPEATEDLY SUGGEST THE SAME KIND OF JOB APPLICANT AS A GOOD FIT,

WILL OVERTIME COME TO BELIEVE THAT THIS IS WHAT A PROMISING EMPLOYEE LOOKS LIKE.

WE ARE ALSO SEEING PRE-EXISTING BIAS IN THIS EXAMPLE: THE AI TOOL WAS TRAINED ON HISTORICAL DATA ABOUT PAST EMPLOYEES, WHO WERE PREDOMINANTLY MALE

[3] Amazon scraps secret AI recruiting tool that showed bias against women. Reuters (2018)

HERE'S ANOTHER EXAMPLE, LATER YET IN THE HIRING PROCESS, PERHAPS DURING A POST-INTERVIEW BACKGROUND CHECK BY A POTENTIAL EMPLOYER -

LATANYA SWEENEY, A COMPUTER SCIENCE PROFESSOR ON THE FACULTY AT HARVARD,

SHOWED THAT GOOGLING FOR AFRICAN-AMERICAN SOUNDING NAMES IS MORE LIKELY TO TRIGGER ADS SUGGESTIVE OF A CRIMINAL RECORD THAN GOOGLING FOR WHITE-SOUNDING NAMES,

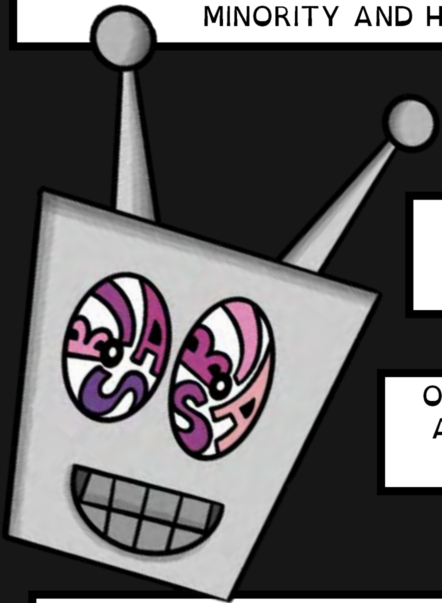
EVEN CONTROLLING FOR WHETHER AN INDIVIDUAL IN FACT HAS A CRIMINAL RECORD! [4]

THIS IS PRE-EXISTING BIAS AT PLAY -

MANIFESTING LONG-STANDING RACIAL PREJUDICES OF SOCIETY.

[4] Racism is Poisoning Online Ad Delivery, Says Harvard Professor. MIT Technology Review (2013)

THE CASES PRESENTED HERE HAVE ONE THING IN COMMON: THEY SHOW THAT AI CAN REINFORCE AND EXACERBATE UNLAWFUL DISCRIMINATION AGAINST MINORITY AND HISTORICALLY DISADVANTAGED GROUPS.



OFTEN THIS IS CALLED OUT AS “BIAS IN AI”.

SO, WHY ARE SOPHISTICATED SYSTEMS THAT AIM TO MAKE HIRING MORE EFFICIENT FAILING AT THIS, AND ARGUABLY MAKING THINGS WORSE?

OF COURSE, THE ISSUES OF BIAS IN EMPLOYMENT ARE NOT NEW. THEY EXHIBITED THEMSELVES IN THE ANALOG ERA AS WELL.

FOR EXAMPLE, IN THEIR WELL-KNOWN 2004 STUDY, MARIANNE BERTRAND AND SENDHIL MULLAINATHAN SENT FICTITIOUS RESUMES TO HELP-WANTED ADS IN BOSTON AND CHICAGO NEWSPAPERS. [5]

TO MANIPULATE PERCEIVED RACE, THEY RANDOMLY ASSIGNED AFRICAN-AMERICAN- OR WHITE-SOUNDING NAMES TO RESUMES.

WHITE NAMES RECEIVE 50 PERCENT MORE CALLBACKS FOR INTERVIEWS.

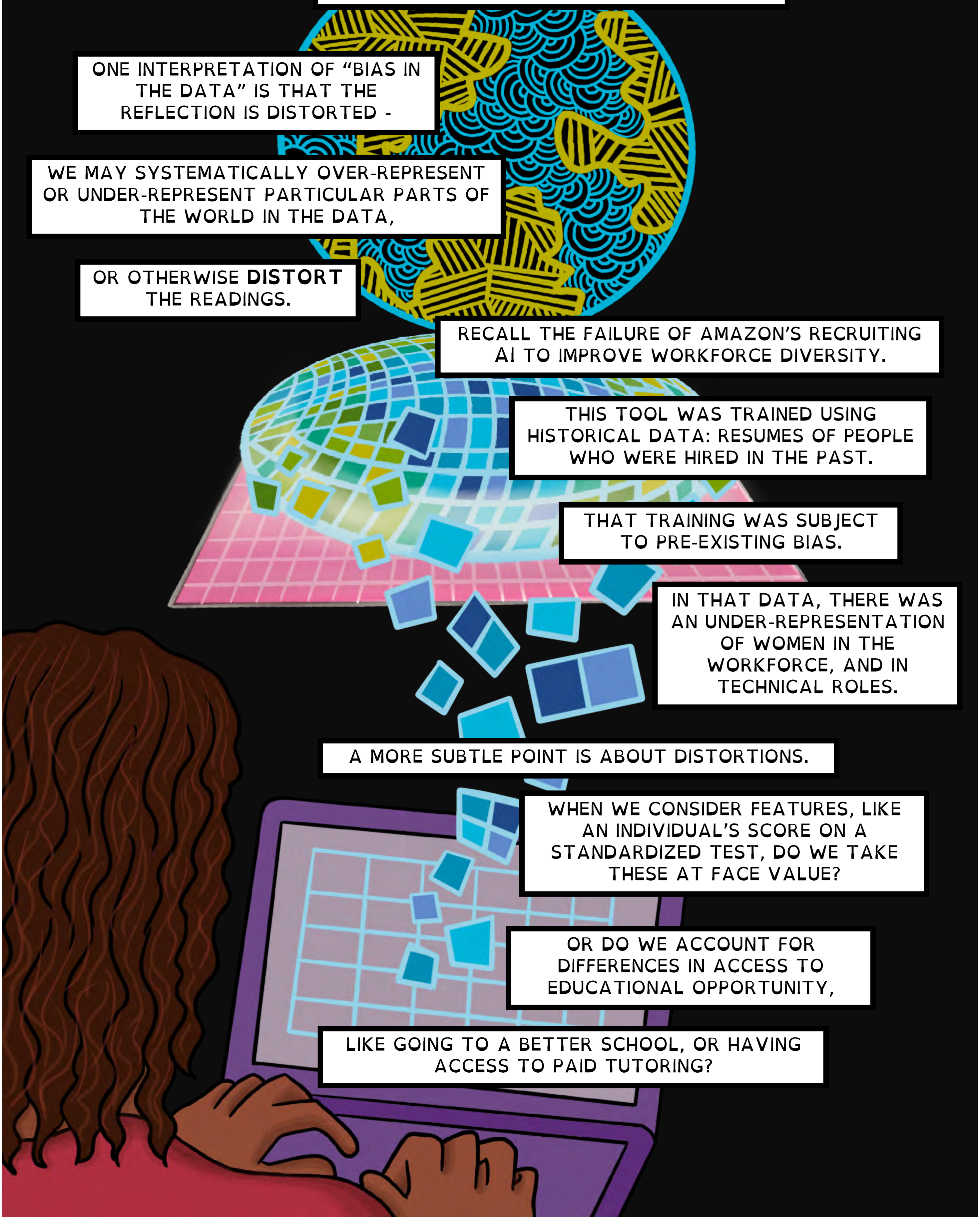
THIS CASE SHOWS THAT BIAS CAN BE DUE TO HUMAN DECISIONS.

[5] Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination.
Marianne Bertrand & Sendhil Mullainathan (2003)

LET'S REVISIT PRE-EXISTING BIAS THAT OFTEN EXHIBITS ITSELF IN THE DATA.

DATA IS AN IMAGE OF THE **WORLD**, ITS **MIRROR REFLECTION**.

WHEN WE THINK ABOUT BIAS IN THE DATA, WE INTERROGATE THIS REFLECTION.



ONE INTERPRETATION OF "BIAS IN THE DATA" IS THAT THE REFLECTION IS DISTORTED -

WE MAY SYSTEMATICALLY OVER-REPRESENT OR UNDER-REPRESENT PARTICULAR PARTS OF THE WORLD IN THE DATA,

OR OTHERWISE **DISTORT** THE READINGS.

RECALL THE FAILURE OF AMAZON'S RECRUITING AI TO IMPROVE WORKFORCE DIVERSITY.

THIS TOOL WAS TRAINED USING HISTORICAL DATA: RESUMES OF PEOPLE WHO WERE HIRED IN THE PAST.

THAT TRAINING WAS SUBJECT TO PRE-EXISTING BIAS.

IN THAT DATA, THERE WAS AN UNDER-REPRESENTATION OF WOMEN IN THE WORKFORCE, AND IN TECHNICAL ROLES.

A MORE SUBTLE POINT IS ABOUT DISTORTIONS.

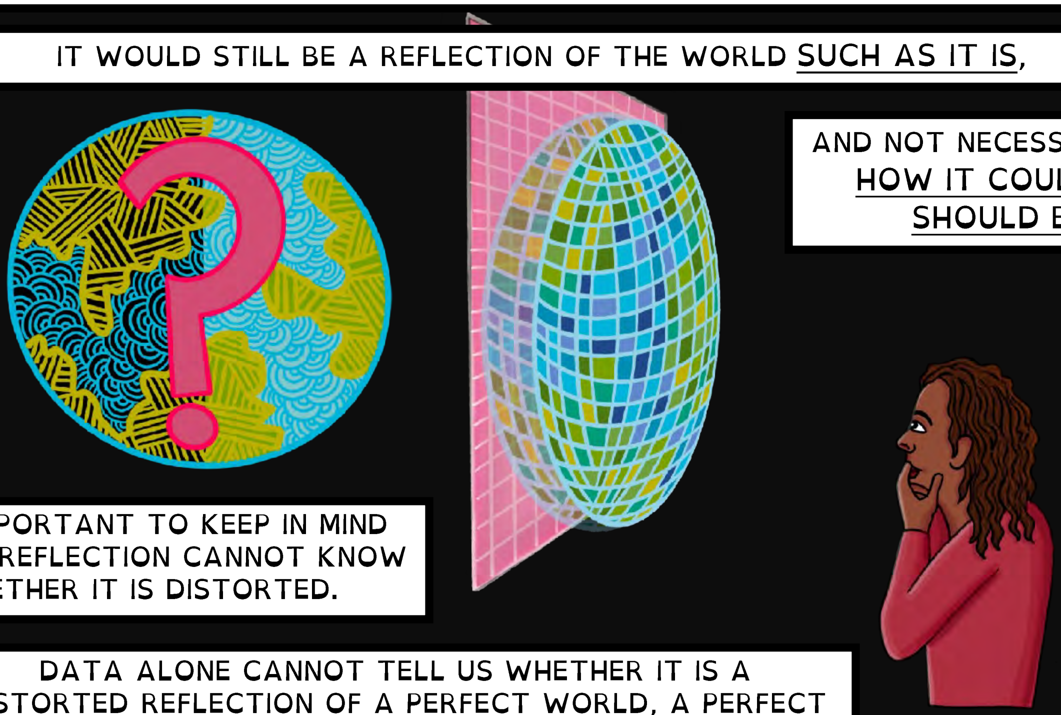
WHEN WE CONSIDER FEATURES, LIKE AN INDIVIDUAL'S SCORE ON A STANDARDIZED TEST, DO WE TAKE THESE AT FACE VALUE?

OR DO WE ACCOUNT FOR DIFFERENCES IN ACCESS TO EDUCATIONAL OPPORTUNITY,

LIKE GOING TO A BETTER SCHOOL, OR HAVING ACCESS TO PAID TUTORING?

ANOTHER INTERPRETATION OF "BIAS IN THE DATA" IS THAT EVEN IF WE WERE ABLE TO REFLECT THE WORLD PERFECTLY IN THE DATA,

IT WOULD STILL BE A REFLECTION OF THE WORLD SUCH AS IT IS,



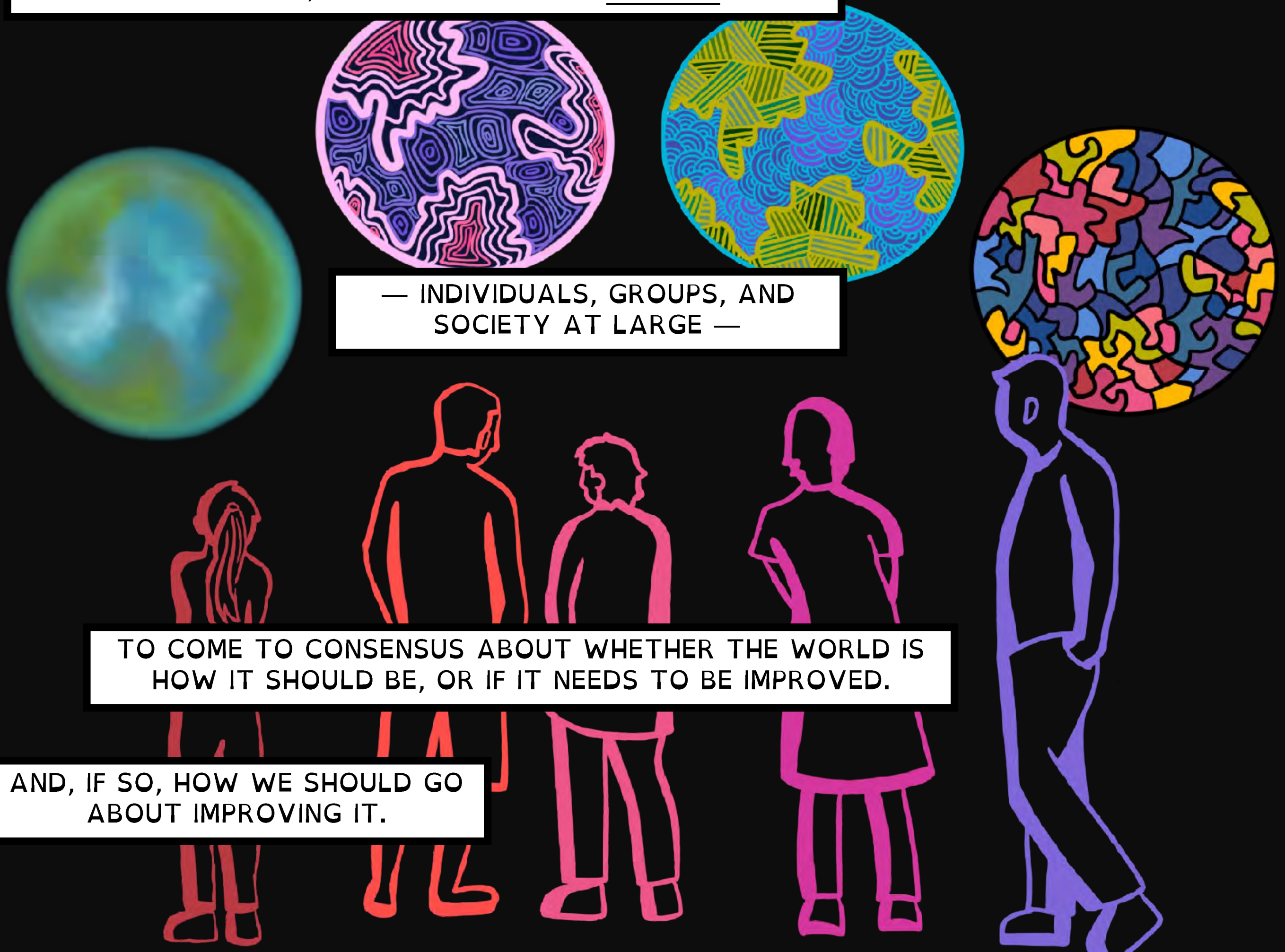
AND NOT NECESSARILY OF HOW IT COULD OR SHOULD BE.

IT IS IMPORTANT TO KEEP IN MIND THAT A REFLECTION CANNOT KNOW WHETHER IT IS DISTORTED.

DATA ALONE CANNOT TELL US WHETHER IT IS A DISTORTED REFLECTION OF A PERFECT WORLD, A PERFECT REFLECTION OF A DISTORTED WORLD,

OR IF THESE DISTORTIONS COMPOUND.

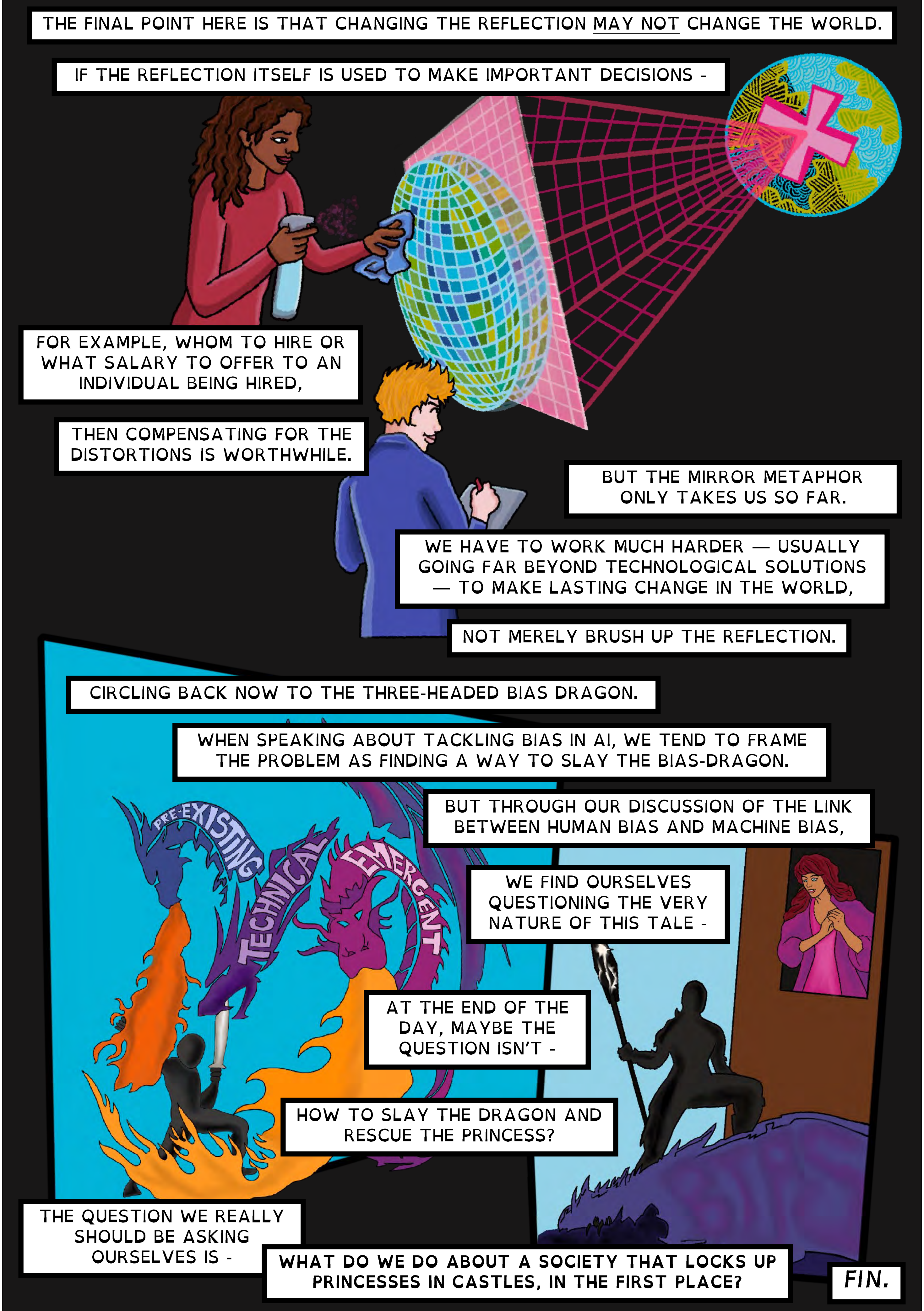
THE SECOND POINT IS THAT IT IS NOT UP TO DATA OR ALGORITHMS, BUT RATHER UP TO **PEOPLE**



— INDIVIDUALS, GROUPS, AND SOCIETY AT LARGE —

TO COME TO CONSENSUS ABOUT WHETHER THE WORLD IS HOW IT SHOULD BE, OR IF IT NEEDS TO BE IMPROVED.

AND, IF SO, HOW WE SHOULD GO ABOUT IMPROVING IT.


THE FINAL POINT HERE IS THAT CHANGING THE REFLECTION MAY NOT CHANGE THE WORLD.
IF THE REFLECTION ITSELF IS USED TO MAKE IMPORTANT DECISIONS -
FOR EXAMPLE, WHOM TO HIRE OR WHAT SALARY TO OFFER TO AN INDIVIDUAL BEING HIRED,
THEN COMPENSATING FOR THE DISTORTIONS IS WORTHWHILE.
BUT THE MIRROR METAPHOR ONLY TAKES US SO FAR.
WE HAVE TO WORK MUCH HARDER — USUALLY GOING FAR BEYOND TECHNOLOGICAL SOLUTIONS — TO MAKE LASTING CHANGE IN THE WORLD,
NOT MERELY BRUSH UP THE REFLECTION.
CIRCLING BACK NOW TO THE THREE-HEADED BIAS DRAGON.
WHEN SPEAKING ABOUT TACKLING BIAS IN AI, WE TEND TO FRAME THE PROBLEM AS FINDING A WAY TO SLAY THE BIAS-DRAGON.
PRE-EXISTING
TECHNICAL
EMERGENT
BUT THROUGH OUR DISCUSSION OF THE LINK BETWEEN HUMAN BIAS AND MACHINE BIAS,
WE FIND OURSELVES QUESTIONING THE VERY NATURE OF THIS TALE -
AT THE END OF THE DAY, MAYBE THE QUESTION ISN'T -
HOW TO SLAY THE DRAGON AND RESCUE THE PRINCESS?
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.

Bias in Computer Systems

BATYA FRIEDMAN
Colby College and The Mina Institute
and
HELEN NISSENBAUM
Princeton University

From an analysis of actual cases, three categories of bias in computer systems have been developed: preexisting, technical, and emergent. Preexisting bias has its roots in social institutions, practices, and attitudes. Technical bias arises from technical constraints or considerations. Emergent bias arises in a context of use. Although others have pointed to bias in particular computer systems and have noted the general problem, we know of no comparable work that examines this phenomenon comprehensively and which offers a framework for understanding and remedying it. We conclude by suggesting that freedom from bias should be counted among the select set of criteria—including reliability, accuracy, and efficiency—according to which the quality of systems in use in society should be judged.

Categories and Subject Descriptors: D.2.0 [**Software**]: Software Engineering; H.1.2 [**Information Systems**]: User/Machine Systems; K.4.0 [**Computers and Society**]: General

General Terms: Design, Human Factors

Additional Key Words and Phrases: Bias, computer ethics, computers and society, design methods, ethics, human values, standards, social computing, social impact, system design, universal design, values

INTRODUCTION

To introduce what bias in computer systems might look like, consider the case of computerized airline reservation systems, which are used widely by travel agents to identify and reserve airline flights for their customers. These reservation systems seem straightforward. When a travel agent types in a customer's travel requirements, the reservation system searches

This research was funded in part by the Clare Boothe Luce Foundation.
Earlier aspects of this work were presented at the 4S/EASST Conference, Goteborg, Sweden, August 1992, and at InterCHI '93, Amsterdam, April 1993. An earlier version of this article appeared as Tech. Rep. CSLI-94-188, CSLI, Stanford University.

Authors' addresses: B. Friedman, Department of Mathematics and Computer Science, Colby College, Waterville, ME 04901; email: b_friedm@colby.edu; H. Nissenbaum, University Center for Human Values, Marx Hall, Princeton University, Princeton, NJ 08544; email: helen@phoenix.princeton.edu.

a database of flights and retrieves all reasonable flight options that meet or come close to the customer's requirements. These options then are ranked according to various criteria, giving priority to nonstop flights, more direct routes, and minimal total travel time. The ranked flight options are displayed for the travel agent. In the 1980s, however, most of the airlines brought before the Antitrust Division of the United States Justice Department allegations of anticompetitive practices by American and United Airlines whose reservation systems—Sabre and Apollo, respectively—dominated the field. It was claimed, among other things, that the two reservations systems are biased [Schrifin 1985].

One source of this alleged bias lies in Sabre's and Apollo's algorithms for controlling search and display functions. In the algorithms, preference is given to "on-line" flights, that is, flights with all segments on a single carrier. Imagine, then, a traveler who originates in Phoenix and flies the first segment of a round-trip overseas journey to London on American Airlines, changing planes in New York. All other things being equal, the British Airlines' flight from New York to London would be ranked lower than the American Airlines' flight from New York to London even though in both cases a traveler is similarly inconvenienced by changing planes and checking through customs. Thus, the computer systems systematically downgrade and, hence, are biased against international carriers who fly few, if any, internal U.S. flights, and against internal carriers who do not fly international flights [Fotos 1988; Ott 1988].

Critics also have been concerned with two other problems. One is that the interface design compounds the bias in the reservation systems. Lists of ranked flight options are displayed screen by screen. Each screen displays only two to five options. The advantage to a carrier of having its flights shown on the first screen is enormous since 90% of the tickets booked by travel agents are booked by the first screen display [Taib 1990]. Even if the biased algorithm and interface give only a small percent advantage overall to one airline, it can make the difference to its competitors between survival and bankruptcy. A second problem arises from the travelers' perspective. When travelers contract with an independent third party—a travel agent—to determine travel plans, travelers have good reason to assume they are being informed accurately of their travel options; in many situations, that does not happen.

As Sabre and Apollo illustrate, biases in computer systems can be difficult to identify let alone remedy because of the way the technology engages and extenuates them. Computer systems, for instance, are comparatively inexpensive to disseminate, and thus, once developed, a biased system has the potential for widespread impact. If the system becomes a standard in the field, the bias becomes pervasive. If the system is complex, and most are, biases can remain hidden in the code, difficult to pinpoint or explicate, and not necessarily disclosed to users or their clients. Furthermore, unlike in our dealings with biased individuals with whom a potential victim can negotiate, biased systems offer no equivalent means for appeal.

Although others have pointed to bias in particular computer systems and have noted the general problem [Johnson and Mulvey 1993; Moor 1985], we know of no comparable work that focuses exclusively on this phenomenon and examines it comprehensively.

In this article, we provide a framework for understanding bias in computer systems. From an analysis of actual computer systems, we have developed three categories: preexisting bias, technical bias, and emergent bias. Preexisting bias has its roots in social institutions, practices, and attitudes. Technical bias arises from technical constraints or considerations. Emergent bias arises in a context of use. We begin by defining bias and explicating each category and then move to case studies. We conclude with remarks about how bias in computer systems can be remedied.

1. WHAT IS A BIASED COMPUTER SYSTEM?

In its most general sense, the term bias means simply "slant." Given this undifferentiated usage, at times the term is applied with relatively neutral content. A grocery shopper, for example, can be "biased" by not buying damaged fruit. At other times, the term bias is applied with significant moral meaning. An employer, for example, can be "biased" by refusing to hire minorities. In this article we focus on instances of the latter, for if one wants to develop criteria for judging the quality of systems in use—which we do—then criteria must be delineated in ways that speak robustly yet precisely to relevant social matters. Focusing on bias of moral import does just that.

Accordingly, we use the term bias to refer to computer systems that *systematically* and *unfairly discriminate* against certain individuals or groups of individuals in favor of others. A system discriminates unfairly if it denies an opportunity or a good or if it assigns an undesirable outcome to an individual or group of individuals on grounds that are unreasonable or inappropriate. Consider, for example, an automated credit advisor that assists in the decision of whether or not to extend credit to a particular applicant. If the advisor denies credit to individuals with consistently poor payment records we do not judge the system to be biased because it is reasonable and appropriate for a credit company to want to avoid extending credit privileges to people who consistently do not pay their bills. In contrast, a credit advisor that systematically assigns poor credit ratings to individuals with ethnic surnames discriminates on grounds that are not relevant to credit assessments and, hence, discriminates unfairly.

Two points follow. First, unfair discrimination alone does not give rise to bias unless it occurs systematically. Consider again the automated credit advisor. Imagine a random glitch in the system which changes in an isolated case information in a copy of the credit record for an applicant who happens to have an ethnic surname. The change in information causes a downgrading of this applicant's rating. While this applicant experiences unfair discrimination resulting from this random glitch, the applicant could have been anybody. In a repeat incident, the same applicant or others with similar ethnicity would not be in a special position to be singled out. Thus, while the system is prone to random error, it is not biased.

Second, systematic discrimination does not establish bias unless it is joined with an unfair outcome. A case in point is the Persian Gulf War, where United States Patriot missiles were used to detect and intercept Iraqi Scud missiles. At least one software error identified during the war contributed to systematically poor performance by the Patriots [Gao 1992]. Calculations used to predict the location of a Scud depended in complex ways on the Patriots' internal clock. The longer the Patriot's continuous running time, the greater the imprecision in the calculation. The deaths of at least 28 Americans in Dhahran can be traced to this software error, which systematically degraded the accuracy of Patriot missiles. While we are not minimizing the serious consequence of this systematic computer error, it falls outside of our analysis because it does not involve unfairness.

2. FRAMEWORK FOR ANALYZING BIAS IN COMPUTER SYSTEMS

We derived our framework by examining actual computer systems for bias. Instances of bias were identified and characterized according to their source, and then the characterizations were generalized to more abstract categories. These categories were further refined by their application to other instances of bias in the same or additional computer systems. In most cases, our knowledge of particular systems came from the published literature. In total, we examined 17 computer systems from diverse fields including banking, commerce, computer science, education, medicine, and law.

The framework that emerged from this methodology is comprised of three overarching categories—preexisting bias, technical bias, and emergent bias. Table I contains a detailed description of each category. In more general terms, they can be described as follows.

2.1 Preexisting Bias

Preexisting bias has its roots in social institutions, practices, and attitudes. When computer systems embody biases that exist independently, and usually prior to the creation of the system, then we say that the system embodies preexisting bias. Preexisting biases may originate in society at large, in subcultures, and in formal or informal, private or public organizations and institutions. They can also reflect the personal biases of individuals who have significant input into the design of the system, such as the client or system designer. This type of bias can enter a system either through the explicit and conscious efforts of individuals or institutions, or implicitly and unconsciously, even in spite of the best of intentions. For example, imagine an expert system that advises on loan applications. In determining an applicant's credit risk, the automated loan advisor negatively weights applicants who live in "undesirable" locations, such as low-income or high-crime neighborhoods, as indicated by their home addresses (a practice referred to as "red-lining"). To the extent the program

Table I. Categories of Bias in Computer System Design

These categories describe ways in which bias can arise in the design of computer systems. The illustrative examples portray plausible cases of bias.

1. Preexisting Bias

Preexisting bias has its roots in social institutions, practices, and attitudes.

When computer systems embody biases that exist independently, and usually prior to the creation of the system, then the system exemplifies preexisting bias. Preexisting bias can enter a system either through the explicit and conscious efforts of individuals or institutions, or implicitly and unconsciously, even in spite of the best of intentions.

1.1. Individual

Bias that originates from individuals who have significant input into the design of the system, such as the client commissioning the design or the system designer (e.g., a client embeds personal racial biases into the specifications for loan approval software).

1.2 Societal

Bias that originates from society at large, such as from organizations (e.g., industry), institutions (e.g., legal systems), or culture at large (e.g., gender biases present in the larger society that lead to the development of educational software that overall appeals more to boys than girls).

2. Technical Bias

Technical bias arises from technical constraints or technical considerations.

2.1 Computer Tools

Bias that originates from a limitation of the computer technology including hardware, software, and peripherals (e.g., in a database for matching organ donors with potential transplant recipients certain individuals retrieved and displayed on initial screens are favored systematically for a match over individuals displayed on later screens).

2.2 Decontextualized Algorithms

Bias that originates from the use of an algorithm that fails to treat all groups fairly under all significant conditions (e.g., a scheduling algorithm that schedules airplanes for take-off relies on the alphabetic listing of the airlines to rank order flights ready within a given period of time).

2.3 Random Number Generation

Bias that originates from imperfections in pseudorandom number generation or in the misuse of pseudorandom numbers (e.g., an imperfection in a random-number generator used to select recipients for a scarce drug leads systematically to favoring individuals toward the end of the database).

2.4 Formalization of Human Constructs

Bias that originates from attempts to make human constructs such as discourse, judgments, or intuitions amenable to computers: when we quantify the qualitative, discretize the continuous, or formalize the nonformal (e.g., a legal expert system advises defendants on whether or not to plea bargain by assuming that law can be spelled out in an unambiguous manner that is not subject to human and humane interpretations in context).

Table I. *Continued*

These categories describe ways in which bias can arise in the design of computer systems. The illustrative examples portray plausible cases of bias.

3. Emergent Bias

Emergent bias arises in a context of use with real users. This bias typically emerges some time after a design is completed, as a result of changing societal knowledge, population, or cultural values. User interfaces are likely to be particularly prone to emergent bias because interfaces by design seek to reflect the capacities, character, and habits of prospective users. Thus, a shift in context of use may well create difficulties for a new set of users.

3.1 New Societal Knowledge

Bias that originates from the emergence of new knowledge in society that cannot be or is not incorporated into the system design (e.g., a medical expert system for AIDS patients has no mechanism for incorporating cutting-edge medical discoveries that affect how individuals with certain symptoms should be treated).

3.2 Mismatch between Users and System Design

Bias that originates when the population using the system differs on some significant dimension from the population assumed as users in the design.

3.2.1 Different Expertise

Bias that originates when the system is used by a population with a different knowledge base from that assumed in the design (e.g., an ATM with an interface that makes extensive use of written instructions—"place the card, magnetic tape side down, in the slot to your left"—is installed in a neighborhood with primarily a nonliterate population).

3.2.2 Different Values

Bias that originates when the system is used by a population with different values than those assumed in the design (e.g., educational software to teach mathematics concepts is embedded in a game situation that rewards individualistic and competitive strategies, but is used by students with a cultural background that largely eschews competition and instead promotes cooperative endeavors).

embeds the biases of clients or designers who seek to avoid certain applicants on the basis of group stereotypes, the automated loan advisor's bias is preexisting.

2.2 Technical Bias

In contrast to preexisting bias, technical bias arises from the resolution of issues in the technical design. Sources of technical bias can be found in several aspects of the design process, including limitations of computer tools such as hardware, software, and peripherals; the process of ascribing social meaning to algorithms developed out of context; imperfections in pseudorandom number generation; and the attempt to make human constructs amenable to computers, when we quantify the qualitative, discretize the continuous, or formalize the nonformal. As an illustration, consider again the case of Sabre and Apollo described above. A technical constraint imposed by the size of the monitor screen forces a piecemeal presentation of flight options and, thus, makes the algorithm chosen to rank flight options critically important. Whatever ranking algorithm is used, if it systematically places certain airlines' flights on initial screens and other airlines' flights on later screens, the system will exhibit technical bias.

2.3 Emergent Bias

While it is almost always possible to identify preexisting bias and technical bias in a system design at the time of creation or implementation, emergent bias arises only in a context of use. This bias typically emerges some time after a design is completed, as a result of changing societal knowledge, population, or cultural values. Using the example of an automated airline reservation system, envision a hypothetical system designed for a group of airlines all of whom serve national routes. Consider what might occur if that system was extended to include international airlines. A flight-ranking algorithm that favors on-line flights when applied in the original context with national airlines leads to no systematic unfairness. However, in the new context with international airlines, the automated system would place these airlines at a disadvantage and, thus, comprise a case of emergent bias. User interfaces are likely to be particularly prone to emergent bias because interfaces by design seek to reflect the capacities, character, and habits of prospective users. Thus, a shift in context of use may well create difficulties for a new set of users.

3. APPLICATIONS OF THE FRAMEWORK

We now analyze actual computer systems in terms of the framework introduced above. It should be understood that the systems we analyze are by and large good ones, and our intention is not to undermine their integrity. Rather, our intention is to develop the framework, show how it can identify and clarify our understanding of bias in computer systems, and establish its robustness through real-world cases.

3.1 The National Resident Match Program (NRMP)

The NRMP implements a centralized method for assigning medical school graduates their first employment following graduation. The centralized method of assigning medical students to hospital programs arose in the 1950s in response to the chaotic job placement process and on-going failure of hospitals and students to arrive at optimal placements. During this early period the matching was carried out by a mechanical card-sorting process, but in 1974 electronic data processing was introduced to handle the entire matching process. (For a history of the NRMP, see Graettinger and Peranson [1981a].) After reviewing applications and interviewing students, hospital programs submit to the centralized program their ranked list of students. Students do the same for hospital programs. Hospitals and students are not permitted to make other arrangements with one another or to attempt to directly influence each others' rankings prior to the match.

Inherent Trade-Offs in the Fair Determination of Risk Scores

Jon Kleinberg [*] Sendhil Mullainathan [†] Manish Raghavan [‡]

Abstract

Recent discussion in the public sphere about algorithmic classification has involved tension between competing notions of what it means for a probabilistic classification to be fair to different groups. We formalize three fairness conditions that lie at the heart of these debates, and we prove that except in highly constrained special cases, there is no method that can satisfy these three conditions simultaneously. Moreover, even satisfying all three conditions approximately requires that the data lie in an approximate version of one of the constrained special cases identified by our theorem. These results suggest some of the ways in which key notions of fairness are incompatible with each other, and hence provide a framework for thinking about the trade-offs between them.

1 Introduction

There are many settings in which a sequence of people comes before a decision-maker, who must make a judgment about each based on some observable set of features. Across a range of applications, these judgments are being carried out by an increasingly wide spectrum of approaches ranging from human expertise to algorithmic and statistical frameworks, as well as various combinations of these approaches.

Along with these developments, a growing line of work has asked how we should reason about issues of bias and discrimination in settings where these algorithmic and statistical techniques, trained on large datasets of past instances, play a significant role in the outcome. Let us consider three examples where such issues arise, both to illustrate the range of relevant contexts, and to surface some of the challenges.

A set of example domains. First, at various points in the criminal justice system, including decisions about bail, sentencing, or parole, an officer of the court may use quantitative *risk tools* to assess a defendant's probability of recidivism — future arrest — based on their past history and other attributes. Several recent analyses have asked whether such tools are mitigating or exacerbating the sources of bias in the criminal justice system; in one widely-publicized report, Angwin et al. analyzed a commonly used statistical method for assigning risk scores in the criminal justice system — the COMPAS risk tool — and argued that it was biased against African-American defendants [2, 23]. One of their main contentions was that the tool's errors were asymmetric: African-American defendants were more likely to be incorrectly labeled as higher-risk than they actually were, while white defendants were more likely to be incorrectly labeled as lower-risk than they actually were. Subsequent analyses raised methodological objections to this report, and also observed that despite the COMPAS risk tool's errors, its estimates of the probability of recidivism are equally well calibrated to the true outcomes for both African-American and white defendants [1, 10, 13, 17].

[*] Cornell University

[†] Harvard University

[‡] Cornell University

Second, in a very different domain, researchers have begun to analyze the ways in which different genders and racial groups experience advertising and commercial content on the Internet differently [9, 26]. We could ask, for example: if a male user and female user are equally interested in a particular product, does it follow that they're equally likely to be shown an ad for it? Sometimes this concern may have broader implications, for example if women in aggregate are shown ads for lower-paying jobs. Other times, it may represent a clash with a user's leisure interests: if a female user interacting with an advertising platform is interested in an activity that tends to have a male-dominated viewership, like professional football, is the platform as likely to show her an ad for football as it is to show such an ad to an interested male user?

A third domain, again quite different from the previous two, is medical testing and diagnosis. Doctors making decisions about a patient's treatment may rely on tests providing probability estimates for different diseases and conditions. Here too we can ask whether such decision-making is being applied uniformly across different groups of patients [16, 27], and in particular how medical tests may play a differential role for conditions that vary widely in frequency between these groups.

Providing guarantees for decision procedures. One can raise analogous questions in many other domains of fundamental importance, including decisions about hiring, lending, or school admissions [24], but we will focus on the three examples above for the purposes of this discussion. In these three example domains, a few structural commonalities stand out. First, the algorithmic estimates are often being used as "input" to a larger framework that makes the overall decision — a risk score provided to a human expert in the legal and medical instances, and the output of a machine-learning algorithm provided to a larger advertising platform in the case of Internet ads. Second, the underlying task is generally about classifying whether people possess some relevant property: recidivism, a medical condition, or interest in a product. We will refer to people as being *positive instances* if they truly possess the property, and *negative instances* if they do not. Finally, the algorithmic estimates being provided for these questions are generally not pure yes-no decisions, but instead probability estimates about whether people constitute positive or negative instances.

Let us suppose that we are concerned about how our decision procedure might operate differentially between two groups of interest (such as African-American and white defendants, or male and female users of an advertising system). What sorts of guarantees should we ask for as protection against potential bias?

A first basic goal in this literature is that the probability estimates provided by the algorithm should be *well-calibrated*: if the algorithm identifies a set of people as having a probability z of constituting positive instances, then approximately a z fraction of this set should indeed be positive instances [8, 14]. Moreover, this condition should hold when applied separately in each group as well [13]. For example, if we are thinking in terms of potential differences between outcomes for men and women, this means requiring that a z fraction of men and a z fraction of women assigned a probability z should possess the property in question.

A second goal focuses on the people who constitute positive instances (even if the algorithm can only imperfectly recognize them): the average score received by people constituting positive instances should be the same in each group. We could think of this as *balance for the positive class*, since a violation of it would mean that people constituting positive instances in one group receive consistently lower probability estimates than people constituting positive instances in another group. In our initial criminal justice example, for instance, one of the concerns raised was that white defendants who went on to commit future crimes were assigned risk scores corresponding to lower probability estimates in aggregate; this is a violation of the condition here. There is a completely analogous property with respect to negative instances, which we could call *balance for the negative class*. These balance conditions can be viewed as generalizations of the notions that both groups should have equal false negative and false positive rates.

It is important to note that balance for the positive and negative classes, as defined here, is distinct in crucial ways from the requirement that the average probability estimate globally over *all* members of the two groups be equal. This latter global requirement is a version of *statistical parity* [12, 4, 21, 22]. In some cases statistical parity is a central goal (and in some it is legally mandated), but the examples considered so far suggest that classification and risk assessment are much broader activities where statistical parity is often neither feasible nor desirable. Balance for the positive and negative classes, however, is a goal that can be discussed independently of statistical parity, since these two balance conditions simply ask that once we condition on the "correct" answer for a person, the chance of making a mistake on them should not depend on which group they belong to.

The present work: Trade-offs among the guarantees. Despite their different formulations, the calibration condition and the balance conditions for the positive and negative classes intuitively all seem to be asking for variants of the same general goal — that our probability estimates should have the same effectiveness regardless of group membership. One might therefore hope that it would be feasible to achieve all of them simultaneously.

Our main result, however, is that these conditions are in general incompatible with each other; they can only be simultaneously satisfied in certain highly constrained cases. Moreover, this incompatibility applies to *approximate* versions of the conditions as well.

In the remainder of this section we formulate this main result precisely, as a theorem building on a model that makes the discussion thus far more concrete.

1.1 Formulating the Goal

Let's start with some basic definitions. As above, we have a collection of people each of whom constitutes either a positive instance or a negative instance of the classification problem. We'll say that the *positive class* consists of the people who constitute positive instances, and the negative class consists of the people who constitute negative instances. For example, for criminal defendants, the positive class could consist of those defendants who will be arrested again within some fixed time window, and the negative class could consist of those who will not. The positive and negative classes thus represent the "correct" answer to the classification problem; our decision procedure does not know them, but is trying to estimate them.

Feature vectors. Each person has an associated *feature vector* σ, representing the data that we know about them. Let p_σ denote the fraction of people with feature vector σ who belong to the positive class. Conceptually, we will picture that while there is variation within the set of people who have feature vector σ, this variation is invisible to whatever decision procedure we apply; all people with feature vector σ are indistinguishable to the procedure. Our model will assume that the value p_σ for each σ is known to the procedure.[1]

Groups. Each person also belongs to one of two *groups*, labeled 1 or 2, and we would like our decisions to be unbiased with respect to the members of these two groups.[2] In our examples, the two groups could correspond to different races or genders, or other cases where we want to look for the possibility of bias between them. The two groups have different distributions over feature vectors: a person of group t has a probability $a_{t\sigma}$ of exhibiting the feature vector σ. However, people of each group have the same probability

[1] Clearly the case in which the value of p_σ is unknown is an important version of the problem as well; however, since our main results establish strong limitations on what is achievable, these limitations are only stronger because they apply even to the case of known p_σ.

[2] We focus on the case of two groups for simplicity of exposition, but it is straightforward to extend all of our definitions to the case of more than two groups.

p_σ of belonging to the positive class provided their feature vector is σ. In this respect, σ contains all the relevant information available to us about the person's future behavior; once we know σ, we do not get any additional information from knowing their group as well.[3]

Risk Assignments. We say that an *instance* of our problem is specified by the parameters above: a feature vector and a group for each person, with a value p_σ for each feature vector, and distributions $\{a_{t\sigma}\}$ giving the frequency of the feature vectors in each group.

Informally, risk assessments are ways of dividing people up into sets based on their feature vectors σ (potentially using randomization), and then assigning each set a probability estimate that the people in this set belong to the positive class. Thus, we define a *risk assignment* to consist of a set of "bins" (the sets), where each bin is labeled with a *score* v_b that we intend to use as the probability for everyone assigned to bin b. We then create a rule for assigning people to bins based on their feature vector σ; we allow the rule to divide people with a fixed feature vector σ across multiple bins (reflecting the possible use of randomization). Thus, the rule is specified by values $X_{\sigma b}$: a fraction $X_{\sigma b}$ of all people with feature vector σ are assigned to bin b. Note that the rule does not have access to the group t of the person being considered, only their feature vector σ. (As we will see, this does not mean that the rule is incapable of exhibiting bias between the two groups.) In summary, a risk assignment is specified by a set of bins, a score for each bin, and values $X_{\sigma b}$ that define a mapping from people with feature vectors to bins.

Fairness Properties for Risk Assignments. Within the model, we now express the three conditions discussed at the outset, each reflecting a potentially different notion of what it means for the risk assignment to be "fair."

(A) *Calibration within groups* requires that for each group t, and each bin b with associated score v_b, the expected number of people from group t in b who belong to the positive class should be a v_b fraction of the expected number of people from group t assigned to b.

(B) *Balance for the negative class* requires that the average score assigned to people of group 1 who belong to the negative class should be the same as the average score assigned to people of group 2 who belong to the negative class. In other words, the assignment of scores shouldn't be systematically more inaccurate for negative instances in one group than the other.

(C) *Balance for the positive class* symmetrically requires that the average score assigned to people of group 1 who belong to the positive class should be the same as the average score assigned to people of group 2 who belong to the positive class.

Why Do These Conditions Correspond to Notions of Fairness?. All of these are natural conditions to impose on a risk assignment; and as indicated by the discussion above, all of them have been proposed as versions of fairness. The first one essentially asks that the scores mean what they claim to mean, even when considered separately in each group. In particular, suppose a set of scores lack the first property for some bin b, and these scores are given to a decision-maker; then if people of two different groups both belong to bin b, the decision-maker has a clear incentive to treat them differently, since the lack of calibration within groups on bin b means that these people have different aggregate probabilities of belonging to the positive class. Another way of stating the property of calibration within groups is to say that, conditioned on the bin to which an individual is assigned, the likelihood that the individual is a member of the positive class is independent of the group to which the individual belongs. This means we are justified in treating people

[3] As we will discuss in more detail below, the assumption that the group provides no additional information beyond σ does not restrict the generality of the model, since we can always consider instances in which people of different groups never have the same feature vector σ, and hence σ implicitly conveys perfect information about a person's group.

with the same score comparably with respect to the outcome, rather than treating people with the same score differently based on the group they belong to.

The second and third ask that if two individuals in different groups exhibit comparable future behavior (negative or positive), they should be treated comparably by the procedure. In other words, a violation of, say, the second condition would correspond to the members of the negative class in one group receiving consistently higher scores than the members of the negative class in the other group, despite the fact that the members of the negative class in the higher-scoring group have done nothing to warrant these higher scores.

We can also interpret some of the prior work around our earlier examples through the lens of these conditions. For example, in the analysis of the COMPAS risk tool for criminal defendants, the critique by Angwin et al. focused on the risk tool's violation of conditions (B) and (C); the counter-arguments established that it satisfies condition (A). While it is clearly crucial for a risk tool to satisfy (A), it may still be important to know that it violates (B) and (C). Similarly, to think in terms of the example of Internet advertising, with male and female users as the two groups, condition (A) as before requires that our estimates of ad-click probability mean the same thing in aggregate for men and women. Conditions (B) and (C) are distinct; condition (C), for example, says that a female user who genuinely wants to see a given ad should be assigned the same probability as a male user who wants to see the ad.

1.2 Determining What is Achievable: A Characterization Theorem

When can conditions (A), (B), and (C) be simultaneously achieved? We begin with two simple cases where it's possible.

- *Perfect prediction.* Suppose that for each feature vector σ, we have either $p_\sigma = 0$ or $p_\sigma = 1$. This means that we can achieve perfect prediction, since we know each person's class label (positive or negative) for certain. In this case, we can assign all feature vectors σ with $p_\sigma = 0$ to a bin b with score $v_b = 0$, and all σ with $p_\sigma = 1$ to a bin b' with score $v_{b'} = 1$. It is easy to check that all three of the conditions (A), (B), and (C) are satisfied by this risk assignment.
- *Equal base rates.* Suppose, alternately, that the two groups have the same fraction of members in the positive class; that is, the average value of p_σ is the same for the members of group 1 and group 2. (We can refer to this as the *base rate* of the group with respect to the classification problem.) In this case, we can create a single bin b with score equal to this average value of p_σ, and we can assign everyone to bin b. While this is not a particularly informative risk assignment, it is again easy to check that it satisfies fairness conditions (A), (B), and (C).

Our first main result establishes that these are in fact the only two cases in which a risk assignment can achieve all three fairness guarantees simultaneously.

Theorem 1.1 *Consider an instance of the problem in which there is a risk assignment satisfying fairness conditions (A), (B), and (C). Then the instance must either allow for perfect prediction (with p_σ equal to 0 or 1 for all σ) or have equal base rates.*

Thus, in every instance that is more complex than the two cases noted above, there will be some natural fairness condition that is violated by any risk assignment. Moreover, note that this result applies regardless of how the risk assignment is computed; since our framework considers risk assignments to be arbitrary functions from feature vectors to bins labeled with probability estimates, it applies independently of the method — algorithmic or otherwise — that is used to construct the risk assignment.

The conclusions of the first theorem can be relaxed in a continuous fashion when the fairness conditions are only approximate. In particular, for any $\varepsilon > 0$ we can define ε-approximate versions of each of conditions (A), (B), and (C) (specified precisely in the next section), each of which requires that the corresponding equalities between groups hold only to within an error of ε. For any $\delta > 0$, we can also define a δ-approximate version of the equal base rates condition (requiring that the base rates of the two groups be within an additive δ of each other) and a δ-approximate version of the perfect prediction condition (requiring that in each group, the average of the expected scores assigned to members of the positive class is at least $1 - \delta$; by the calibration condition, this can be shown to imply a complementary bound on the average of the expected scores assigned to members of the negative class).

In these terms, our approximate version of Theorem 1.1 is the following.

Theorem 1.2 *There is a continuous function f, with $f(x)$ going to 0 as x goes to 0, so that the following holds. For all $\varepsilon > 0$, and any instance of the problem with a risk assignment satisfying the ε-approximate versions of fairness conditions (A), (B), and (C), the instance must satisfy either the $f(\varepsilon)$-approximate version of perfect prediction or the $f(\varepsilon)$-approximate version of equal base rates.*

Thus, anything that approximately satisfies the fairness constraints must approximately look like one of the two simple cases identified above.

Finally, in connection to Theorem 1.1, we note that when the two groups have equal base rates, then one can ask for the most accurate risk assignment that satisfies all three fairness conditions (A), (B), and (C) simultaneously. Since the risk assignment that gives the same score to everyone satisfies the three conditions, we know that at least one such risk assignment exists; hence, it is natural to seek to optimize over the set of all such assignments. We consider this algorithmic question in the final technical section of the paper.

To reflect a bit further on our main theorems and what they suggest, we note that our intention in the present work isn't to make a recommendation on how conflicts between different definitions of fairness should be handled. Nor is our intention to analyze which definitions of fairness are violated in particular applications or datasets. Rather, our point is to establish certain unavoidable trade-offs between the definitions, regardless of the specific context and regardless of the method used to compute risk scores. Since each of the definitions reflect (and have been proposed as) natural notions of what it should mean for a risk score to be fair, these trade-offs suggest a striking implication: that outside of narrowly delineated cases, any assignment of risk scores can in principle be subject to natural criticisms on the grounds of bias. This is equally true whether the risk score is determined by an algorithm or by a system of human decision-makers.

Special Cases of the Model. Our main results, which place strong restrictions on when the three fairness conditions can be simultaneously satisfied, have more power when the underlying model of the input is more general, since it means that the restrictions implied by the theorems apply in greater generality. However, it is also useful to note certain special cases of our model, obtained by limiting the flexibility of certain parameters in intuitive ways. The point is that our results apply *a fortiori* to these more limited special cases.

First, we have already observed one natural special case of our model: cases in which, for each feature vector σ, only members of one group (but not the other) can exhibit σ. This means that σ contains perfect information about group membership, and so it corresponds to instances in which risk assignments would have the potential to use knowledge of an individual's group membership. Note that we can convert any instance of our problem into a new instance that belongs to this special case as follows. For each feature vector σ, we create two new feature vectors $\sigma^{(1)}$ and $\sigma^{(2)}$; then, for each member of group 1 who had feature vector σ, we assign them $\sigma^{(1)}$, and for each member of group 2 who had feature vector σ, we assign them

$\sigma^{(2)}$. The resulting instance has the property that each feature vector is associated with members of only one group, but it preserves the essential aspects of the original instance in other respects.

Second, we allow risk assignments in our model to split people with a given feature vector σ over several bins. Our results also therefore apply to the natural special case of the model with *integral* risk assignments, in which all people with a given feature σ must go to the same bin.

Third, our model is a generalization of binary classification, which only allows for 2 bins. Note that although binary classification does not explicitly assign scores, we can consider the probability that an individual belongs to the positive class given that they were assigned to a specific bin to be the score for that bin. Thus, our results hold in the traditional binary classification setting as well.

Data-Generating Processes. Finally, there is the question of where the data in an instance of our problem comes from. Our results do not assume any particular process for generating the positive/negative class labels, feature vectors, and group memberships; we simply assume that we are given such a collection of values (regardless of where they came from), and then our results address the existence or non-existence of certain risk assignments for these values.

This increases the generality of our results, since it means that they apply to any process that produces data of the form described by our model. To give an example of a natural generative model that would produce instances with the structure that we need, one could assume that each individual starts with a "hidden" class label (positive or negative), and a feature vector σ is then probabilistically generated for this individual from a distribution that can depend on their class label and their group membership. (If feature vectors produced for the two groups are disjoint from one another, then the requirement that the value of p_σ is independent of group membership given σ necessarily holds.) Since a process with this structure produces instances from our model, our results apply to data that arises from such a generative process.

It is also interesting to note that the basic set-up of our model, with the population divided across a set of feature vectors for which race provides no additional information, is in fact a very close match to the information one gets from the output of a well-calibrated risk tool. In this sense, one setting for our model would be the problem of applying post-processing to the output of such a risk tool to ensure additional fairness guarantees. Indeed, since much of the recent controversy about fair risk scores has involved risk tools that are well-calibrated but lack the other fairness conditions we consider, such an interpretation of the model could be a useful way to think about how one might work with these tools in the context of a broader system.

1.3 Further Related Work

Mounting concern over discrimination in machine learning has led to a large body of new work seeking to better understand and prevent it. Barocas and Selbst survey a range of ways in which data-analysis algorithms can lead to discriminatory outcomes [3], and review articles by Romei and Ruggieri [25] and Zliobaite [30] survey data-analytic and algorithmic methods for measuring discrimination.

Kamiran and Calders [21] and Hajian and Domingo-Ferrer [18] seek to modify datasets to remove any information that might permit discrimination. Similarly, Zemel et al. look to learn fair intermediate representations of data while preserving information needed for classification [29]. Joseph et al. consider how fairness issues can arise during the process of learning, modeling this using a multi-armed bandit framework [20].

Fair prediction with disparate impact:
A study of bias in recidivism prediction instruments

Alexandra Chouldechova [*]

Last revised: February 8, 2017

Abstract

Recidivism prediction instruments (RPI's) provide decision makers with an assessment of the likelihood that a criminal defendant will reoffend at a future point in time. While such instruments are gaining increasing popularity across the country, their use is attracting tremendous controversy. Much of the controversy concerns potential discriminatory bias in the risk assessments that are produced. This paper discusses several fairness criteria that have recently been applied to assess the fairness of recidivism prediction instruments. We demonstrate that the criteria cannot all be simultaneously satisfied when recidivism prevalence differs across groups. We then show how disparate impact can arise when a recidivism prediction instrument fails to satisfy the criterion of error rate balance.

Keywords: disparate impact; bias; recidivism prediction; risk assessment; fair machine learning

1 Introduction

Risk assessment instruments are gaining increasing popularity within the criminal justice system, with versions of such instruments being used or considered for use in pre-trial decision-making, parole decisions, and in some states even sentencing[1,2,3]. In each of these cases, a high-risk classification—particularly a high-risk misclassification—may have a direct adverse impact on a criminal defendant's outcome. If the use of RPI's is to become commonplace, it is especially important to ensure that the instruments are free from discriminatory biases that could result in unethical practices and inequitable outcomes for different groups.

In a recent widely popularized investigation conducted by a team at ProPublica, Angwin et al.[4] studied an RPI called COMPAS[a], concluding that it is biased against black defendants. The authors

[*] Heinz College, Carnegie Mellon University

[a] COMPAS[5] is a risk assessment instrument developed by Northpointe Inc.. Of the 22 scales that COMPAS provides, the Recidivism risk and Violent Recidivism risk scales are the most widely used. The empirical results in this paper are based on decile scores coming from the COMPAS Recidivism risk scale.

found that the likelihood of a non-recidivating black defendant being assessed as high risk is nearly twice that of white defendants. Similarly, the likelihood of a recidivating black defendant being assessed as low risk is nearly half that of white defendants. In technical terms, these findings indicate that the COMPAS instrument has considerably higher false positive rates and lower false negative rates for black defendants than for white defendants.

ProPublica's analysis has met with much criticism from both the academic community and from the Northpointe corporation. Much of the criticism has focussed on the particular choice of fairness criteria selected for the investigation. Flores et al.[6] argue that the correct approach for assessing RPI bias is instead to check for *calibration*, a fairness criterion that they show COMPAS satisfies. Northpointe in their response[7] argue for a still different approach that checks for a fairness criterion termed *predictive parity*, which they demonstrate COMPAS also satisfies. We provide precise definitions and a more in-depth discussion of these and other fairness criteria in Section 2.1.

In this paper we show that the differences in false positive and false negative rates cited as evidence of racial bias by Angwin et al.[4] are a direct consequence of applying an RPI that that satisfies predictive parity to a population in which recidivism prevalence[a] differs across groups. Our main contribution is twofold. (1) First, we make precise the connection between the predictive parity criterion and error rates in classification. (2) Next, we demonstrate how using an RPI that has different false postive and false negative rates between groups can lead to disparate impact when individuals assessed as high risk receive stricter penalties. Throughout our discussion we use the term *disparate impact* to refer to settings where a penalty policy has unintended disproportionate adverse impact on a particular group.

It is important to bear in mind that fairness itself—along with the notion of disparate impact—is a social and ethical concept, not a statistical one. A risk prediction instrument that is fair with respect to particular fairness criteria may nevertheless result in disparate impact depending on how and where it is used. In this paper we consider hypothetical use cases in which we are able to directly connect particular fairness properties of an RPI to a measure of disparate impact. We present both theoretical and empirical results to illustrate how disparate impact can arise.

1.1 Outline of paper

We begin in Section 2 by providing some background on several of the different fairness criteria that have appeared in recent literature. We then proceed to demonstrate that an instrument that satisfies predictive parity cannot have equal false positive and negative rates across groups when the recidivism prevalence differs across those groups. In Section 3 we analyse a simple risk assessment-based sentencing policy and show how differences in false positive and false negative rates can result in disparate impact under this policy. In Section 3.3 we back up our theoretical analysis by presenting some empirical results based on the data made available by the ProPublica investigators. We conclude with a discussion of the issues that biased data presents for the arguments put forth in this paper.

[a] *Prevalence*, also termed the *base rate*, is the proportion of individuals who recidivate in a given population.

1.2 Data description and setup

The empirical results in this paper are based on the Broward County data made publicly available by ProPublica[8]. This data set contains COMPAS recidivism risk decile scores, 2-year recidivism outcomes, and a number of demographic and crime-related variables on individuals who were scored in 2013 and 2014. We restrict our attention to the subset of defendants whose race is recorded as African-American (b) or Caucasian (w).[a] After applying the same data pre-processing and filtering as reported in the ProPublica analysis, we are left with a data set on $n = 6150$ individuals, of whom $n_b = 3696$ are African-American and $n_c = 2454$ are Caucasian.

2 Assessing fairness

2.1 Background

We begin by with some notation. Let $S = S(x)$ denote the risk score based on covariates $X = x \in \mathbb{R}^p$, with higher values of S corresponding to higher levels of assessed risk. We will interchangeably refer to S as a *score* or an *instrument*. For simplicity, our discussion of fairness criteria will focus on a setting where there exist just two groups. We let $R \in \{b, w\}$ denote the group to which an individual belongs, and do not preclude R from being one of the elements of X. We denote the outcome indicator by $Y \in \{0, 1\}$, with $Y = 1$ indicating that the given individual goes on to recidivate. Lastly, we introduce the quantity s_{HR}, which denotes the high-risk score threshold. Defendants whose score S exceeds s_{HR} will be referred to as *high-risk*, while the remaining defendants will be referred to as *low-risk*.

With this notation in hand, we now proceed to define and discuss several fairness criteria that commonly appear in the literature, beginning with those mentioned in the introduction. We indicate cases where a given criterion is known to us to also commonly appear under some other name. All of the criteria presented below can also be assessed *conditionally* by further conditioning on some covariates in X. We discuss this point in greater detail in Section 3.1.

Definition 1 (Calibration)**.** A score $S = S(x)$ is said to be *well-calibrated* if it reflects the same likelihood of recidivism irrespective of the individuals' group membership. That is, if for all values of s,

$$\mathbb{P}(Y = 1 \mid S = s, R = b) = \mathbb{P}(Y = 1 \mid S = s, R = w). \tag{2.1}$$

Within the educational and psychological testing and assessment literature, the notion of *calibration* features among the widely accepted and adopted standards for empirical fairness assessment. In this literature, an instrument that is *well-calibrated* is referred to as being *free from predictive bias*. This criterion has recently been applied to the PCRA[b] instrument, with initial findings suggesting that calibration is satisfied with respect race[10,11], but not with respect to gender[12]. In

[a] There are 6 racial groups represented in the data. 85% of individuals are either African-American or Caucasian.

[b] The Post Conviction Risk Assessment (PCRA) tool was developed by the Administrative Office of the United States Courts for the purpose of improving "the effectiveness and efficiency of post-conviction supervision"[9]

their response to the ProPublica investigation, Flores et al.[6] verify that COMPAS is well-calibrated using logistic regression modeling.

Definition 2 (Predictive parity)**.** A score $S = S(x)$ satisfies *predictive parity* at a threshold s_{HR} if the likelihood of recidivism among high-risk offenders is the same regardless of group membership. That is, if,

$$\mathbb{P}(Y = 1 \mid S > s_{\text{HR}}, R = b) = \mathbb{P}(Y = 1 \mid S > s_{\text{HR}}, R = w). \tag{2.2}$$

Predictive parity at a given threshold s_{HR} amounts to requiring that the *positive predictive value* (PPV) of the classifier $\hat{Y} = \mathbb{1}_{S > s_{\text{HR}}}$ be the same across groups. While predictive parity and calibration look like very similar criteria, well-calibrated scores can fail to satisfy predictive parity at a given threshold. This is because the relationship between (2.2) and (2.1) depends on the conditional distribution of $S \mid R = r$, which can differ across groups in ways that result in PPV imbalance. In the simple case where S itself is binary, a score that is well-calibrated will also satisfy predictive parity. Northpointe's refutation[7] of the ProPublica analysis shows that COMPAS satisfies predictive parity for threshold choices of interest.

Definition 3 (Error rate balance)**.** A score $S = S(x)$ satisfies *error rate balance* at a threshold s_{HR} if the false positive and false negative error rates are equal across groups. That is, if,

$$\mathbb{P}(S > s_{\text{HR}} \mid Y = 0, R = b) = \mathbb{P}(S > s_{\text{HR}} \mid Y = 0, R = w)\ , \quad \text{and} \tag{2.3}$$

$$\mathbb{P}(S \leq s_{\text{HR}} \mid Y = 1, R = b) = \mathbb{P}(S \leq s_{\text{HR}} \mid Y = 1, R = w), \tag{2.4}$$

where the expressions in the first line are the group-specific false positive rates, and those in the second line are the group-specific false negative rates.

ProPublica's analysis considered a threshold of $s_{\text{HR}} = 4$, which they showed leads to considerable imbalance in both false positive and false negative rates. While this choice of cutoff met with some criticism, we will see later in this section that error rate imbalance persists—indeed, must persist—for any choice of cutoff at which the score satisfies the predictive parity criterion. Error rate balance is also closely connected to the notions of *equalized odds* and *equal opportunity* as introduced in the recent work of Hardt et al.[13].

Definition 4 (Statistical parity)**.** A score $S = S(x)$ satisfies *statistical parity* at a threshold s_{HR} if the proportion of individuals classified as high-risk is the same for each group. That is, if,

$$\mathbb{P}(S > s_{\text{HR}} \mid R = b) = \mathbb{P}(S > s_{\text{HR}} \mid R = w) \tag{2.5}$$

Statistical parity also goes by the name of *equal acceptance rates*[14] or *group fairness*[15], though it should be noted that these terms are in many cases not used synonymously. While our discussion focusses primarily on first three fairness criteria, statistical parity is widely used within the machine learning community and may be the criterion with which many readers are most familiar[16,17]. Statistical parity is well-suited to contexts such as employment or admissions, where it may be desirable or required by law or regulation to employ or admit individuals in equal proportion across racial, gender, or geographical groups. It is, however, a difficult criterion to motivate in the recidivism prediction setting, and thus will not be further considered in this work.

2.2 Further related work

Though the study of discrimination in decision making and predictive modeling is rapidly evolving, it also has a long and rich multidisciplinary history. Romei and Ruggieri[18] provide an excellent overview of some of the work in this broad subject area. The recent work of Barocas and Selbst[19] offers a broad examination of algorithmic fairness framed within the context of anti-discrimination laws governing employment practices. Hannah-Moffat[20], Skeem[21], and Monahan and Skeem[22] examine legal and ethical issues relating specifically to the use of risk assessment instruments in sentencing, citing the potential for race and gender discrimination as a major concern.

In work concurrent with our own, several other researchers have also investigated the compatibility of different notions of fairness. Kleinberg et al.[23] show that calibration cannot be satisfied simultaneously with the fairness criteria of *balance for the negative class* and *balance for the positive class*. Translated into the present context, the latter criteria require that the average score assigned to non-recidivists (the negative class) should be the same for both groups, and that the same should hold among recidivists (the positive class). The work of Corbett-Davies et al.[24] closely parallels the results that we present in Section 2.3, reaching the same conclusion regarding the incompatibility of predictive parity and error rate balance in the setting of unequal prevalence.

2.3 Predictive parity, false positive rates, and false negative rates

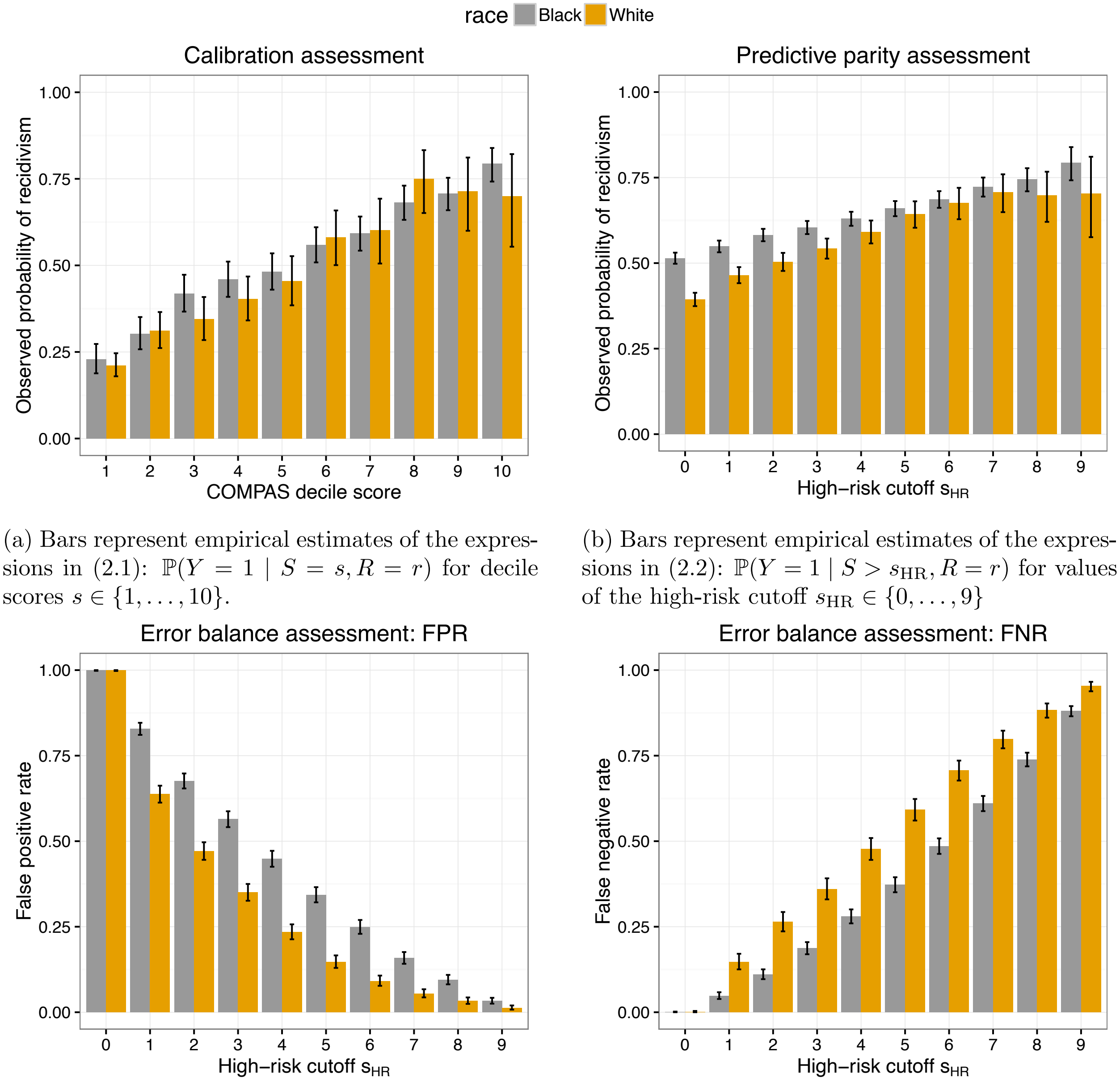
In this section we present our first main result, which establishes that predictive parity is incompatible with error rate balance when prevalence differs across groups. To better motivate the discussion, we begin by presenting an empirical fairness assessment of the COMPAS RPI. Figure 1 shows plots of the observed recidivism rates and error rates corresponding to the fairness notions of calibration, predictive parity, and error rate balance. We see that the COMPAS RPI is (approximately) well-calibrated, and also satisfies predictive parity provided that the high-risk cutoff s_{HR} is 4 or greater. However, COMPAS fails on both false positive and false negative error rate balance across the range of high-risk cutoffs.

Angwin et al.[4] focussed on a high-risk cutoff of $s_{\mathrm{HR}} = 4$ for their analysis, which some critics have argued is too low, suggesting that $s_{\mathrm{HR}} = 7$ is more suitable. As can be seen from Figures 1c and 1d, significant error rate imbalance persists at this cut-off as well. Moreover, the error rates achieved at so high a cutoff are at odds with evidence suggesting that the use of RPI's is of interest in settings where false negatives have a higher cost than false positives, with relative cost estimates ranging from 2.6 to upwards of 15.[25,26]

As we now proceed to show, the error rate imbalance exhibited by COMPAS is not a coincidence, nor can it be remedied in the present context. When the recidivism prevalence–i.e., the base rate $\mathbb{P}(Y = 1 \mid R = r)$—differs across groups, any instrument that satisfies predictive parity at a given threshold s_{HR} *must* have imbalanced false positive or false negative errors rates at that threshold. To understand why predictive parity and error rate balance are mutually exclusive in the setting of unequal recidivism prevalence, it is instructive to think of how these quantities are all related.

Given a particular choice of s_{HR}, we can summarize an instrument's performance in terms of a confusion matrix, as shown in Table 1 below.
All of the fairness metrics presented in Section 2.1 can be thought of as imposing constraints on

the values (or the distribution of values) in this table. Another constraint—one that we have no direct control over—is imposed by the recidivism prevalence within groups. It is not difficult to



(a) Bars represent empirical estimates of the expressions in (2.1): $\mathbb{P}(Y = 1 \mid S = s, R = r)$ for decile scores $s \in \{1, \ldots, 10\}$.

(b) Bars represent empirical estimates of the expressions in (2.2): $\mathbb{P}(Y = 1 \mid S > s_{\mathrm{HR}}, R = r)$ for values of the high-risk cutoff $s_{\mathrm{HR}} \in \{0, \ldots, 9\}$

(c) Bars represent observed false positive rates, which are empirical estimates of the expressions in (2.3): $\mathbb{P}(S > s_{\mathrm{HR}} \mid Y = 0, R = r)$ for values of the high-risk cutoff $s_{\mathrm{HR}} \in \{0, \ldots, 9\}$

(d) Bars represent observed false negative rates, which are empirical estimates of the expressions in (2.4): $\mathbb{P}(S \leq s_{\mathrm{HR}} \mid Y = 1, R = r)$ for values of the high-risk cutoff $s_{\mathrm{HR}} \in \{0, \ldots, 9\}$

Figure 1: Empirical assessment of the COMPAS RPI according to three of the fairness criteria presented in Section 2.1. Error bars represent 95% confidence intervals. These Figures confirm that COMPAS is (approximately) well-calibrated, satisfies predictive parity for high-risk cutoff values of 4 or higher, but fails to have error rate balance.

	Low-Risk	High-Risk
$Y = 0$	TN	FP
$Y = 1$	FN	TP

Table 1: T/F denote True/False and N/P denote Negative/Positive. For instance, FP is the number of false positives: individuals who are classified as high-risk but who do not reoffend.

show that the prevalence (p), positive predictive value (PPV), and false positive and negative error rates (FPR, FNR) are related via the equation

$$\mathrm{FPR} = \frac{p}{1-p}\frac{1-\mathrm{PPV}}{\mathrm{PPV}}(1-\mathrm{FNR}). \tag{2.6}$$

From this simple expression we can see that if an instrument satisfies predictive parity—that is, if the PPV is the same across groups—but the prevalence differs between groups, the instrument cannot achieve equal false positive and false negative rates across those groups.

This observation enables us to better understand why we observe such large discrepancies in FPR and FNR between black and white defendants in Figure 1. The recidivism rate among black defendants in the data is 51%, compared to 39% for White defendants. Thus at any threshold s_{HR} where the COMPAS RPI satisfies predictive parity, equation (2.6) tells us that some level of imbalance in the error rates must exist. Since not all of the fairness criteria can be satisfied at the same time, it becomes important to understand the potential impact of failing to satisfy particular criteria. This question is explored in the context of a hypothetical risk-based sentencing framework in the next section.

3 Assessing impact

In this section we show how differences in false positive and false negative rates can result in disparate impact under policies where a high-risk assessment results in a stricter penalty for the defendant. Such situations may arise when risk assessments are used to inform bail, parole, or sentencing decisions. In Pennsylvania and Virginia, for instance, statutes permit the use of RPI's in sentencing, provided that the sentence ultimately falls within accepted guidelines[1]. We use the term "penalty" somewhat loosely in this discussion to refer to outcomes both in the pre-trial and post-conviction phase of legal proceedings. For instance, even though pre-trial outcomes such as the amount at which bail is set are not punitive in a legal sense, we nevertheless refer to bail amount as a "penalty" for the purpose of our discussion.

There are notable cases where RPI's are used for the express purpose of informing risk reduction efforts. In such settings, individuals assessed as high risk receive what may be viewed as a benefit rather than a penalty. The PCRA score, for instance, is intended to support precisely this type of decision-making at the federal courts level[11]. Our analysis in this section specifically addresses use cases where high-risk individuals receive stricter penalties.

To begin, consider a setting in which guidelines indicate that a defendant is to receive a penalty