Introduction and Algorithmic Fairness

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

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

MIRROR, MIRROR

TERMS OF USE

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:

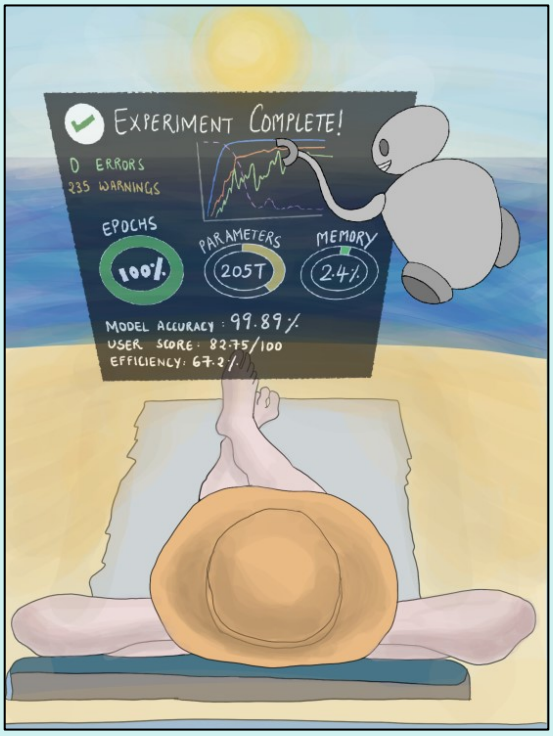
Please direct any queries about using elements from this comic to themachinelearnist@gmail.com and cc stoyanovich@nyu.edu

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...

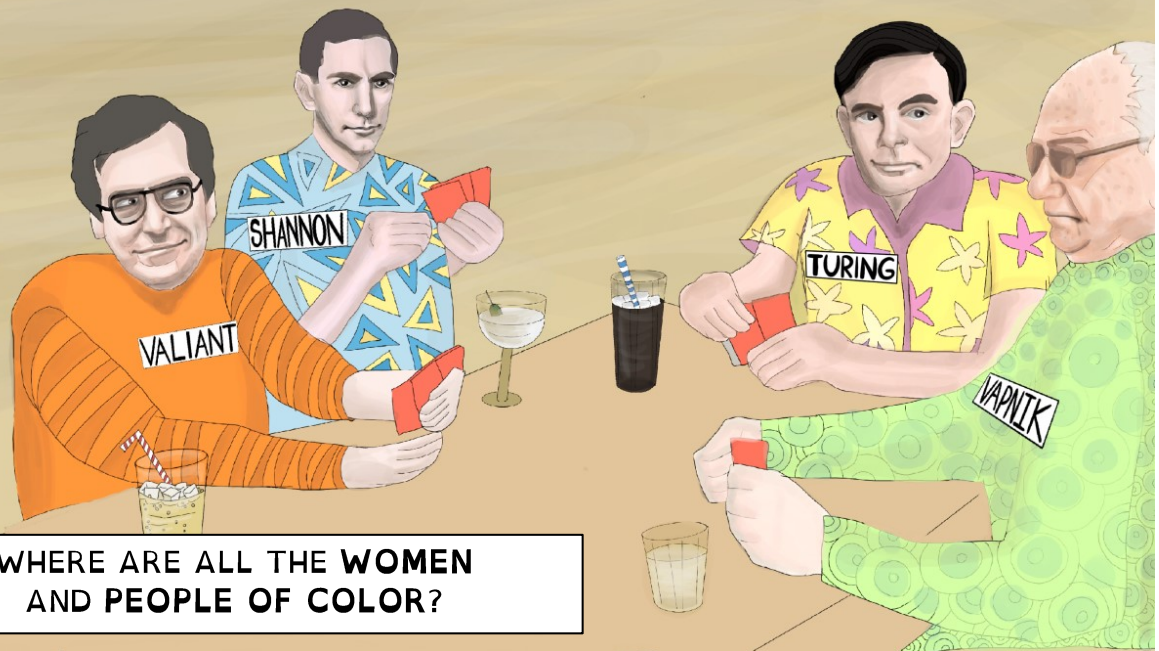
THEY WERE RIGHT. THEY DID IT.

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?)

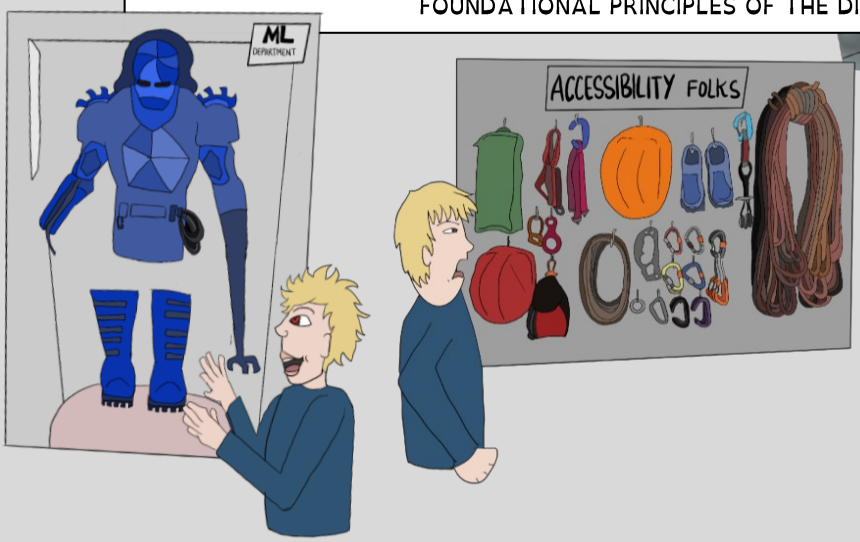
AI IS THE SHINIEST TOY ON THE BLOCK AND SO, INEVITABLY, ALL THE **MONEY-MAGPIES** HAVE COME FLOCKING.

BUDGET CUTS

WHAAA

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.

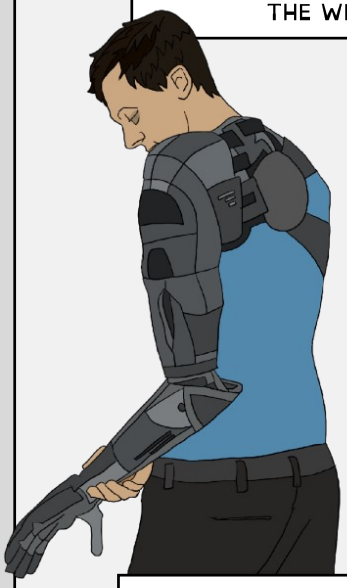
LET'S GIVE ***HCI*** A MOMENT IN THE **SPOTLIGHT**, SHALL WE?

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...

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.

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

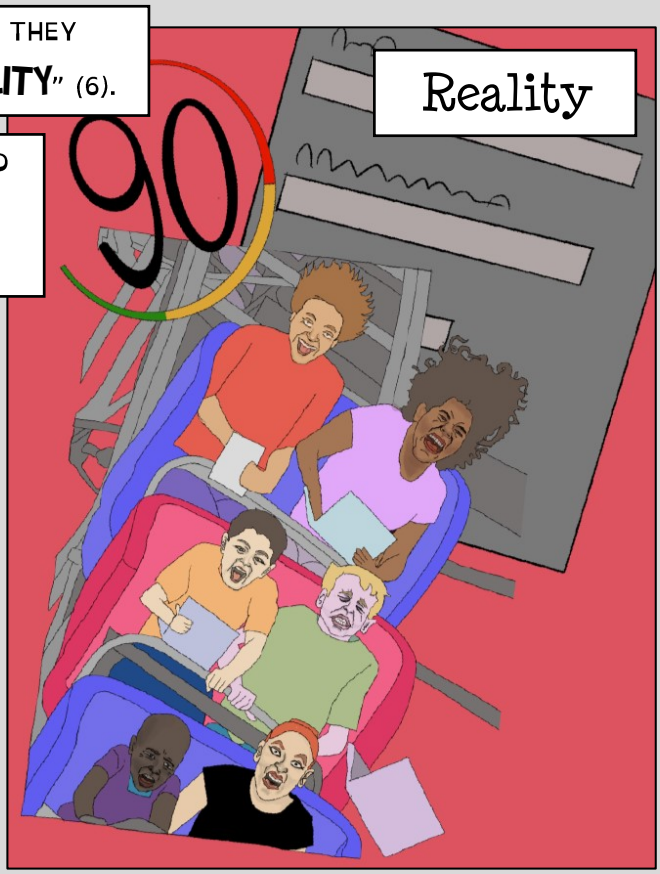
(WHO ARE WE BUILDING MODELS FOR?)

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''?**

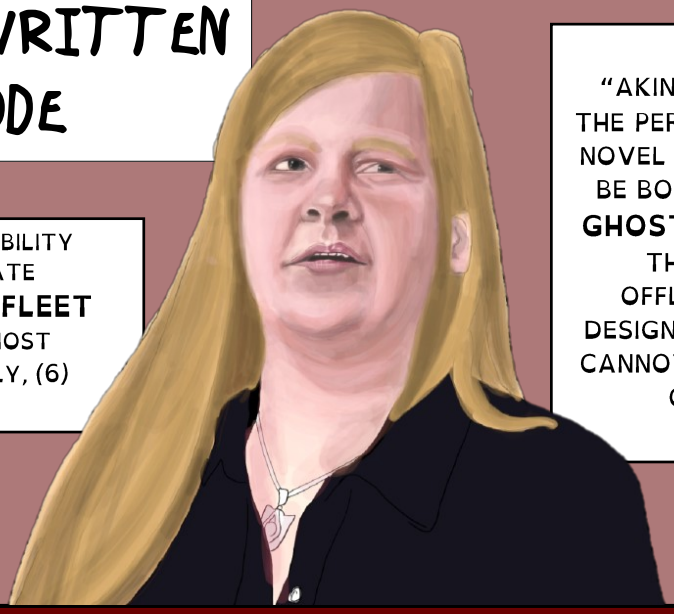
WHEN PLATFORMS ARE NOT DESIGNED FOR EVERYONE, THEY GIVE OFF THE STENCH OF "**ENCODED INHOSPITALITY**" (6).

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.



GHOSTWRITTEN CODE

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



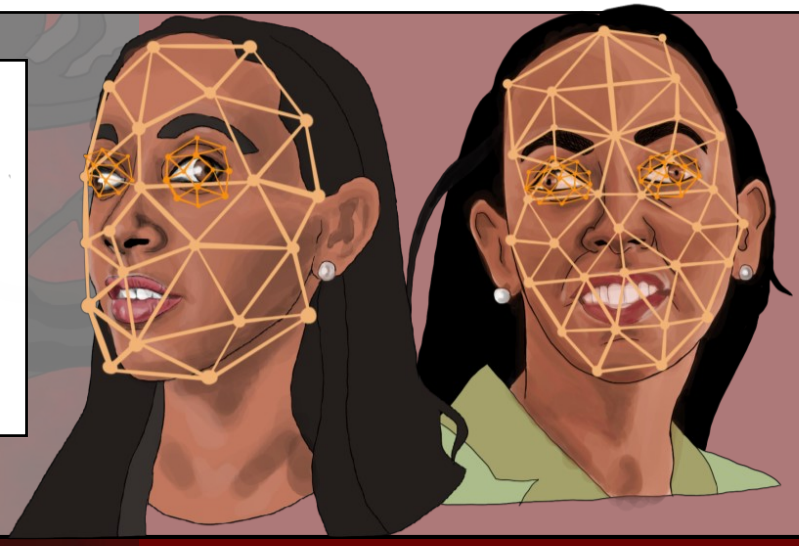
"AKIN TO HOW A **GHOSTWRITER** IS THE PERSON WHO IS PAID TO COMPOSE A NOVEL THAT SOMEONE ELSE COULD NOT BE BOTHERED TO WRITE THEMSELVES, **GHOSTWRITTEN CODE** IS SOFTWARE THAT THE ORGANIZATION HAS OFFLOADED ON PROGRAMMERS TO DESIGN FOR USERS THAT THE COMPANY CANNOT BE BOTHERED TO ENGAGE WITH OR EMPLOY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

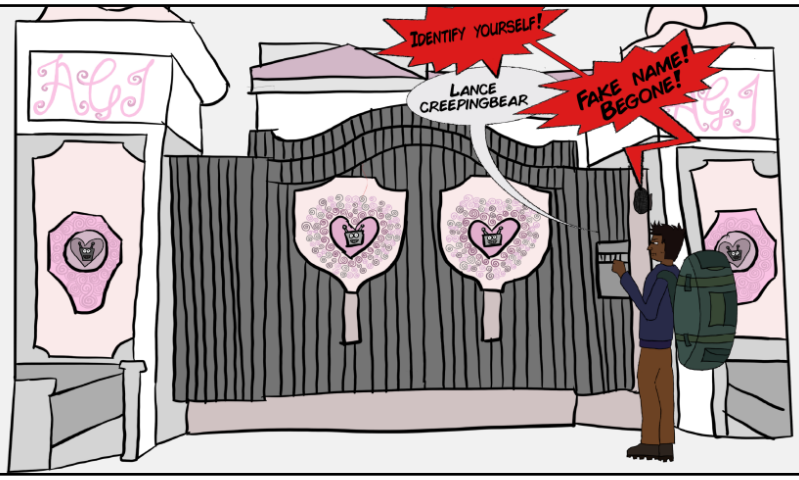
HAVING BEEN TRAINED ON THE FACIAL DYNAMICS OF SIGHTED WHITE PEOPLE, FACIAL RECOGNITION TECHNOLOGY PEDDLES AN ABLEIST AND RACIST NARRATIVE.

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?

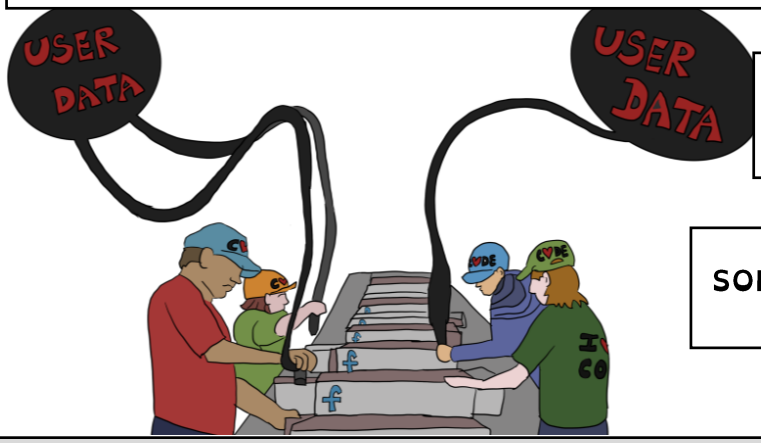
TAKE FACEBOOK'S **"REAL NAME" POLICY** THAT INDISCRIMINATELY TARGETED NATIVE AMERICANS (8)

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**

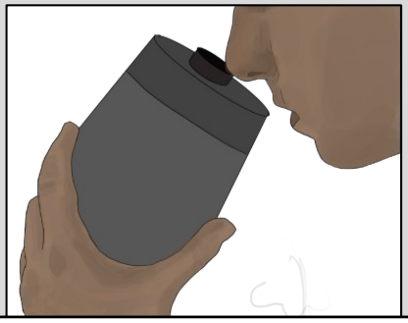
PART 3: THE POISONING

(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...

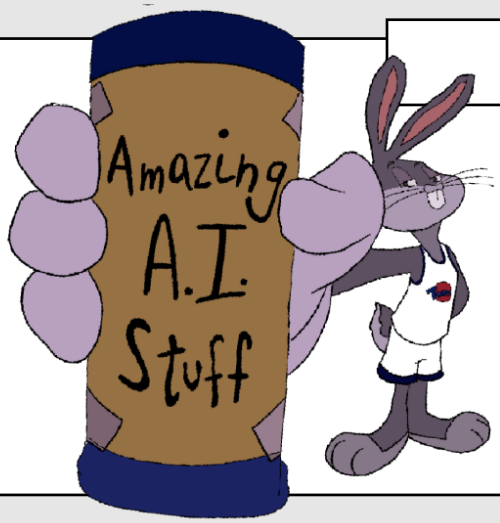


YOU MIGHT NOT WANT TO DRINK THAT!

IT'S SNAKE OIL!

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

WHAT IS AI-SNAKE OIL?

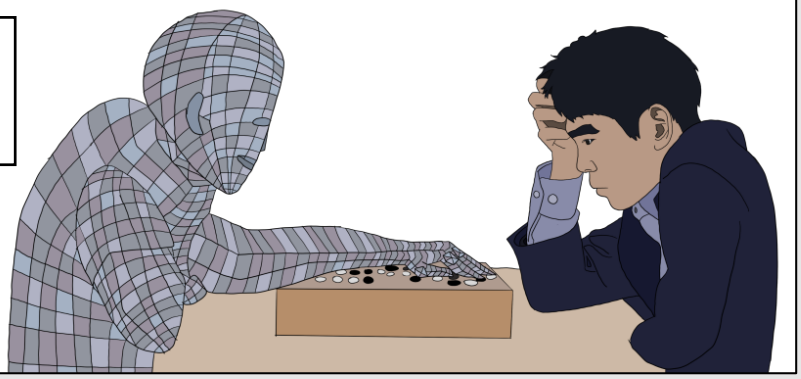


SNAKE OIL IS THE MYSTICAL SUBSTANCE THAT IS CREATED BY TAKING EQUAL PARTS MEDIA HYPE AND PUBLIC MISINFORMATION AND STIRRING THEM INTO A POTION, WITH AN IRRESISTIBLE LABEL THAT SCREAMS "DATA" AND "INTELLIGENCE"

... 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!**

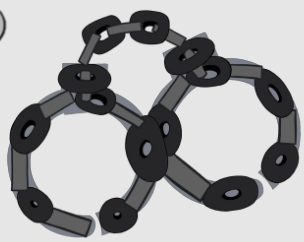


PERCEPTION TASKS, SUCH AS **FACIAL RECOGNITION**, THAT ARE INTERTWINED WITH THE **SOCIAL, POLITICAL AND CULTURAL UNDERPINNINGS** OF THE DATA ON WHICH THEY WERE TRAINED, ARE FAR MORE TOXIC.

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.

SOTA-CHASING

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**

DATA + COMPUTE

STATISTICAL LEARNING THEORY

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.

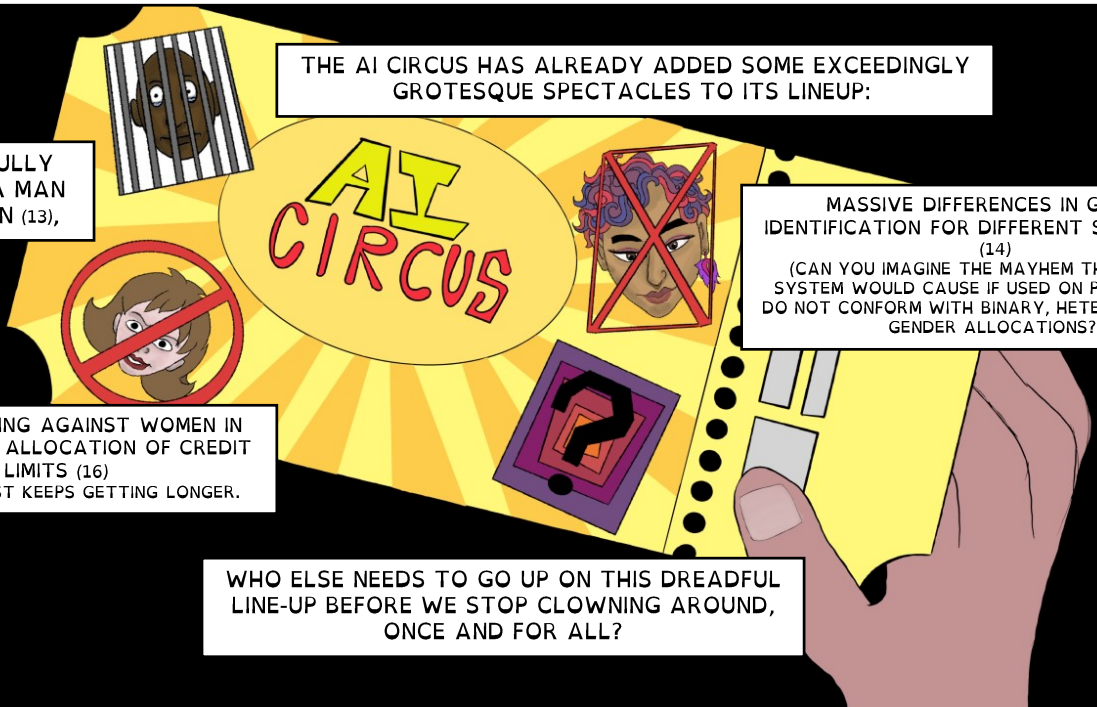
AND THIS IS NEVER THE MAJORITY.

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.

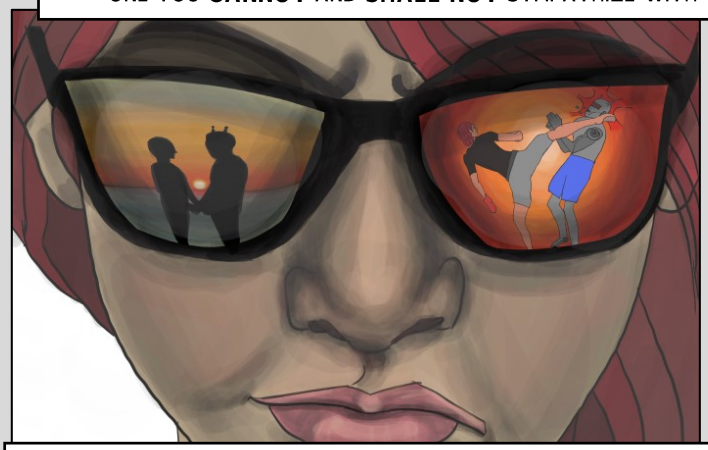
WHO ELSE NEEDS TO GO UP ON THIS DREADFUL LINE-UP BEFORE WE STOP CLOWNING AROUND, ONCE AND FOR ALL?

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
WAY OF THE FUTURE
CANCEL TECH
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.

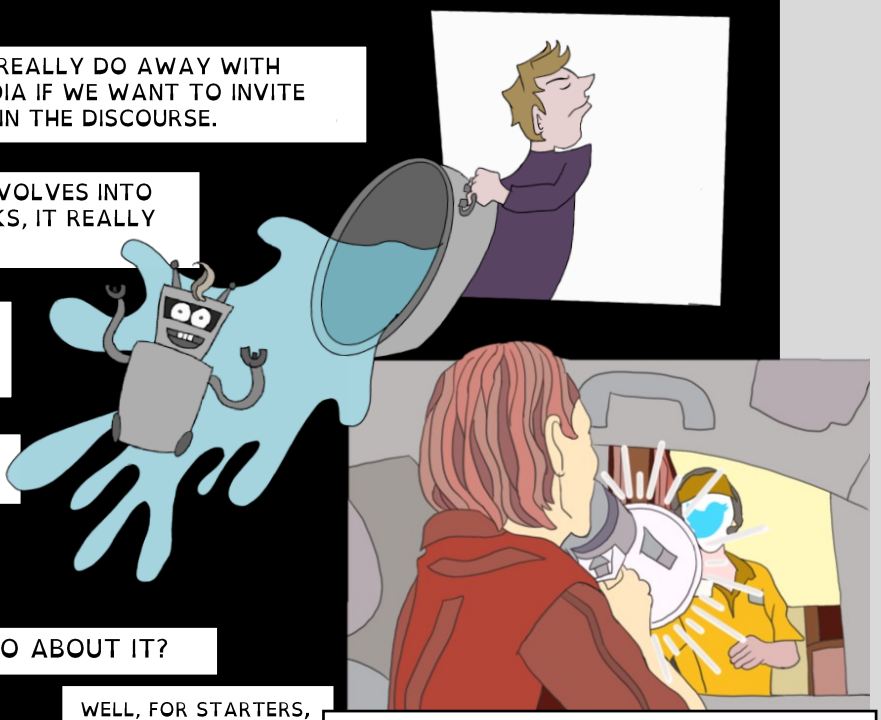
MAYBE WE NEED TO STOP REACTING TO EVERYTHING WE READ AND INSTEAD TAKE A MOMENT TO RE-READ, THINK DEEPLY AND THEN RESPOND.

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.**

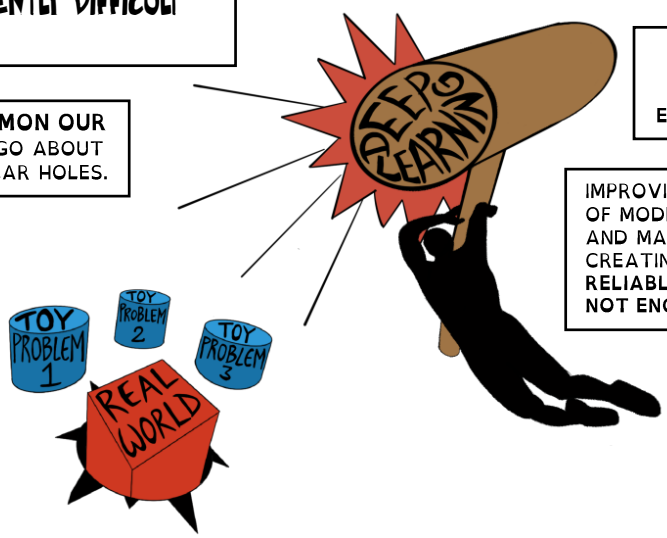
WE HAVE BEEN **DANGEROUSLY CONFLATING** EXPERTISE IN TRAINING AND DEPLOYING A MODEL WITH DOMAIN EXPERTISE.

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.**

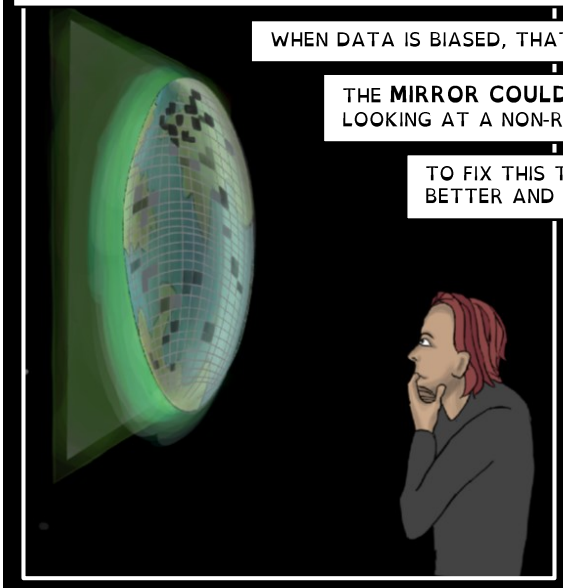
BUT WE AREN'T THERE YET.

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

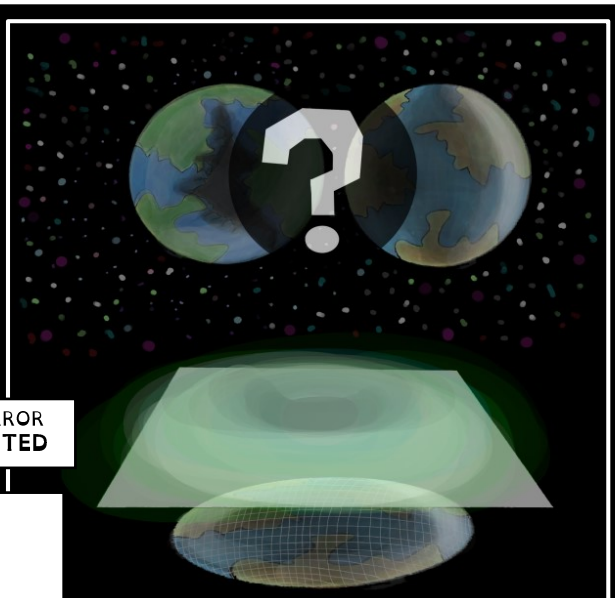
DATA IS A MIRROR REFLECTION OF THE WORLD (18)



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** *DOES NOT* 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.**

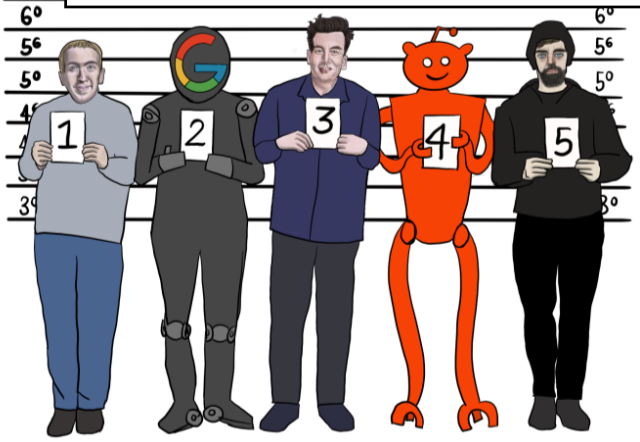
PREDICTION IS DIFFICULT, ESPECIALLY ABOUT THE ***FUTURE***! (19)

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.

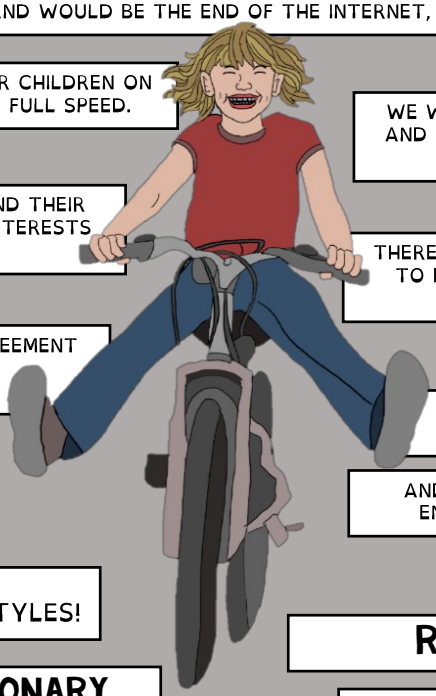
AND SO WHILE TECH COMPANIES AND THEIR CELEBRITY CEOS PROTECT THEIR INTERESTS BY BAD MOUTHING REGULATION,

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

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.

V/S

RISK-BASED

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**!

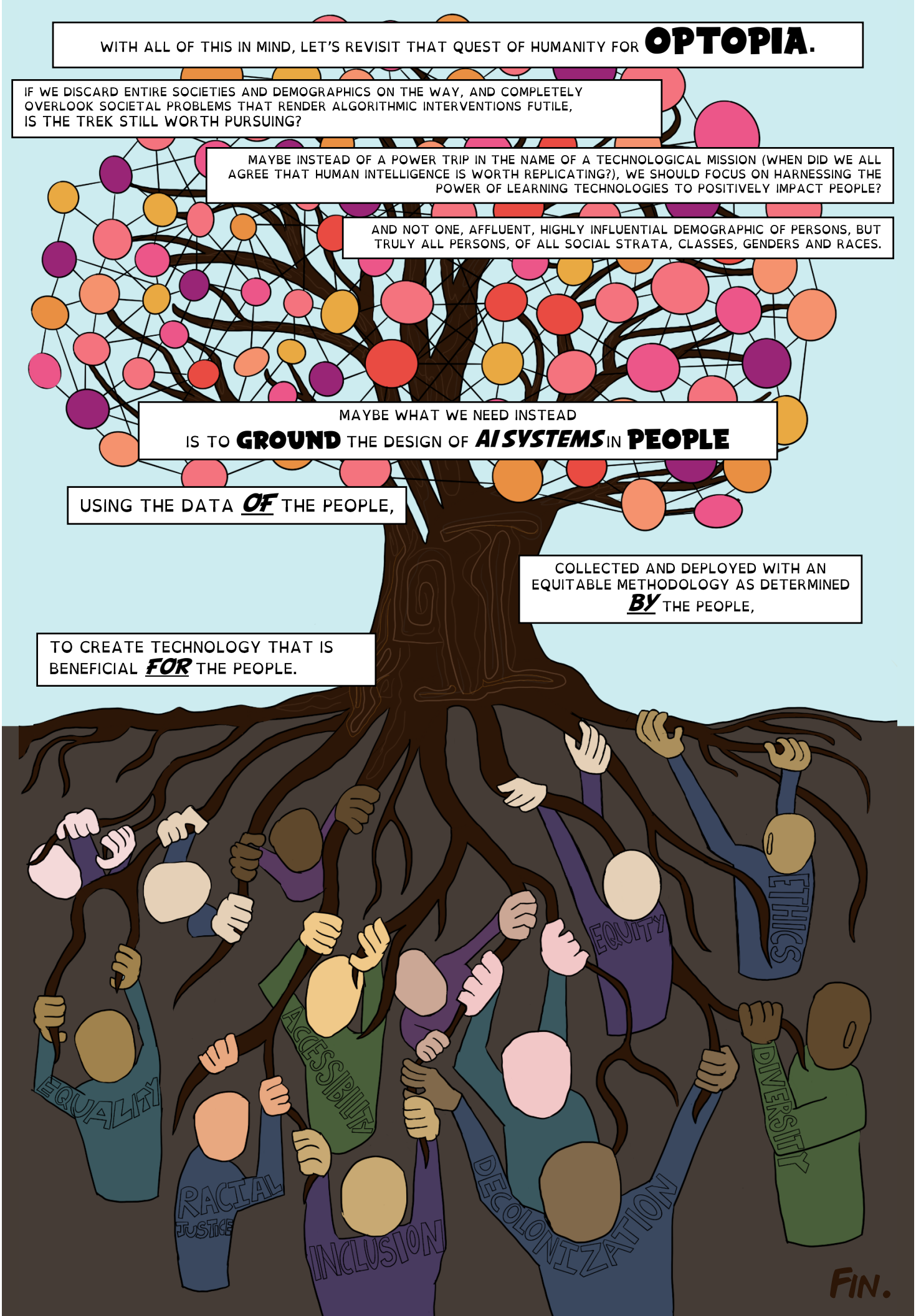
TAKE THE **NYC AUTOMATED DECISION SYSTEMS (ADS) TASK FORCE**, THE FIRST OF ITS KIND IN THE U.S., ENVISIONED TO BE THE BEACON FOR **TRANSPARENCY** AND **EXPERT INSIGHT** INTO THE USE OF ALGORITHMS TO AID DECISION-MAKING BY CITY AGENCIES. (20)

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.
EQUALITY
ACCESSIBILITY
EQUITY
ETHICS
RACIAL JUSTICE
INCLUSION
DECOLONIZATION
DIVERSITY
FIN.

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

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.

.png .pdf .jpg .jpeg .bmp .tiff .tif .gif .eps .ps .eps.gz .ps.gz .eps.Z

Week 2: Algorithmic Fairness

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 word-2vec 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 bi-

ases. 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.

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

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

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.

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.

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.

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 work-

shop; 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. C

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

Algorithmic Fairness

DANA PESSACH*, Department of Industrial Engineering, Tel-Aviv University, Israel
EREZ SHMUELI, Department of Industrial Engineering, Tel-Aviv University, Israel

An increasing number of decisions regarding the daily lives of human beings are being controlled by artificial intelligence (AI) algorithms in spheres ranging from healthcare, transportation, and education to college admissions, recruitment, provision of loans and many more realms. Since they now touch on many aspects of our lives, it is crucial to develop AI algorithms that are not only accurate but also objective and fair. Recent studies have shown that algorithmic decision-making may be inherently prone to unfairness, even when there is no intention for it. This paper presents an overview of the main concepts of identifying, measuring and improving algorithmic fairness when using AI algorithms. The paper begins by discussing the causes of algorithmic bias and unfairness and the common definitions and measures for fairness. Fairness-enhancing mechanisms are then reviewed and divided into pre-process, in-process and post-process mechanisms. A comprehensive comparison of the mechanisms is then conducted, towards a better understanding of which mechanisms should be used in different scenarios. The paper then describes the most commonly used fairness-related datasets in this field. Finally, the paper ends by reviewing several emerging research sub-fields of algorithmic fairness.

Keywords: Algorithmic Bias, Algorithmic Fairness, Fairness-Aware Machine Learning.

1 INTRODUCTION

Nowadays, an increasing number of decisions are being controlled by artificial intelligence (AI) algorithms, with increased implementation of automated decision-making systems in business and government applications. The motivation for an automated learning model is clear – we expect algorithms to perform better than human beings for several reasons: First, algorithms may integrate much more data than a human may grasp and take many more considerations into account. Second, algorithms can perform complex computations much faster than human beings. Third, human decisions are subjective, and they often include biases.

Hence, it is a common belief that using an automated algorithm makes decisions more objective or fair. However, this is unfortunately not the case since AI algorithms are not always as objective as we would expect. The idea that AI algorithms are free from biases is wrong since the assumption that the data injected into the models are unbiased is wrong. More specifically, a prediction model may actually be inherently biased since it learns and preserves historical biases [85].

Since many automated decisions (including which individuals will receive jobs, loans, medication, bail or parole) can significantly impact people's lives, there is great importance in assessing and improving the ethics of the decisions made by these automated systems. Indeed, in recent years, the concern for algorithm fairness has made headlines. One of the most common examples was in the field of criminal justice, where recent revelations have shown that an algorithm used by the United States criminal justice system had falsely predicted future criminality among African-Americans at twice the rate as it predicted for white people [6, 36]. In another case of a hiring application, it was recently exposed that Amazon discovered that their AI hiring system was discriminating against female candidates, particularly for software development and technical positions. One suspected reason for this is that most recorded historical data were for male software developers [40]. In a

*Corresponding author

Authors' addresses: Dana Pessach, Department of Industrial Engineering, Tel-Aviv University, P.O. Box 39040, 6997801, Tel-Aviv, Israel, danapessach@gmail.com; Erez Shmueli, Department of Industrial Engineering, Tel-Aviv University, P.O. Box 39040, 6997801, Tel-Aviv, Israel, shmueli@tau.ac.il.

different scenario in advertising, it was shown that Google's ad-targeting algorithm had proposed higher-paying executive jobs more for men than for women [41, 125].

These lines of evidence and concerns about algorithmic fairness have led to growing interest in the literature on defining, evaluating and improving fairness in AI algorithms (see, for example, [15, 37, 57, 68]). It is important to note, however, that the task of improving fairness of AI algorithms is not trivial since there exists an inherent trade-off between accuracy and fairness. That is, as we pursue a higher degree of fairness, we may compromise accuracy (see, for example, [85]).

In contrast to other recent surveys in this field [37, 57], our paper proposes a comprehensive and up-to-date overview of the field, ranging from definitions and measures of fairness to state-of-the-art fairness-enhancing mechanisms. Our survey also attempts to cover the pros and cons of the various measures and mechanisms, and guide under which setting they should be used. Finally, a major goal of this survey is to highlight and discuss emerging areas of research that are expected to grow in the upcoming years. Overall, this survey provides the relevant knowledge to enable new researchers to enter the field, inform current researchers on rapidly evolving sub-fields, and provide practitioners the necessary tools to apply the results.

The rest of this paper is structured as follows: Section 2 discusses the potential causes of algorithmic unfairness; Section 3 presents definitions and measures of fairness and their trade-offs; Section 4 reviews fairness mechanisms and methods and a comparison of the mechanisms, focusing on the pros and cons of each mechanism; Section 5 outlines commonly used fairness-related datasets; Section 6 presents several emerging research sub-fields of algorithmic fairness; and Section 7 provides concluding remarks and sketches several open challenges for future research.

2 POTENTIAL CAUSES OF UNFAIRNESS

The literature has indicated several causes that may lead to unfairness in machine learning [37, 99]:

- Biases already included in the datasets used for learning, which are based on biased device measurements, historically biased human decisions, erroneous reports or other reasons. Machine learning algorithms are essentially designed to replicate these biases.
- Biases caused by missing data, such as missing values or sample/selection biases, which result in datasets that are not representative of the target population.
- Biases that stem from algorithmic objectives, which aim at minimizing overall aggregated prediction errors and therefore benefit majority groups over minorities.
- Biases caused by "proxy" attributes for sensitive attributes. Sensitive attributes differentiate privileged and unprivileged groups, such as race, gender and age, and are typically not legitimate for use in decision making. Proxy attributes are non-sensitive attributes that can be exploited to derive sensitive attributes. In the case that the dataset contains proxy attributes, the machine learning algorithm can implicitly make decisions based on the sensitive attributes under the cover of using presumably legitimate attributes [11].

To illustrate the last cause mentioned above, consider the example depicted in Figure 1. The figure illustrates a case of SAT scores for two sub-populations: a privileged one and an unprivileged one.

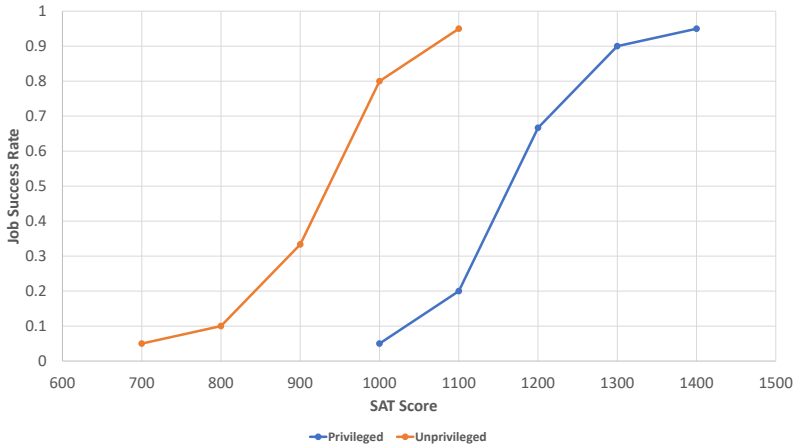


Fig. 1. If the SAT scores were used for hiring, then unprivileged candidates with high potential would be excluded, whereas lower potential candidates from the privileged group would be hired instead

In this illustration, SAT scores may be used to predict the probability of job success when hiring candidates since the higher the SAT score is, the higher the probability of success. However, unprivileged candidates with SAT scores of approximately 1100 perform just as well as privileged candidates with SAT scores of 1400 since they may have encountered more challenging pathways to achieve their scores. In other words, if the SAT scores were used for hiring, unprivileged candidates with high potential would be excluded, whereas lower potential candidates from the privileged group would be hired instead.

3 FAIRNESS DEFINITIONS AND MEASURES

This section presents some general legal notions for discrimination followed by a survey of the most common measures for algorithmic fairness, and the inevitable trade-offs between them.

3.1 Definitions of Discrimination in Legal Domains

The legal domain has introduced two main definitions of discrimination: i) **disparate treatment** [11, 151]: intentionally treating an individual differently based on his/her membership in a protected class (*direct discrimination*); ii) **disparate impact** [11, 119]: negatively affecting members of a protected class more than others even if by a seemingly neutral policy (*indirect discrimination*).

Put in our context, it is important to note that algorithms trained with data that do not include sensitive attributes (i.e., attributes that explicitly identify the protected and unprotected groups) are unlikely to produce *disparate treatment*, but may still induce unintentional discrimination in the form of *disparate impact* [85].

3.2 Measures of Algorithmic Bias

This section presents the most prominent measures of algorithmic fairness in machine learning classification tasks. We refer the readers to the Appendix and specifically to Table 4 for a review of additional, less popular measures used in the literature.

(1) **Disparate impact** [54] – This measure was designed to mathematically represent the legal notion of *disparate impact*. It requires a high ratio between the positive prediction rates of both groups. This ensures that the proportion of the positive predictions is similar across groups. For example, if a positive prediction represents acceptance for a job, the condition

requires the proportion of accepted applicants to be similar across groups. Formally, this measure is computed as follows:

$$\frac{P[\hat{Y} = 1|S \neq 1]}{P[\hat{Y} = 1|S = 1]} \geq 1 - \varepsilon \tag{1}$$

where S represents the protected attribute (e.g., race or gender), $S = 1$ is the privileged group, and $S \neq 1$ is the unprivileged group. $\hat{Y} = 1$ means that the prediction is positive. Let us note that if $\hat{Y} = 1$ represents acceptance (e.g., for a job), then the condition requires the acceptance rates to be similar across groups. A higher value of this measure represents more similar rates across groups and therefore more fairness. Note that this notion relates to the "80 percent rule" in disparate impact law [54], which requires that the acceptance rate for any race, sex, or ethnic group be at least 80% of the rate for the group with the highest rate.

(2) **Demographic parity** – This measure is similar to *disparate impact*, but the difference is taken instead of the ratio [28, 44]. This measure is also commonly referred to as *statistical parity*. Formally, this measure is computed as follows:

$$\left|P[\hat{Y} = 1|S = 1] - P[\hat{Y} = 1|S \neq 1]\right| \leq \varepsilon \tag{2}$$

A lower value of this measure indicates more similar acceptance rates and therefore better fairness. *Demographic parity* (and *disparate impact*) ensure that the positive prediction is assigned to the two groups at a similar rate.

One disadvantage of these two measures is that a fully accurate classifier may be considered unfair, when the base rates (i.e., the proportion of actual positive outcomes) of the various groups are significantly different. Moreover, in order to satisfy *demographic parity*, two similar individuals may be treated differently since they belong to two different groups – such treatment is prohibited by law in some cases (note that this notion also corresponds to the practice of *affirmative action* [58]).

(3) **Equalized odds** – This measure was designed by [65] to overcome the disadvantages of measures such as *disparate impact* and *demographic parity*. The measure computes the difference between the false positive rates (FPR), and the difference between the true positive rates (TPR) of the two groups. Formally, this measure is computed as follows:

$$\left|P[\hat{Y} = 1|S = 1, Y = 0] - P[\hat{Y} = 1|S \neq 1, Y = 0]\right| \leq \varepsilon \tag{3}$$

$$\left|P[\hat{Y} = 1|S = 1, Y = 1] - P[\hat{Y} = 1|S \neq 1, Y = 1]\right| \leq \varepsilon \tag{4}$$

Where the upper formula requires the absolute difference in the FPR of the two groups to be bounded by ε, and the lower formula requires the absolute difference in the TPR of the two groups to be bounded ε. Smaller differences between groups indicate better fairness. In contrast to *demographic parity* and *disparate impact* measures, a fully accurate classifier will necessarily satisfy the two *equalized odds* constraints. Nevertheless, since *equalized odds* relies on the actual ground truth (i.e., Y), it assumes that the base rates of the two groups are representative and were not obtained in a biased manner.

One use case that demonstrates the effectiveness of this measure investigated the COMPAS [1] algorithm used in the United States criminal justice system. For predicting recidivism,

although its accuracy was similar for both groups (African-Americans and Caucasians), it was discovered that the *odds* were different. It was discovered that the system had falsely predicted future criminality (FPR) among African-Americans at twice the rate predicted for white people [6]; importantly, the algorithm also induced the opposite error, significantly underestimating future crimes among Caucasians (FNR).

(4) **Equal opportunity** – This requires true positive rates (TPRs) to be similar across groups (meaning the probability of an individual with a positive outcome to have a positive prediction) [65]. This measure is similar to equalized odds but focuses on the true positive rates only. This measure is mathematically formulated as follows:

$$\left|P[\hat{Y} = 1|S \neq 1, Y = 1] - P[\hat{Y} = 1|S = 1, Y = 1]\right| \leq \varepsilon \tag{5}$$

Let us note that following the equality in terms of only one type of error (e.g., true positives) will increase the disparity in terms of the other error [112]. Moreover, according to [38], this measure may be problematic when base rates differ between groups.

Thus far, we have mapped the most common *group* notions of fairness, which require parity of some statistical measure across groups. The literature has additionally indicated *individual* notions of fairness. It is alternatively possible to match other measures such as accuracy, error rates or calibration values between groups (see the Appendix and specifically Table 4). *Group* definitions of fairness, such as *demographic parity, disparate impact, equalized odds* and *equalized opportunity*, consider fairness with respect to the whole group, as opposed to *individual* notions of fairness.

(5) **Individual fairness** – This requires that similar individuals will be treated similarly. Similarity may be defined with respect to a particular task [44, 73]. Individual fairness may be described as follows:

$$\left|P(\hat{Y}^{(i)} = y|X^{(i)}, S^{(i)}) - P(\hat{Y}^{(j)} = y|X^{(j)}, S^{(j)})\right| \leq \varepsilon;\ if\ d(i,j) \approx 0 \tag{6}$$

where i and j denote two individuals, $S^{(\cdot)}$ refers to the individuals' sensitive attributes and $X^{(\cdot)}$ refers to their associated features. $d(i,j)$ is a distance metric between individuals that can be defined depending on the domain such that similarity is measured according to an intended task. This measure considers other individual attributes for defining fairness, rather than just the sensitive attributes. However, note that in order to define similarity between individuals, a similarity metric needs to be defined, which is not trivial. This measure, in addition to assuming a similarity metric, also requires some assumptions regarding the relationship between features and labels (see, for example, [37]).

3.3 Trade-offs

Determining the right measure to be used must take into account the proper legal, ethical, and social context. As demonstrated above, different measures exhibit different advantages and disadvantages. Next, we highlight the main trade-offs that exist between different notions of fairness, and the inherent trade-off between fairness and accuracy.

Fairness measures trade-offs

Interestingly, several recent studies have shown that it is not possible to satisfy multiple notions of fairness simultaneously [15, 36, 38, 39, 56, 85, 112]. For example, when base rates differ between

groups, it is not possible to have a classifier that equalizes both calibration and odds (except for trivial cases such as a classifier that assigns all examples to a single class). Additionally, there is also evidence for incompatibility between equalized accuracy and equalized odds, as in the COMPAS criminal justice use case [6, 15].

[112] recommends that in light of the inherent incompatibility between equalized calibration and equalized odds, practical implications requires choosing only one of these goals according to the specific application's requirements. We recommend that any selected measure of algorithmic fairness be considered in the appropriate legal, social and ethical contexts.

Table 1. Measures and Definitions for Algorithmic Fairness

Measure	Paper	Description	Type	Uses Actual Outcome	Uses Sensitive Attribute	Type of Actual Outcome	Type of Sensitive Attribute	Equivalent Notions
Disparate Impact	[54]	High ratio between positive prediction rates of both groups	Group	✗	✓	-	Binary	For $\varepsilon = 0.2$ relates to the "80 percent rule" in disparate impact law [54]
Demographic Parity	[28], [44]	Similar positive prediction rates between groups	Group	✗	✓	-	Binary	• Statistical parity [44]; • Group fairness [44]; • Equal acceptance rates [36, 136, 152]; • Discrimination score [28]
Equal Opportunity	[65]	Requires that TPRs are similar across groups	Group	✓	✓	Binary	Binary	• Equal true positive rate (TPR); • Mathematically equal TPRs will induce equal false negative rates (FNRs) (see [39, 136]); • False negative error rate balance [36, 136];
Equalized Odds	[65]	Requires that FPRs (1-TNR) and TPRs (1-FNR) are similar across groups	Group	✓	✓	Binary	Binary	• Disparate mistreatment [142]; • Error rate balance [36]; • Conditional procedure accuracy equality [15]
Fairness through Awareness	[44]	Requires that similar individuals will have similar classifications. Similarity can be defined with respect to a specific task	Individual	✗	✗	-	-	Individual fairness

Fairness-accuracy trade-off

The literature extensively discusses the inherent trade-off between accuracy and fairness - as we pursue a higher degree of fairness, we may compromise accuracy (see for example [85]). A theoretical analysis of the trade-off between fairness and accuracy was studies in [39] and [91]. Since then, many papers have empirically supported the existence of this trade-off (for example, [12, 57, 101]). Generally, the aspiration of a fairness-aware algorithm is to achieve a model that allows for higher fairness without significantly compromising the accuracy or other alternative notions of utility.

Table 1 presents a summary of the measures presented in this section. For further reading about algorithmic fairness measures, we refer the reader to [38], [85], and [136].

4 FAIRNESS-ENHANCING MECHANISMS

Numerous recent papers have proposed mechanisms to enhance fairness in machine learning algorithms. These mechanisms are typically categorized into three types: pre-process, in-process, and post-process. The following three subsections review studies in each one of these categories. The fourth subsection is devoted for comparing the three mechanism types and providing guidelines on when each type should be used.

4.1 Pre-Process Mechanisms

Mechanisms in this category involve changing the training data before feeding it into a machine learning algorithm. Preliminary mechanisms, such as the ones proposed by [76] and [95] proposed changing the labels of some instances or reweighing them before training to make the classification fairer. Typically, the labels that are changed are related to samples that are closer to the decision boundary since these are the ones that are most likely to be discriminated. More recent mechanisms suggest modifying feature representations, so that a subsequent classifier will be fairer [30, 54, 94, 122, 144].

For example, [54] suggest modifying the features in the dataset so that the distributions for both privileged and unprivileged groups become similar, and therefore, making it more difficult for the algorithm to differentiate between the two groups. A tuning parameter λ was provided for controlling the trade-off between fairness and accuracy (λ=0 indicates no fairness considerations, while λ=1 maximizes fairness). [8, 35] use the same notion of fair representation learning and applies it for fair clustering, and [122] applies it for fair dimensionality reduction (PCA). For more fair representation learning using *adversarial learning*, see section 6.2.

Note that this approach to achieving fairness is somewhat related to the field of data *compression* [131, 144]. It is also very closely related to *privacy* research since both fairness and privacy can be enhanced by removing or obfuscating the sensitive information, with the adversary goal of minimal data distortion [47, 83].

4.2 In-Process Mechanisms

These mechanisms involve modifying the machine learning algorithms to account for fairness during the training time [4, 12, 13, 28, 59, 79, 138, 142, 143].

For example, [79] suggest adding a regularization term to the objective function that penalizes the mutual information between the sensitive feature and the classifier predictions. A tuning parameter η was provided to modulate the trade-off between fairness and accuracy.

[142], [143] and [138] suggest adding constraints to the classification model that require satisfying a proxy for *equalized odds* [138, 142] or *disparate impact* [143]. [138] also show that there exist difficult computational challenges in learning a fair classifier based on *equalized odds*.

[12] and [13] suggest incorporating penalty terms into the objective function that enforce matching proxies of FPR and FNR. [77] suggest adjusting a decision tree split criterion to maximize information gain between the split attribute and the class label while minimizing information gain with respect to the sensitive attribute. [144] combine fair representation learning with an in-process model by applying a multi-objective loss function based on logistic regression, and [94] apply this notion using a variational autoencoder.

[113] suggest using the notion of *privileged learning*[1] for improving performance in cases where the sensitive information is available at training time but not at testing time. They add constraints and regularization components to the privileged learning support vector machine (SVM) model proposed by [135]. They combine the sensitive attributes as privileged information that is known only at training time, and they additionally use a maximum mean discrepancy (MMD) criterion [62] to encourage the distributions to be similar across privileged and unprivileged groups.

[14] propose a convex in-process fairness mechanism for regression tasks and use three regularization terms that include variations of individual fairness, group fairness and a combined hybrid fairness penalty term. [5] propose an in-process minimax optimization formulation for enhancing fairness in regression tasks based on the suggested design of [4] for classification tasks. They use two fairness metrics adjusted for regression tasks. One is an adjusted *demographic parity* measure, which requires the predictor to be independent of the sensitive attribute as measured by the cumulative distribution function (CDF) of the protected group compared to the CDF of the general population [5] using the *Kolmogorov-Smirnov statistic* [90]. The second measure is the bounded group loss (BGL), which requires that the prediction error of all groups remain below a predefined level [5].

4.3 Post-Process Mechanisms

These mechanisms perform post-processing of the output scores of the classifier to make decisions fairer [39, 46, 65, 101]. For example, [65] propose a technique for flipping some decisions of a classifier to enhance equalized odds or equalized opportunity. [39] and [101] similarly suggest selecting separate thresholds for each group separately, in a manner that maximizes accuracy and minimizes demographic parity. [46] propose a decoupling technique to learn a different classifier for each group. They additionally combine a *transfer learning* technique with their procedure to learn from out-of-group samples (to read more about transfer learning, see [108]).

Table 2 presents a summary of the pre-process, in-process and post-process mechanisms for algorithmic fairness discussed in this section. These methods were designed for the task of classification. Fairness mechanisms for other learning tasks are discussed in section 6.

[1] Privileged learning is designed to improve performance by using additional information, denoted as the âĂIJprivileged information,âĂİ which is present only in the training stage and not in the testing stage [135].

Table 2. Pre-Process, In-Process and Post-Process Mechanisms for Algorithmic Fairness

Paper	Mechanism Type	Base Algorithm	Optimization Measure	Evaluation Measure	Method Name	Datasets
[79]	In-Process	Logistic regression	Mutual information between prediction and sensitive attribute	Normalized prejudice index	Prejudice Remover Regularizer	Adult (test only)
[54]	Pre-Process	Any	Earth moving distance	Disparate impact	Removing Disparate Impact	• Adult • German
[143]	In-Process	Decision boundary-based	Covariance (between sensitive attributes and distance to the decision boundary)	Disparate impact	Fairness Constraints	ProPublica
[142]	In-Process	Decision boundary-based	Proxy for equalized odds	Equalized odds	Removing Disparate Mistreatment	ProPublica
[13]	In-Process	Decision boundary-based	Proxy for equalized odds	Equalized odds	Penalizing Unfairness	• ProPublica • Adult • Loans • Admissions
[77]	In-Process, Post-Process	In-Process - decision tree; Post-Process - any algorithm	Information gain	Demographic parity	• Discrimination Aware Tree Construction (new split criterion); • Relabeling (post-process)	• Adult • Communities • Dutch census
[76]	Pre-Process	Any score-based	Acceptance probabilities, distance from boundary	Demographic parity	• Massaging • Reweighing • Sampling • Suppression	• German • Adult • Communities • Dutch
[28]	In-Process, Post-Process	Naive Bayes	Acceptance probabilities	Demographic parity	• Modifying Naive Bayes • Two Naive Bayes • Expectation Maximization	Adult (test only)
[95]	Pre-Process	Any	Conditional statistical parity	Conditional statistical parity	Discrimination Prevention with KNN	• Adult • Communities • German
[144]	Pre-Process + In-Process	Logistic regression	Demographic parity	Demographic parity	Learning Fair Representations	• Adult • German • Heritage health
[59]	In-Process	SVM	Disparate impact, Equal opportunity	Disparate impact	Dataset Constraints	Adult
[65]	Post-Process	Any score-based	Equalized odds	Equalized odds	Equality of Opportunity in Supervised Learning	FICO scores [65]
[39]	Post-Process	Any score-based	• Demographic parity • Conditional statistical parity • Predictive parity	• Demographic parity • Conditional statistical parity • Predictive parity	Cost of Fairness	ProPublica
[138]	In-Process + Post-Process	Convex linear	Equalized odds	Equalized odds	Learning Non-Discriminatory Predictors	-
[113]	In-Process	SVM	Maximum mean discrepancy (MMD)	• Equalized odds • Overall accuracy equality	Recycling Privileged Learning and Distribution Matching	• ProPublica • Adult
[30]	Pre-Process	Any	• Disparate impact • Individual fairness	Disparate impact	Optimized Preprocessing for Discrimination Prevention	• ProPublica • Adult
[46]	In-Process + Post-Process	Any score-based	• Demographic parity • Equalized odds	• Demographic parity • Equalized odds	Decoupled Classifiers	ImageNet [42]
[101]	Post-Process	Any score-based	• Demographic parity • Equal opportunity	• Demographic parity • Equal opportunity	Plugin Approach	-
[4]	In-Process	Any	• Demographic parity • Equalized odds	• Demographic parity • Equalized odds	Reductions Approach	• ProPublica • Adult • Dutch • Admissions
[94]	Pre-Process + In-Process	Any	Maximum mean discrepancy (MMD)	• Demographic parity • Mean difference	Variational Fair Autoencoder	• Adult • German • Heritage health

4.4 Which Mechanism to Use?

The different mechanism types present respective advantages and disadvantages. Pre-process mechanisms can be advantageous since they can be used with any classification algorithm. However, they may harm the explainability of the results. Moreover, since they are not tailored for a specific classification algorithm, there is high uncertainty with regard to the level of accuracy obtained at the end of the process.

Similar to pre-process mechanisms, post-process mechanisms may be used with any classification algorithm. However, due to the relatively late stage in the learning process in which they are applied, post-process mechanisms typically obtain inferior results [138]. In a post-process mechanism, it may be easier to fully remove bias types such as *disparate impact*; however, this is not always the desired measure, and it could be considered as discriminatory since it deliberately damages accuracy for some individuals in order to compensate others (this is also related to the controversies in the legal and economical field of *affirmative action*, see [58]). Specifically, post-process mechanisms may treat differently two individuals who are similar across all features except for the group to which they belong. This approach requires the decision maker at the end of the loop to possess the information of the group to which individuals belong (this information may be unavailable due to legal or privacy reasons).

In-process mechanisms are beneficial since they can explicitly impose the required trade-off between accuracy and fairness in the objective function [138]. However, such mechanisms are tightly coupled with the machine algorithm itself.

Hence, we see that the selection of method depends on the availability of the ground truth, the availability of the sensitive attributes at test time, and on the desired definition of fairness, which can also vary from one application to another.

Several preliminary attempts were made in order to understand which methods are best for use. The study in [64] was a first effort in comparing several fairness mechanisms previously proposed in the literature [28, 54, 79, 143]. The analysis focuses on binary classification with binary sensitive attributes. The authors have demonstrated that the performances of the methods vary across datasets, and there was no conclusively dominating method.

Another study by [116] has shown as a preliminary benchmark that in several cases, in-process mechanisms perform better than pre-process mechanisms, and for other cases, they do not, leading to the conclusion that there is a need for much more extensive experiments.

A recent empirical study [57] has provided a benchmark analysis of several fairness-aware methods and compared the fairness-accuracy trade-offs obtained by these methods. The authors have tested the performances of these methods across different measures of fairness and across different datasets. They have concluded that there was no single method that outperformed the others in all cases and that the results depend on the fairness measure, on the dataset, and on changes in the train-test splits.

More research is required for developing robust fairness mechanisms and metrics or, alternatively, for finding the adequate mechanism and metric for each scenario. For instance, the conclusions reached when considering missing data might be very different than those reached when all information is available [74, 99]. [74] explore the limitations of measuring fairness when the membership in a protected group is not available in the data. [99] have tested imputation strategies to deal with the fairness of partially missing examples in the dataset. They have shown that rows containing missing values may be more fair than the rest and therefore suggest imputation rather than deletion of these data. [110] find that when there is an evident selection bias in the

data, meaning that there is an extreme under-representation of unprivileged groups, pre-process mechanisms can outperform in-process mechanisms.

5 FAIRNESS-RELATED DATASETS

In this section, we review the most commonly used datasets in the literature of algorithmic fairness.

ProPublica risk assessment dataset

The ProPublica dataset includes data from the COMPAS risk assessment system (see [1, 6, 88]).

This dataset was previously extensively used for fairness analysis in the field of criminal justice risk [15]. The dataset includes 6,167 individuals, and the features in the dataset include number of previous felonies, charge degree, age, race and gender. The target variable indicates whether an inmate recidivated (was arrested again) within two years after release from prison.

As for the sensitive variable, this dataset was previously used with two variations – the first when race was considered as the sensitive attribute and the second when gender was considered as the sensitive attribute [13, 30, 52, 57, 99].

Adult income dataset

The Adult dataset is a publicly available dataset in the UCI repository [43] based on 1994 US census data. The goal of this dataset is to successfully predict whether an individual earns more or less than 50,000$ per year based on features such as occupation, marital status, and education. The sensitive attributes in this dataset includes age [94], gender [144] and race [57, 99, 143].

This dataset is used with several different preprocessing procedures. For example, the dataset of [143] includes 45,222 individuals after preprocessing (48,842 before preprocessing).

German credit dataset

The German dataset is a publicly available dataset in the UCI repository [43] that includes information of individuals from a German bank in 1994.

The goal of this dataset is to predict whether an individual should receive a good or bad credit risk score based on features such as employment, housing, savings, and age. The sensitive attributes in this dataset include gender [57, 94] and age [75, 144]. This dataset is significantly smaller, with only 1,000 individuals with 20 attributes.

Ricci promotion dataset

The Ricci dataset includes the results of an exam administered to 118 individuals to determine which of them would receive a promotion. The dataset originated from a case that was brought to the United States Supreme Court [102, 120]. The goal of this dataset is to successfully predict whether an individual receives a promotion based on features that were tested in the exam, as well as the current position of each individual. The sensitive attribute in this dataset is race.

Mexican poverty dataset

The Mexican poverty dataset includes poverty estimation for determining whether to match households with social programs. The data originated from a survey of 70,305 households in 2016 [71]. The target feature is poverty level, and there are 183 features. This dataset was studied, for example, in [106]. The authors studied two sensitive features: young and old families; urban and rural areas.

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

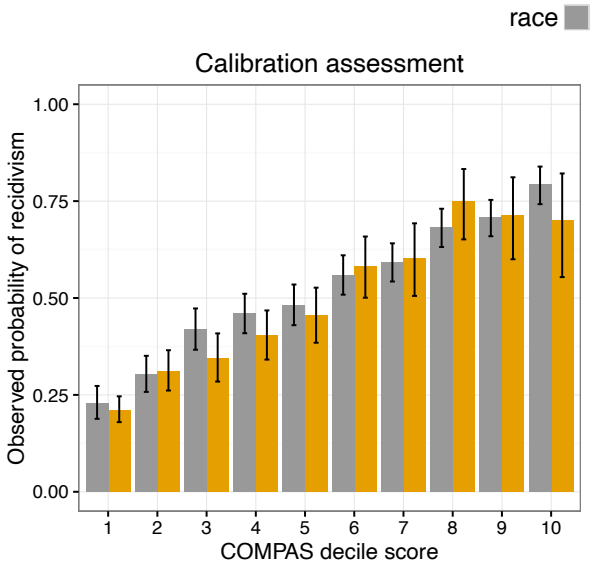
In this section we present our first main result, which establishes that predictive parity is incompatible with error rate balance when prevalence differs across groups. To better motivate the discussion, we begin by presenting an empirical fairness assessment of the COMPAS RPI. Figure 1 shows plots 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 $s_{\text{HR}} = 4$ for their analysis, which some critics have argued is too low, suggesting that $s_{\text{HR}} = 7$ 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]

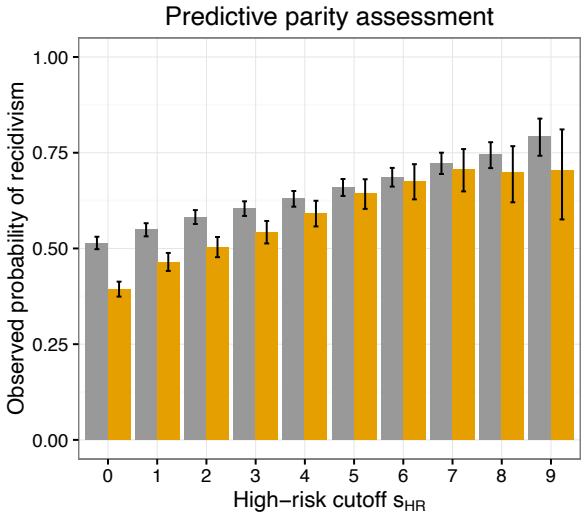
As we now proceed to show, the error rate imbalance exhibited by COMPAS is not a coincidence, nor can it be remedied in the present context. When the recidivism prevalence–i.e., the base rate $\mathbb{P}(Y = 1 \mid R = r)$—differs across groups, any instrument that satisfies predictive parity at a given 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

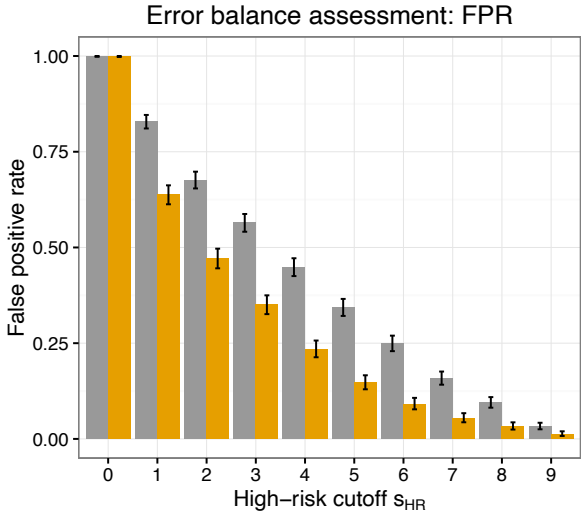
the values (or the distribution of values) in this table. Another constraint—one that we have no direct control over—is imposed by the recidivism prevalence within groups. It is not difficult to



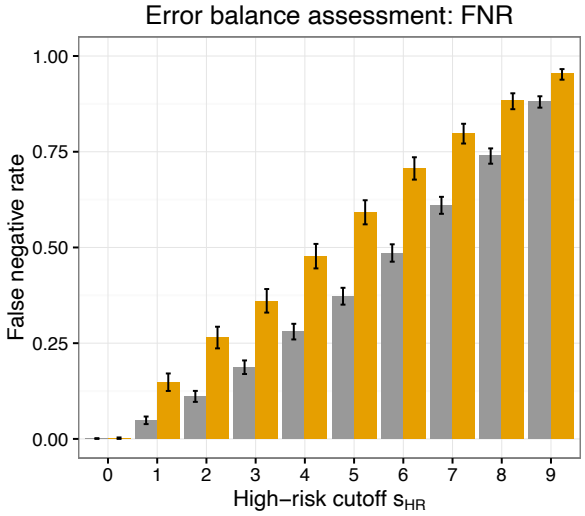
(a) Bars represent empirical estimates of the expressions in (2.1): $\mathbb{P}(Y = 1 \mid S = s, R = r)$ for decile scores $s \in \{1, \ldots, 10\}$.



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



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



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

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

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

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

Week 3: Fairness and Causality

RUTH FREMSON/NYT/REDUX/EYEVINE

A migrant farm worker has her fingerprints scanned so that she can register for a national identity card in India.

The long road to fairer algorithms

Matt J. Kusner & Joshua R. Loftus

Build models that identify and mitigate the causes of discrimination.

An algorithm deployed across the United States is now known to underestimate the health needs of black patients[1]. The algorithm uses health-care costs as a proxy for health needs. But black patients' health-care costs have historically been lower because systemic racism has impeded their access to treatment – not because they are healthier.

This example illustrates how machine learning and artificial intelligence can maintain and amplify inequity. Most algorithms exploit crude correlations in data. Yet these correlations are often by-products of more salient social relationships (in the health-care example, treatment that is inaccessible is, by definition, cheaper), or chance occurrences that will not replicate.

To identify and mitigate discriminatory relationships in data, we need models that capture or account for the causal pathways that give rise to them. Here we outline what is required to build models that would allow us to explore ethical issues underlying seemingly objective analyses. Only by unearthing the true causes of discrimination can we build algorithms that correct for these.

Causal models

Models that account for causal pathways have three advantages. These 'causal models' are: tailored to the data at hand; allow us to account for quantities that aren't observed; and address shortcomings in current concepts of fairness (see 'Fairness four ways').

A causal model[2] represents how data are generated, and how variables might change in response to interventions. This can be shown as a graph in which each variable is a node and arrows represent the causal connections between them. Take, for example, a data set about who gets a visa to work in a country. There is information about the country each person comes from, the work they do, their religion and whether or not they obtained a visa (see 'Three causal tests', part 1).

This model says that the country of origin directly influences a person's religion and whether they obtain a visa; so, too, do religion and type of work. Having a causal model allows us to address questions related to ethics, such as does religion influence the visa process?

But because many different causal models could have led to a particular observed data set, it is not generally possible to identify the right causal model from that data set alone[3]. For example, without any extra assumptions, data generated from the causal graph described here could seem identical to those from a graph in which religion is no longer linked to visa granting. A modeller must therefore also leverage experiments and expert knowledge, and probe assumptions.

Experiments can help in identifying factors that affect fairness. For example, a modeller wishing to explore whether ethnicity would affect treatment recommendations made online by health-care professionals could create two patient profiles that differ only in some respect that relates to ethnicity. For instance, one profile could have a name common to Americans of Chinese descent, and the other a name common to Americans of African descent. If the treatment recommendations are the same, then names can be ruled out as a source of bias, and the model can be stress-tested in another way.

Few aspects of a deep, multifaceted concept can be tested as easily as changing a name. This means that experimental evidence can underestimate the effects of discrimination. Integration of expert knowledge, particularly

from the social sciences and including qualitative methods, can help to overcome such limitations. This knowledge can be used to, for example, inform the modeller of variables that might be influential but unobserved (lighter circles in 'Three causal tests'), or to determine where to put arrows.

Assumptions about unobserved variables that might alter the predictions of a model need to be clearly stated. This is particularly important when experiments cannot be run or more detailed expert knowledge is not available. For example, if 'health-care access' is not observed in a model attempting to predict 'health need', then it is crucial to identify any potential impacts it might have on 'health costs' as well as how it is affected by 'ethnicity'.

This need for context and metadata makes causal models harder to build than non-causal ones. It can also make them a more powerful way to explore ethical questions.

Three tests

Causal models can test the fairness of predictive algorithms in three ways.

Counterfactuals. A causal model allows us to ask and answer questions such as 'Had the past been different, would the present or future have changed?' In the visa example (see 'Three causal tests', part 1), algorithmic biases could be smoked out by tweaking parts of the model to explore, for instance: 'Had individual X been Christian, would this algorithm have granted them a visa?' A researcher could then identify what pieces of information an algorithm could use to achieve counterfactual fairness[4]: the algorithm's output would not change regardless of the individual's religion. For example, if the algorithm used just work and not country of origin or religion, it would satisfy counterfactual fairness.

Sensitivity. In many settings, unknowns alter knowns – data we can observe are influenced by data we cannot. Consider a causal model for a trial setting (see 'Three causal tests', part 2).

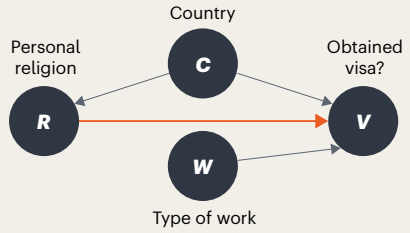
This model shows how two independent sets of unobserved quantities, structural racism and jury racism, can unfairly lead to a guilty verdict. Although researchers often cannot precisely identify unobserved variables, they can reason about how sensitive a model is to them. For instance, they can explore how sensitive our estimate of the causal link between legal representation and guilty verdict is to different levels of jury racism. Simulations of the worst-case bias scenarios (that is, when jury racism is highest) can then be used to alter jury selection to minimize the bias.

Impacts. Data-driven decisions can have long-term consequences and spillover effects. These effects might not be obvious, especially in the standard machine-learning paradigm of predicting one short-term outcome. But carefully designed causal models can help researchers to use 'interventions' to probe the ripple effects of decisions far into the future[5,6]. For instance, the models can help regulatory agencies to understand how changing a scholarship algorithm influences who is accepted into law school (see 'Three causal tests', part 3). In this example, a single parent might need a scholarship so that they can reduce the hours they need to spend at a job, leaving them more time for study. That boosts their grades and therefore influences their chances of being admitted to law school. This complex chain can be explored using causal models.

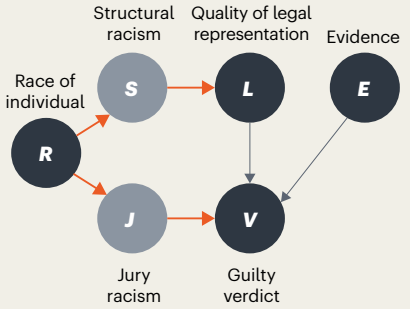
THREE CAUSAL TESTS

Algorithmic fairness can be examined in different ways.

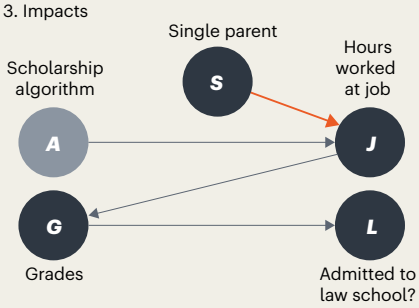
1. Counterfactuals



2. Sensitivity



3. Impacts



Five steps

Causal models are powerful tools, but they must be used appropriately. They are only models, and will thus fail to capture important aspects of the real world. Here we offer some guidelines on using them wisely.

Collaborate across fields. Researchers in statistics and machine learning need to know more about the causes of unfairness in society. They should work closely with those in disciplines such as law, social sciences and the humanities. This will help them to incorporate the context of the data used to train the algorithms. For example, scholars should meet at interdisciplinary workshops and conferences. One such is next year's Association for Computing and Machinery (ACM) conference on Fairness, Accountability and Transparency to derive a set of causal models for setting bail price and for immigration decisions.

A great example of such collaborations is one between information scientist Solon Barocas at Cornell University in Ithaca, New York, and attorney Andrew Selbst at the Data & Society Research Institute in New York City. They described how current law is unable to deal with algorithmic bias[7]. Partly in response to this work, machine-learning researchers have launched a large subfield, known as algorithmic fairness, that looks into ways of removing bias from data. And we and other researchers now use causal models to quantify discrimination due to data.

Partner with stakeholders. Predictive algorithms should be developed with people they are likely to affect. Stakeholders are best placed to provide input on difficult ethical questions and historical context. One example is the work by statistician Kristian Lum at the Human Rights Data Analysis Group in San Francisco, California, which investigates criminal-justice algorithms[8]. Such algorithms decide whether to detain or release arrested individuals and how high to set their bail, yet they are known to be biased. Lum has invited people affected by such decisions to speak at academic conferences attended by people who research these algorithms. This has led to closer collaboration, including the tutorial 'Understanding the context and consequences of pre-trial detention' presented at the 2018 ACM conference on Fairness, Accountability and Transparency in New York. So far, most stakeholder work has focused on criminal justice. Another setting that would benefit from it is mortgage lending.

We propose that a rotating, interdisciplinary panel of stakeholders investigates the impacts of algorithmic decisions, for example as part of a new international regulatory institute.

Make the workforce equitable. Women and people from minority groups are under-represented in the fields of statistics and machine learning. This directly contributes to the creation of unfair algorithms. For example, if facial detection software struggles to detect faces of black people[9], it is likely the algorithm was trained largely on data representing white people. Initiatives such as Black in AI (go.nature.com/38pbcaa) or Women in Machine Learning

Fairness four ways

A flurry of work has conceptualized fairness. Here are some of the most popular, and ways in which causal models offer alternatives.

Fairness through unawareness[12]. This method works by removing any data that are considered *prima facie* to be unfair. For example, for an algorithm used by judges making parole decisions, fairness through unawareness could dictate that data on ethnic origin should be removed when training this algorithm, whereas data on the number of previous offences can be used. But most data are biased. For instance, number of previous offences can bear the stamp of historical racial bias in policing, as can the use of plea bargaining (pleading guilty being more likely to reduce a sentence than arguing innocence)[13]. This can leave researchers with a hard choice: either remove all data or keep biased data.

Alternatively, causal models can directly quantify how data are biased.

Demographic parity[14]. A predictive algorithm satisfies demographic parity if, on average, it gives the same predictions to different groups. For example, a university-admissions algorithm would satisfy demographic parity for gender if 50% of its offers went to women and 50% to men. It is currently more common in law to relax demographic parity so that predictions aren't necessarily equal, but are not too imbalanced. Specifically, the US Equal Employment Opportunity Commission states that fair employment should satisfy the 80% rule: the acceptance rate for any group should be no less than 80% of that of the highest-accepted group. For instance, if 25% of women were offered jobs, and this is the highest acceptance rate, then at least 20% of men must be offered jobs[4]. One criticism of demographic parity is that it might not make sense to use it in certain settings, such as a fair arrest rate for violent crimes (men are significantly more likely to commit acts of violence)[15].

Instead, one could require that counterfactual versions of the same individual should get the same prediction[4].

Equality of opportunity[16]. This is the principle of giving the same beneficial predictions to individuals in each group. Consider a predictive algorithm that grants loans only to individuals who have paid back previous loans. It satisfies 'disability-based equality of opportunity' if it grants loans to the same percentage of individuals who both pay back and have a disability as it does to those who pay back and who do not have a disability. However, being able to pay back a loan in the first place can be affected by bias: discriminatory employers might be less likely to hire a person with a disability, which can make it harder for that person to pay back a loan. This societal unfairness is not captured by equality of opportunity.

A causal model could be used to quantify the bias and estimate an unbiased version of loan repayment.

Individual fairness[17]. This concept states that similar individuals should get similar predictions. If two people are alike except for their sexual orientation, say, an algorithm that displays job advertisements should display the same jobs to both. The main issue with this concept is how to define similar. In this example, training data will probably have been distorted by the fact that one in five individuals from sexual or gender minorities report discrimination against them in hiring, promotions and pay[18]. Thus similarity is hard to define, which makes individual fairness hard to use in practice.

In causal modelling, counterfactuals offer a natural way to define a similar individual. **M.J.K. & J.R.L.**

(go.nature.com/2s5km5g) are positive steps.

And we can go further. Causal models can themselves help to address the field's 'pipeline problem' by identifying where unfairness enters the process and which interventions can increase the participation of under-represented groups without shifting the burden to extra work for role models in those groups. Academic institutions should critically evaluate and use these models for fairer admissions in fields related to artificial intelligence.

Identify when algorithms are inappropriate. Statistics and machine learning are not all-powerful. Some problems should not be solved by expanding data-gathering capabilities and automating decisions. For example, a more accurate model for predictive policing won't solve many of the ethical concerns related to the criminal legal system. In fact, these methods can mask structural issues, including the fact that many neighbourhoods are policed by people who do not live in them[10]. This disconnect means that police officers might not be invested in the community they police or the people they arrest.

There are red flags when demographics, such as ethnic origin, influence nearly every piece of information in a causal graph, or when previous attempts to address a bias failed because people strategically changed behaviours in response. In these cases, an algorithmic solution would paper over a system that needs fundamental change.

Foment criticism. A vibrant culture of feedback is essential. Researchers need to continually question their models, evaluation techniques and assumptions. Useful as causal models are, they should be scrutinized intensely: bad models can make discrimination worse[11]. At the very least, a scientist should check whether a model has the right data to make causal claims, and how much these claims would change when the assumptions are relaxed.

Algorithms are increasingly used to make potentially life-changing decisions about people. By using causal models to formalize our understanding of discrimination, we must build these algorithms to respect the ethical standards required of human decision makers.

The authors

Matt J. Kusner is an associate professor in the Department of Computer Science at University College London, and a fellow at the Alan Turning Institute, London, UK. **Joshua R. Loftus** is an assistant professor in the Department of Technology, Operations, and Statistics at New York University, New York, USA.
e-mails: matt.kusner@gmail.com; loftus@nyu.edu

1. Obermeyer, Z. *et al. Science* **366**, 447–453 (2019).
2. Pearl, J. *Causality: Models, Reasoning, and Inference* (Cambridge Univ. Press, 2000).
3. Spirtes, P. *et al. Causation, Prediction, and Search* (MIT Press, 2000).
4. Kusner, M. J., Loftus, J., Russell, C. & Silva, R. In *Advances in Neural Information Processing Systems* 4066–4076 (MIT Press, 2017).
5. Liu, L. T. *et al.* In *International Conference on Machine Learning* 3150–3158 (ACM, 2018).
6. Kusner, M., Russell, C., Loftus, J. & Silva, R. *Proc. Machine Learning Res.* **97**, 3591–3600 (2019).
7. Barocas, S. & Selbst, A. D. *Calif. L. Rev.* **104**, 671 (2016).
8. Lum, K. *Nature Hum. Behav.* **1**, 0141 (2017).
9. Simon, M. 'HP looking into claim webcams can't see black people.' (CNN Tech, 23 December 2009).
10. McManus, H. D. *et al. Race Justice* https://doi.org/10.1177/2153368719849486 (2019).
11. Kilbertus, N. *et al.* 'The Sensitivity of Counterfactual Fairness to Unmeasured Confounding'. In *Uncertainty in Artificial Intelligence* (AUAI, 2019).
12. Grgic-Hlaca, N. *et al.* 'The case for process fairness in learning: Feature selection for fair decision making.' *NeurIPS Symposium on Machine Learning and the Law* (2016).
13. Wilford, M. M. & Khairalla, A. in *Social Sciences Contributions to the Real Legal System* Ch. 7, 132 (2019).
14. Zafar, M. B., Valera, I., Rogriguez, M. G. & Gummadi, K. P. In *Artificial Intelligence and Statistics* 962–970 (2017).
15. Dobash, R. E., Dobash, R. P., Cavanagh, K. & Lewis, R. *Violence Against Women* **10**, 577–605 (2004).
16. Hardt, M., Price, E. & Srebro, N. 'Equality of opportunity in supervised learning'. In *Advances in Neural Information Processing Systems* 3315–3323 (2016).
17. Dwork, C. *et al.* 'Fairness through awareness'. In *Proc. 3rd Innov. Theoret. Comp. Sci. Conf.* 214–226 (2012).
18. Pizer, J. C. *et al. Loy. LAL Rev.* **45**, 715 (2011).

RESEARCH ARTICLE

ECONOMICS

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and affecting millions of patients, exhibits significant racial bias: At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise. We suggest that the choice of convenient, seemingly effective proxies for ground truth can be an important source of algorithmic bias in many contexts.

There is growing concern that algorithms may reproduce racial and gender disparities via the people building them or through the data used to train them (*1*–*3*). Empirical work is increasingly lending support to these concerns. For example, job search ads for highly paid positions are less likely to be presented to women (*4*), searches for distinctively Black-sounding names are more likely to trigger ads for arrest records (*5*), and image searches for professions such as CEO produce fewer images of women (*6*). Facial recognition systems increasingly used in law enforcement perform worse on recognizing faces of women and Black individuals (*7*, *8*), and natural language processing algorithms encode language in gendered ways (*9*).

Empirical investigations of algorithmic bias, though, have been hindered by a key constraint: Algorithms deployed on large scales are typically proprietary, making it difficult for independent researchers to dissect them. Instead, researchers must work "from the outside," often with great ingenuity, and resort to clever workarounds such as audit studies. Such efforts can document disparities, but understanding how and why they arise—much less figuring out what to do about them—is difficult without greater access to the algorithms themselves. Our understanding of a mechanism therefore typically relies on theory or exercises with researcher-created algorithms (*10*–*13*). Without an algorithm's training data, objective function, and prediction methodology, we can only guess as to the actual mechanisms for the important algorithmic disparities that arise.

In this study, we exploit a rich dataset that provides insight into a live, scaled algorithm deployed nationwide today. It is one of the largest and most typical examples of a class of commercial risk-prediction tools that, by industry estimates, are applied to roughly 200 million people in the United States each year. Large health systems and payers rely on this algorithm to target patients for "high-risk care management" programs. These programs seek to improve the care of patients with complex health needs by providing additional resources, including greater attention from trained providers, to help ensure that care is well coordinated. Most health systems use these programs as the cornerstone of population health management efforts, and they are widely considered effective at improving outcomes and satisfaction while reducing costs (*14*–*17*). Because the programs are themselves expensive—with costs going toward teams of dedicated nurses, extra primary care appointment slots, and other scarce resources—health systems rely extensively on algorithms to identify patients who will benefit the most (*18*, *19*).

Identifying patients who will derive the greatest benefit from these programs is a challenging causal inference problem that requires estimation of individual treatment effects. To solve this problem, health systems make a key assumption: Those with the greatest care needs will benefit the most from the program. Under this assumption, the targeting problem becomes a pure prediction policy problem (*20*). Developers then build algorithms that rely on past data to build a predictor of future health care needs.

Our dataset describes one such typical algorithm. It contains both the algorithm's predictions as well as the data needed to understand its inner workings: that is, the underlying ingredients used to form the algorithm (data, objective function, etc.) and links to a rich set of outcome data. Because we have the inputs, outputs, and eventual outcomes, our data allow us a rare opportunity to quantify racial disparities in algorithms and isolate the mechanisms by which they arise. It should be emphasized that this algorithm is not unique. Rather, it is emblematic of a generalized approach to risk prediction in the health sector, widely adopted by a range of for- and non-profit medical centers and governmental agencies (*21*).

Our analysis has implications beyond what we learn about this particular algorithm. First, the specific problem solved by this algorithm has analogies in many other sectors: The predicted risk of some future outcome (in our case, health care needs) is widely used to target policy interventions under the assumption that the treatment effect is monotonic in that risk, and the methods used to build the algorithm are standard. Mechanisms of bias uncovered in this study likely operate elsewhere. Second, even beyond our particular finding, we hope that this exercise illustrates the importance, and the large opportunity, of studying algorithmic bias in health care, not just as a model system but also in its own right. By any standard—e.g., number of lives affected, life-and-death consequences of the decision—health is one of the most important and widespread social sectors in which algorithms are already used at scale today, unbeknownst to many.

Data and analytic strategy

Working with a large academic hospital, we identified all primary care patients enrolled in risk-based contracts from 2013 to 2015. Our primary interest was in studying differences between White and Black patients. We formed race categories by using hospital records, which are based on patient self-reporting. Any patient who identified as Black was considered to be Black for the purpose of this analysis. Of the remaining patients, those who self-identified as races other than White (e.g., Hispanic) were so considered (data on these patients are presented in table S1 and fig. S1 in the supplementary materials). We considered all remaining patients to be White. This approach allowed us to study one particular racial difference of social and historical interest between patients who self-identified as Black and patients who self-identified as White without another race or ethnicity; it has the disadvantage of not allowing for the study of intersectional racial

[1]School of Public Health, University of California, Berkeley, Berkeley, CA, USA. [2]Department of Emergency Medicine, Brigham and Women's Hospital, Boston, MA, USA. [3]Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA. [4]Mongan Institute Health Policy Center, Massachusetts General Hospital, Boston, MA, USA. [5]Booth School of Business, University of Chicago, Chicago, IL, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: sendhil.mullainathan@chicagobooth.edu

and ethnic identities. Our main sample thus consisted of (i) 6079 patients who self-identified as Black and (ii) 43,539 patients who self-identified as White without another race or ethnicity, whom we observed over 11,929 and 88,080 patient-years, respectively (1 patient-year represents data collected for an individual patient in a calendar year). The sample was 71.2% enrolled in commercial insurance and 28.8% in Medicare; on average, 50.9 years old; and 63% female (Table 1).

For these patients, we obtained algorithmic risk scores generated for each patient-year. In the health system we studied, risk scores are generated for each patient during the enrollment period for the system's care management program. Patients above the 97th percentile are automatically identified for enrollment in the program. Those above the 55th percentile are referred to their primary care physician, who is provided with contextual data about the patients and asked to consider whether they would benefit from program enrollment.

Many existing metrics of algorithmic bias may apply to this scenario. Some definitions focus on calibration [i.e., whether the realized value of some variable of interest Y matches the risk score R (*2*, *22*, *23*)]; others on statistical parity of some decision D influenced by the algorithm (*10*); and still others on balance of average predictions, conditional on the realized outcome (*22*). Given this multiplicity and the growing recognition that not all conditions can be simultaneously satisfied (*3*, *10*, *22*), we focus on metrics most relevant to the real-world use of the algorithm, which are related to calibration bias [formally, comparing Blacks B and Whites W, $E[Y|R, W] = E[Y|R, B]$ indicates the absence of bias (here, E is the expectation operator)]. The algorithm's stated goal is to predict complex health needs for the purpose of targeting an intervention that manages those needs. Thus, we compare the algorithmic risk score for patient i in year t ($R_{i,t}$), formed on the basis of claims data $X_{i,(t-1)}$ from the prior year, to data on patients' realized health $H_{i,t}$, assessing how well the algorithmic risk score is calibrated across race for health outcomes $H_{i,t}$. We also ask how well the algorithm is calibrated for costs $C_{i,t}$.

To measure H, we link predictions to a wide range of outcomes in electronic health record data, including all diagnoses (in the form of International Classification of Diseases codes) as well as key quantitative laboratory studies and vital signs capturing the severity of chronic illnesses. To measure C, we link predictions to insurance claims data on utilization, including outpatient and emergency visits, hospitalizations, and health care costs. These data, and the rationale for the specific measures of H used in this study, are described in more detail in the supplementary materials.

Health disparities conditional on risk score

We begin by calculating an overall measure of health status, the number of active chronic conditions [or "comorbidity score," a metric used extensively in medical research (*24*) to provide a comprehensive view of a patient's health (*25*)] by race, conditional on algorithmic risk score. Fig. 1A shows that, at the same level of algorithm-predicted risk, Blacks have significantly more illness burden than Whites. We can quantify these differences by choosing one point on the x axis that corresponds to a very-high-risk group (e.g., patients at the 97th percentile of risk score, at which patients are auto-identified for program enrollment), where Blacks have 26.3% more chronic illnesses than Whites (4.8 versus 3.8 distinct conditions; $P < 0.001$).

What do these prediction differences mean for patients? Algorithm scores are a key input to decisions about future enrollment in a care coordination program. So as we might expect, with less-healthy Blacks scored at similar risk scores to more-healthy Whites, we find evidence

Table 1. Descriptive statistics on our sample, by race. BP, blood pressure; LDL, low-density lipoprotein.

	White	Black
n (patient-years)	88,080	11,929
n (patients)	43,539	6079
Demographics		
Age	51.3	48.6
Female (%)	62	69
Care management program		
Algorithm score (percentile)	50	52
Race composition of program (%)	81.8	18.2
Care utilization		
Actual cost	$7540	$8442
Hospitalizations	0.09	0.13
Hospital days	0.50	0.78
Emergency visits	0.19	0.35
Outpatient visits	4.94	4.31
Mean biomarker values		
HbA1c (%)	5.9	6.4
Systolic BP (mmHg)	126.6	130.3
Diastolic BP (mmHg)	75.5	75.7
Creatinine (mg/dl)	0.89	0.98
Hematocrit (%)	40.7	37.8
LDL (mg/dl)	103.4	103.0
Active chronic illnesses (comorbidities)		
Total number of active illnesses	1.20	1.90
Hypertension	0.29	0.44
Diabetes, uncomplicated	0.08	0.22
Arrythmia	0.09	0.08
Hypothyroid	0.09	0.05
Obesity	0.07	0.18
Pulmonary disease	0.07	0.11
Cancer	0.07	0.06
Depression	0.06	0.08
Anemia	0.05	0.10
Arthritis	0.04	0.04
Renal failure	0.03	0.07
Electrolyte disorder	0.03	0.05
Heart failure	0.03	0.05
Psychosis	0.03	0.05
Valvular disease	0.03	0.02
Stroke	0.02	0.03
Peripheral vascular disease	0.02	0.02
Diabetes, complicated	0.02	0.07
Heart attack	0.01	0.02
Liver disease	0.01	0.02

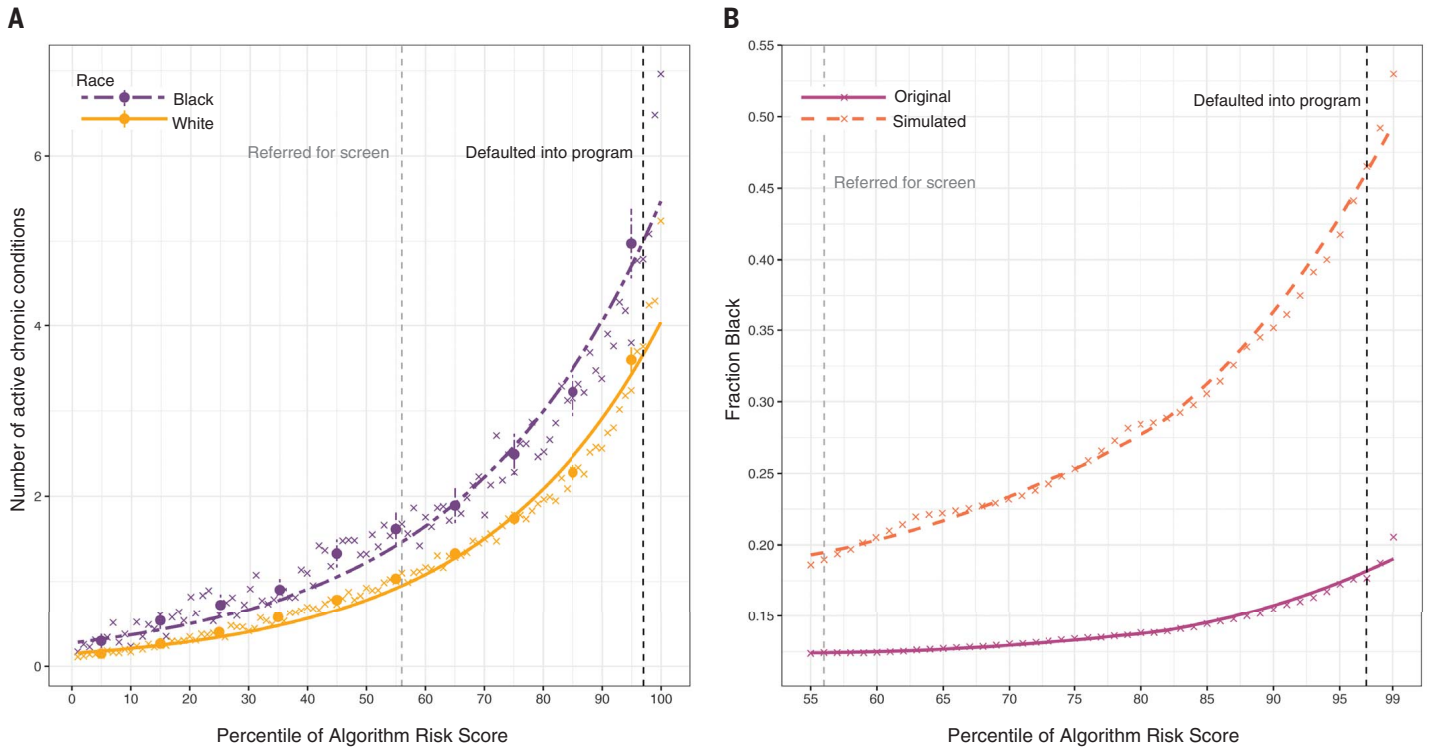


Fig. 1. Number of chronic illnesses versus algorithm-predicted risk, by race. (**A**) Mean number of chronic conditions by race, plotted against algorithm risk score. (**B**) Fraction of Black patients at or above a given risk score for the original algorithm ("original") and for a simulated scenario that removes algorithmic bias ("simulated": at each threshold of risk, defined at a given percentile on the *x* axis, healthier Whites above the threshold are replaced with less healthy Blacks below the threshold, until the marginal patient is equally healthy). The × symbols show risk percentiles by race; circles show risk deciles with 95% confidence intervals clustered by patient. The dashed vertical lines show the auto-identification threshold (the black line, which denotes the 97th percentile) and the screening threshold (the gray line, which denotes the 55th percentile).

of substantial disparities in program screening. We quantify this by simulating a counterfactual world with no gap in health conditional on risk. Specifically, at some risk threshold α, we identify the supramarginal White patient (i) with $R_i > \alpha$ and compare this patient's health to that of the inframarginal Black patient (j) with $R_j < \alpha$. If $H_i > H_j$, as measured by number of chronic medical conditions, we replace the (healthier, but supramarginal) White patient with the (sicker, but inframarginal) Black patient. We repeat this procedure until $H_i = H_j$, to simulate an algorithm with no predictive gap between Blacks and Whites. Fig. 1B shows the results: At all risk thresholds α above the 50th percentile, this procedure would increase the fraction of Black patients. For example, at α = 97th percentile, among those auto-identified for the program, the fraction of Black patients would rise from 17.7 to 46.5%.

We then turn to a more multidimensional picture of the complexity and severity of patients' health status, as measured by biomarkers that index the severity of the most common chronic illnesses in our sample (as shown in Table 1). This allows us to identify patients who might derive a great deal of benefit from care management programs—e.g., patients with severe diabetes who are at risk of catastrophic complications if they do not lower their blood sugar (*18*, *26*). (The materials and methods section describes several experiments to rule out a large effect of the program on these health measures in year t; had there been such an effect, we could not easily use the measures to assess the accuracy of the algorithm's predictions on health, because the program is allocated as a function of algorithm score.) Across all of these important markers of health needs—severity of diabetes, high blood pressure, renal failure, cholesterol, and anemia—we find that Blacks are substantially less healthy than Whites at any level of algorithm predictions, as shown in Fig. 2. Blacks have more-severe hypertension, diabetes, renal failure, and anemia, and higher cholesterol. The magnitudes of these differences are large: For example, differences in severity of hypertension (systolic pressure: 5.7 mmHg) and diabetes [glycated hemoglobin (HbA1c): 0.6%] imply differences in all-cause mortality of 7.6% (*27*) and 30% (*28*), respectively, calculated using data from clinical trials and longitudinal studies.

Mechanism of bias

An unusual aspect of our dataset is that we observe the algorithm's inputs and outputs as well as its objective function, providing us a unique window into the mechanisms by which bias arises. In our setting, the algorithm takes in a large set of raw insurance claims data $X_{i,t-1}$ (features) over the year t – 1: demographics (e.g., age, sex), insurance type, diagnosis and procedure codes, medications, and detailed costs. Notably, the algorithm specifically excludes race.

The algorithm uses these data to predict $Y_{i,t}$ (i.e., the label). In this instance, the algorithm takes total medical expenditures (for simplicity, we denote "costs" C_t) in year t as the label. Thus, the algorithm's prediction on health needs is, in fact, a prediction on health costs.

As a first check on this potential mechanism of bias, we calculate the distribution of realized costs C versus predicted costs R. By this metric, one could call the algorithm unbiased. Fig. 3A shows that, at every level of algorithm-predicted risk, Blacks and Whites have (roughly) the same costs the following year. In other words, the algorithm's predictions are well calibrated across races. For example, at the median risk score, Black patients had costs of $5147 versus $4995 for Whites (U.S. dollars); in the top 5% of algorithm-predicted risk, costs were $35,541 for Blacks versus $34,059 for Whites.

Because these programs are used to target patients with high costs, these results are largely inconsistent with algorithmic bias, as measured by calibration: Conditional on risk score, predictions do not favor Whites or Blacks anywhere in the risk distribution.

To summarize, we find substantial disparities in health conditional on risk but little disparity in costs. On the one hand, this is surprising: Health care costs and health needs are highly correlated, as sicker patients need and receive more care, on average. On the other hand, there are many opportunities for a wedge to creep in between needing health care and receiving health care—and crucially, we find that wedge to be correlated with race, as shown in Fig. 3B. At a given level of health (again measured by number of chronic illnesses), Blacks generate lower costs than Whites—on average, $1801 less per year, holding constant the number of chronic illnesses (or $1144 less, if we instead hold constant the specific individual illnesses that contribute to the sum). Table S2 also shows that Black patients generate very different kinds of costs: for example, fewer inpatient surgical and outpatient specialist costs, and more costs related to emergency visits and dialysis. These results suggest that the driving force behind the bias we detect is that Black patients generate lesser medical expenses, conditional on health, even when we account for specific comorbidities. As a result, accurate prediction of costs necessarily means being racially biased on health.

How might these disparities in cost arise? The literature broadly suggests two main potential channels. First, poor patients face substantial barriers to accessing health care, even when enrolled in insurance plans. Although the population we study is entirely insured, there are many other mechanisms by which poverty can lead to disparities in use of health care: geography and differential access to transportation, competing demands from jobs or child care, or knowledge of reasons to seek care (*29*–*31*). To the extent that race and socioeconomic status are correlated, these factors will differentially affect Black patients. Second, race could affect costs directly via several channels: direct ("taste-based") discrimination, changes to the doctor–patient relationship, or others. A recent trial randomly assigned Black patients to a Black or White primary care provider and found significantly higher uptake of recommended preventive care when the provider was Black (*32*). This is perhaps the most rigorous demonstration of this effect, and it fits with a larger literature on potential mechanisms by which race can affect health care directly. For example, it has long been documented that Black patients have reduced trust in the health care system (*33*), a fact that some studies trace to the revelations of the Tuskegee study and other adverse experiences (*34*). A substantial literature in psychology has documented physicians' differential perceptions of Black patients, in terms of intelligence, affiliation (*35*), or pain tolerance (*36*). Thus, whether it is communication, trust, or bias, something about the interactions of Black patients with the health care system itself leads to reduced use of health care. The collective effect of these many channels is to lower health spending substantially for Black patients, conditional on need—a finding that has been appreciated for at least two decades (*37*).

Problem formulation

Our findings highlight the importance of the choice of the label on which the algorithm is trained. On the one hand, the algorithm manufacturer's choice to predict future costs is reasonable: The program's goal, at least in part, is

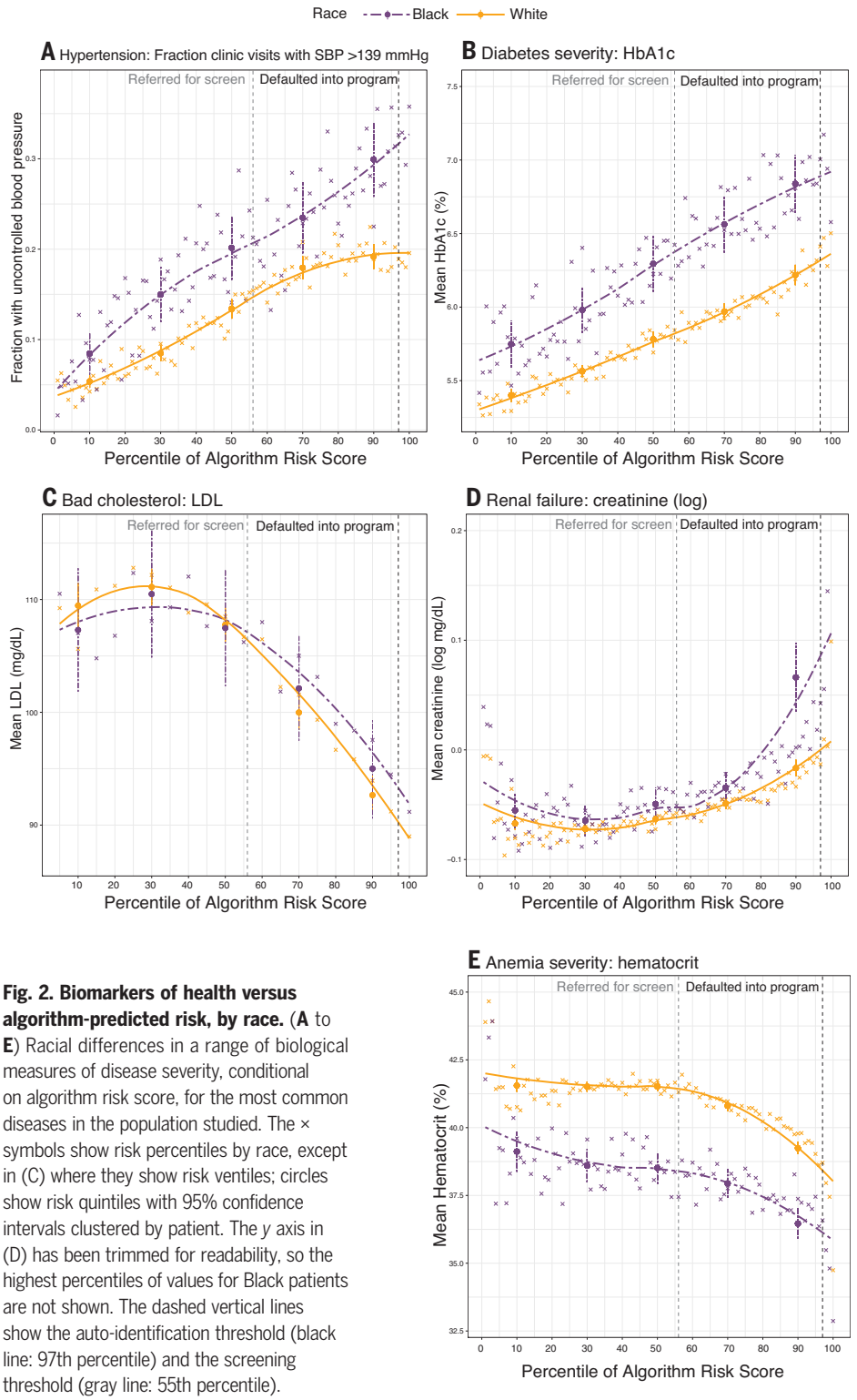


Fig. 2. Biomarkers of health versus algorithm-predicted risk, by race. (**A** to **E**) Racial differences in a range of biological measures of disease severity, conditional on algorithm risk score, for the most common diseases in the population studied. The × symbols show risk percentiles by race, except in (C) where they show risk ventiles; circles show risk quintiles with 95% confidence intervals clustered by patient. The *y* axis in (D) has been trimmed for readability, so the highest percentiles of values for Black patients are not shown. The dashed vertical lines show the auto-identification threshold (black line: 97th percentile) and the screening threshold (gray line: 55th percentile).

to reduce costs, and it stands to reason that patients with the greatest future costs could have the greatest benefit from the program. As noted in the supplementary materials, the manufacturer is not alone. Although the details of individual algorithms vary, the cost label reflects the industry-wide approach. For example, the Society of Actuaries's comprehensive evaluation of the 10 most widely used algorithms, including the particular algorithm we study, used cost prediction as its accuracy metric (*21*). As noted in the report, the enthusiasm for cost prediction is not restricted to industry: Similar algorithms are developed and used by non-profit hospitals, academic groups, and governmental agencies, and are often described in academic literature on targeting population health interventions (*18*, *19*).

On the other hand, future cost is by no means the only reasonable choice. For example, the evidence on care management programs shows that they do not operate to reduce costs globally. Rather, these programs primarily work to prevent acute health decompensations that lead to catastrophic health care utilization (indeed, they actually work to increase other categories of costs, such as primary care and home health assistance; see table S2). Thus avoidable future costs, i.e., those related to emergency visits and hospitalizations, could be a useful label to predict. Alternatively, rather than predicting costs at all, we could simply predict a measure of health; e.g., the number of active chronic health conditions. Because the program ultimately operates to improve the management of these conditions, patients with the most encounters related to them could also be a promising group on which to deploy preventative interventions.

The dilemma of which label to choose relates to a growing literature on "problem formulation" in data science: the task of turning an often amorphous concept we wish to predict into a concrete variable that can be predicted in a given dataset (*38*). Problems in health seem particularly challenging: Health is, by nature, holistic and multidimensional, and there is no single, precise way to measure it. Health care costs, though well measured and readily available in insurance claims data, are also the result of a complex aggregation process with a number of distortions due to structural inequality, incentives, and inefficiency. So although the choice of label is perhaps the single most important decision made in the development of a prediction algorithm, in our setting and in many others, there is often a confusingly large array of different options, each with its own profile of costs and benefits.

Experiments on label choice

Through a series of experiments with our dataset, we can gain some insight into how label choice affects both predictive performance and racial bias. We develop three new predictive algorithms, all trained in the same way, to predict the following outcomes: total cost in year t (this tailors cost predictions to our own dataset rather than the national training set), avoidable cost in year t (due to emergency visits and hospitalizations), and health in year t (measured by the number of chronic conditions that flare up in that year). We train all models in a random ⅔ training set and show all results only from the ⅓ holdout set. Furthermore, as with the original algorithm, we exclude race from the feature set (more details are in the materials and methods).

Table 2 shows the results of these experiments. The first finding is that all algorithms perform reasonably well for predicting not only the outcome on which they were trained but also the other outcomes: The concentration of realized outcomes in those at or above the 97th percentile is notably similar for all algorithms across all outcomes. The largest difference in performance across algorithms is seen for cost prediction: Of all costs in the holdout set, the fraction generated by those at or above the 97th percentile is 16.5% for the cost predictor versus 12.1% for the predictor

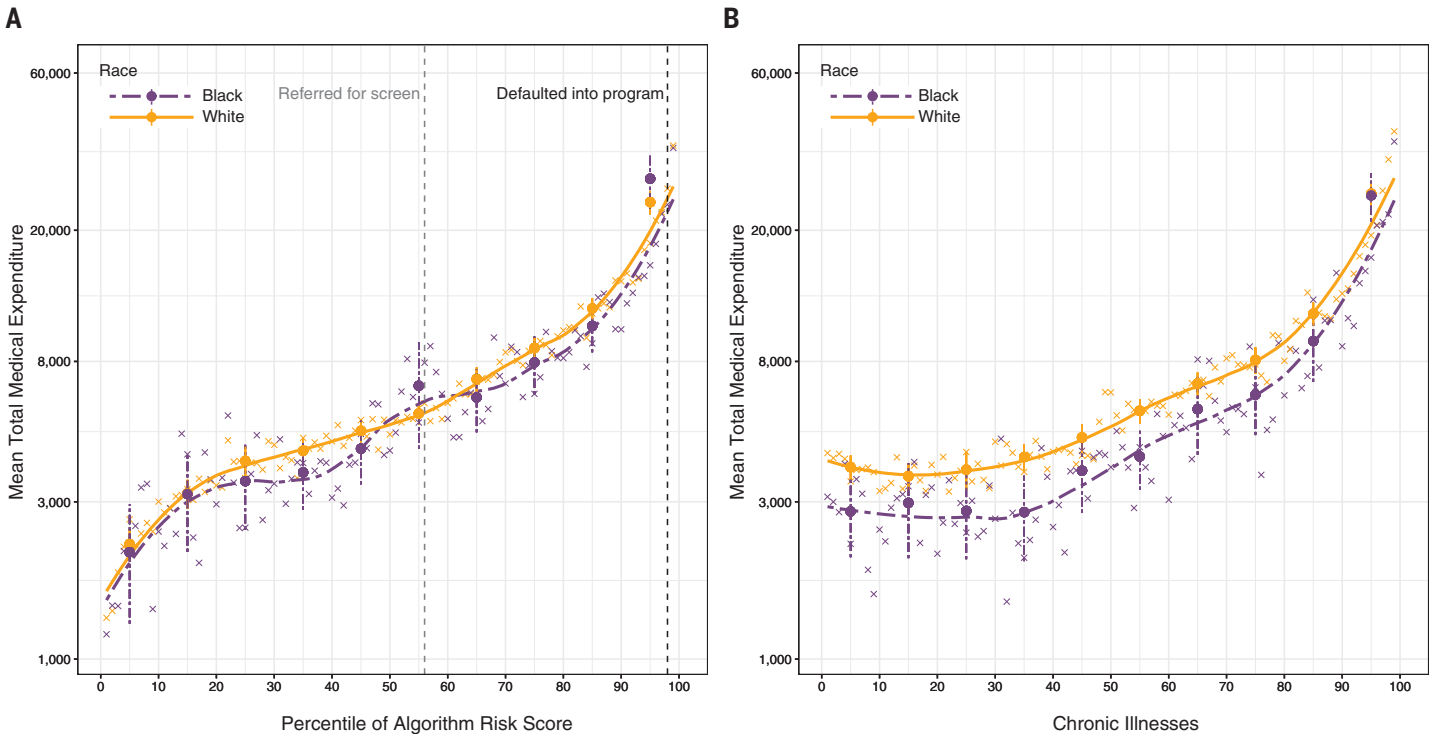


Fig. 3. Costs versus algorithm-predicted risk, and costs versus health, by race. (**A**) Total medical expenditures by race, conditional on algorithm risk score. The dashed vertical lines show the auto-identification threshold (black line: 97th percentile) and the screening threshold (gray line: 55th percentile). (**B**) Total medical expenditures by race, conditional on number of chronic conditions. The × symbols show risk percentiles; circles show risk deciles with 95% confidence intervals clustered by patient. The *y* axis uses a log scale.

of chronic conditions. We then test for label choice bias, defined analogously to calibration bias above: For two algorithms trained to predict Y and Y', and using a threshold τ indexing a (similarly sized) high-risk group, we would test $p[B|R > \tau] = p[B|R' > \tau]$ (here, p denotes probability and B represents Black patients).

We find that the racial composition of this highest-risk group varies far more across algorithms: The fraction of Black patients at or above these risk levels ranges from 14.1% for the cost predictor to 26.7% for the predictor of chronic conditions. Thus, although there could be many reasonable choices of label—all predictions are highly correlated, and any could be justified as a measure of patients' likely benefit from the program—they have markedly different implications in terms of bias, with nearly twofold variation in composition of Black patients in the highest-risk groups.

Relation to human judgment

As noted above, the algorithm is not used for program enrollment decisions in isolation. Rather, it is used as a screening tool, in part to alert primary care doctors to high-risk patients. Specifically, for patients at or above a certain level of predicted risk (the 55th percentile), doctors are presented with contextual information from patients' electronic health records and insurance claims and are prompted to consider enrolling them in the program. Thus, realized enrollment decisions largely reflect how doctors respond to algorithmic predictions, along with other administrative factors related to eligibility (for instance, primary care practice site, residence outside of a nursing home, and continual enrollment in an insurance plan).

Table 3 shows statistics on those enrolled in the program, accounting for 1.3% of observations in our sample: The enrolled individuals are 19.2% Black (versus 11.9% Black in our entire sample) and account for 2.9% of all costs and 3.3% of all active chronic conditions in the population as a whole. We then perform four counterfactual simulations to put these numbers in context; naturally, these simulations use only observable factors, not the many unobserved administrative and human factors that also affect enrollment. First, we calculate the realized program enrollment rate within each percentile of the original algorithm's predicted risk bins and randomly sample patients in each bin for enrollment. This simulation, which mimics "race-blind" enrollment conditional on algorithm score, would yield an enrolled population that is 18.3% Black (versus 19.2% observed; P = 0.8348). Second, rather than randomly sampling, we sample those with the highest predicted number of active chronic conditions within a risk bin (using our experimental algorithm described above); this would yield a population that is 26.9% Black. Finally, we compare this to simply assigning those with the highest predicted costs, or the highest number of active chronic conditions, to the program (also using our own algorithms detailed above), which would yield 17.2 and 29.2% Black patients, respectively. Thus, although doctors do redress a small part of the algorithm's bias, they do so far less than an algorithm trained on a different label.

Discussion

Bias attributable to label choice—the difference between some unobserved optimal prediction and the prediction of an algorithm trained on an observed label—is a useful framework through which to understand bias in algorithms, both

Table 2. Performance of predictors trained on alternative labels. For each new algorithm, we show the label on which it was trained (rows) and the concentration of a given outcome of interest (columns) at or above the 97th percentile of predicted risk. We also show the fraction of Black patients in each group.

Algorithm training label	Concentration in highest-risk patients (SE)						Fraction of Black patients in group with highest risk (SE)	
	Total costs		Avoidable costs		Active chronic conditions			
Total costs	0.165	(0.003)	0.187	(0.003)	0.105	(0.002)	0.141	(0.003)
Avoidable costs	0.142	(0.003)	0.215	(0.003)	0.130	(0.003)	0.210	(0.003)
Active chronic conditions	0.121	(0.003)	0.182	(0.003)	0.148	(0.003)	0.267	(0.003)
Best-to-worst difference	0.044		0.033		0.043		0.126	

Table 3. Doctors' decisions versus algorithmic predictions. For those enrolled in the high-risk care management program (1.3% of our sample), we first show the fraction of the population that is Black, as well as the fraction of all costs and chronic conditions accounted for by these observations. We also show these quantities for four alternative program enrollment rules, which we simulate in our dataset (using the holdout set when we use our experimental predictors). We first calculate the program enrollment rate within each percentile bin of predicted risk from the original algorithm and either (i) randomly sample patients or (ii) sample those with the highest predicted number of active chronic conditions within a bin and assign them to the program. The resultant values are then compared with values obtained by simply assigning the aforementioned 1.3% of our sample with (iii) the highest predicted cost or (iv) the highest number of active chronic conditions to the program.

Population	Fraction Black (SE)		Fraction of all costs (SE)		Fraction of all active chronic conditions (SE)	
Observed program enrollment (1.3%)	0.192	(0.003)	0.029	(0.001)	0.033	(0.001)
Simulated alternative enrollment rules						
Random, in predicted-cost bin	0.183	(0.003)	0.044	(0.002)	0.034	(0.001)
Predicted health, in predicted-cost bin	0.269	(0.003)	0.044	(0.002)	0.064	(0.002)
Highest predicted cost	0.172	(0.003)	0.100	(0.002)	0.047	(0.002)
Worst predicted health	0.292	(0.004)	0.067	(0.002)	0.076	(0.002)

in the health sector and further afield. This is because labels are often measured with errors that reflect structural inequalities (*39*). Within the health sector, using mortality or readmission rates to measure hospital performance penalizes those serving poor or non-White populations (*40*, *41*). Outside of the health arena, credit-scoring algorithms predict outcomes related to income, thus incorporating disparities in employment and salary (*2*). Policing algorithms predict measured crime, which also reflects increased scrutiny of some groups (*42*). Hiring algorithms predict employment decisions or supervisory ratings, which are affected by race and gender biases (*43*). Even retail algorithms, which set pricing for goods at the national level, penalize poorer households, which are subjected to increased prices as a result (*44*).

This mechanism of bias is particularly pernicious because it can arise from reasonable choices: Using traditional metrics of overall prediction quality, cost seemed to be an effective proxy for health yet still produced large biases. After completing the analyses described above, we contacted the algorithm manufacturer for an initial discussion of our results. In response, the manufacturer independently replicated our analyses on its national dataset of 3,695,943 commercially insured patients. This effort confirmed our results—by one measure of predictive bias calculated in their dataset, Black patients had 48,772 more active chronic conditions than White patients, conditional on risk score—illustrating how biases can indeed arise inadvertently.

To resolve the issue, we began to experiment with solutions together. As a first step, we suggested using the existing model infrastructure—sample, predictors (excluding race, as before), training process, and so forth—but changing the label: Rather than future cost, we created an index variable that combined health prediction with cost prediction. This approach reduced the number of excess active chronic conditions in Blacks, conditional on risk score, to 7758, an 84% reduction in bias. Building on these results, we are establishing an ongoing (unpaid) collaboration to convert the results of Table 3 into a better, scaled predictor of multidimensional health measures, with the goal of rolling these improvements out in a future round of algorithm development. Of course, our experience may not be typical of all algorithm developers in this sector. But because the manufacturer of the algorithm we study is widely viewed as an industry leader in data and analytics, we are hopeful that this endeavor will prompt other manufacturers to implement similar fixes.

These results suggest that label biases are fixable. Changing the procedures by which we fit algorithms (for instance, by using a new statistical technique for decorrelating predictors with race or other similar solutions) is not required. Rather, we must change the data we feed the algorithm—specifically, the labels we give it. Producing new labels requires deep understanding of the domain, the ability to identify and extract relevant data elements, and the capacity to iterate and experiment. But there is precedent for all of these functions in the literature and, more concretely, in the private companies that invest heavily in developing new and improved labels to predict factors such as consumer behavior (*45*). In addition, although health—as well as criminal justice, employment, and other socially important areas—presents substantial challenges to measurement, the importance of these sectors emphasizes the value of investing in such research. Because labels are the key determinant of both predictive quality and predictive bias, careful choice can allow us to enjoy the benefits of algorithmic predictions while minimizing their risks.

REFERENCES AND NOTES

1. J. Angwin, J. Larson, S. Mattu, L. Kirchner, "Machine Bias," *ProPublica* (23 May 2016); www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
2. S. Barocas, A. D. Selbst, *Calif. Law Rev.* **104**, 671 (2016).
3. A. Chouldechova, A. Roth, arXiv:1810.08810 [cs.LG] (20 October 2018).
4. A. Datta, M. C. Tschantz, A. Datta, *Proc. Privacy Enhancing Technol.* **2015**, 92–112 (2015).
5. L. Sweeney, *Queue* **11**, 1–19 (2013).
6. M. Kay, C. Matuszek, S. A. Munson, in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (ACM, 2015), pp. 3819–3828.
7. B. F. Klare, M. J. Burge, J. C. Klontz, R. W. Vorder Bruegge, A. K. Jain, *IEEE Trans. Inf. Forensics Security* **7**, 1789–1801 (2012).
8. J. Buolamwini, T. Gebru, in *Proceedings of the Conference on Fairness, Accountability and Transparency* (PMLR, 2018), pp. 77–91.
9. A. Caliskan, J. J. Bryson, A. Narayanan, *Science* **356**, 183–186 (2017).
10. S. Corbett-Davies, S. Goel, arXiv:1808.00023 [cs.CY] (31 July 2018).
11. M. De-Arteaga *et al.*, arXiv:1901.09451 [cs.IR] (27 January 2019).
12. M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian, in *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2015), pp. 259–268.
13. J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, S. Mullainathan, *Q. J. Econ.* **133**, 237–293 (2018).
14. C. S. Hong, A. L. Siegel, T. G. Ferris, *Issue Brief (Commonwealth Fund)* **19**, 1–19 (2014).
15. N. McCall, J. Cromwell, C. Urato, "Evaluation of Medicare Care Management for High Cost Beneficiaries (CMHCB) Demonstration: Massachusetts General Hospital and Massachusetts General Physicians Organization (MGH)" (RTI International, 2010).
16. J. Hsu *et al.*, *Health Aff.* **36**, 876–884 (2017).
17. L. Nelson, "Lessons from Medicare's demonstration projects on disease management and care coordination" (Working Paper 2012-01, Congressional Budget Office, 2012).
18. C. Vogeli *et al.*, *J. Gen. Intern. Med.* **22** (suppl. 3), 391–395 (2007).
19. D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, G. Escobar, *Health Aff.* **33**, 1123–1131 (2014).
20. J. Kleinberg, J. Ludwig, S. Mullainathan, Z. Obermeyer, *Am. Econ. Rev.* **105**, 491–495 (2015).
21. G. Hileman, S. Steele, "Accuracy of claims-based risk scoring models" (Society of Actuaries, 2016).
22. J. Kleinberg, S. Mullainathan, M. Raghavan, arXiv:1609.05807 [cs.LG] (19 September 2016).
23. A. Chouldechova, *Big Data* **5**, 153–163 (2017).
24. V. de Groot, H. Beckerman, G. J. Lankhorst, L. M. Bouter, *J. Clin. Epidemiol.* **56**, 221–229 (2003).
25. J. J. Gagne, R. J. Glynn, J. Avorn, R. Levin, S. Schneeweiss, *J. Clin. Epidemiol.* **64**, 749–759 (2011).
26. A. K. Parekh, M. B. Barton, *JAMA* **303**, 1303–1304 (2010).
27. D. Ettehad *et al.*, *Lancet* **387**, 957–967 (2016).
28. K.-T. Khaw *et al.*, *BMJ* **322**, 15 (2001).
29. K. Fiscella, P. Franks, M. R. Gold, C. M. Clancy, *JAMA* **283**, 2579–2584 (2000).
30. N. E. Adler, K. Newman, *Health Aff.* **21**, 60–76 (2002).
31. N. E. Adler, W. T. Boyce, M. A. Chesney, S. Folkman, S. L. Syme, *JAMA* **269**, 3140–3145 (1993).
32. M. Alsan, O. Garrick, G. C. Graziani, "Does diversity matter for health? Experimental evidence from Oakland" (National Bureau of Economic Research, 2018).
33. K. Armstrong, K. L. Ravenell, S. McMurphy, M. Putt, *Am. J. Public Health* **97**, 1283–1289 (2007).
34. M. Alsan, M. Wanamaker, *Q. J. Econ.* **133**, 407–455 (2018).
35. M. van Ryn, J. Burke, *Soc. Sci. Med.* **50**, 813–828 (2000).
36. K. M. Hoffman, S. Trawalter, J. R. Axt, M. N. Oliver, *Proc. Natl. Acad. Sci. U.S.A.* **113**, 4296–4301 (2016).
37. J. J. Escarce, F. W. Puffer, in *Racial and Ethnic Differences in the Health of Older Americans* (National Academies Press, 1997), chap. 6; www.ncbi.nlm.nih.gov/books/NBK109841/.
38. S. Passi, S. Barocas, arXiv:1901.02547 [cs.CY] (8 January 2019).
39. S. Mullainathan, Z. Obermeyer, *Am. Econ. Rev.* **107**, 476–480 (2017).
40. K. E. Joynt Maddox *et al.*, *Health Serv. Res.* **54**, 327–336 (2019).
41. K. E. Joynt Maddox, M. Reidhead, A. C. Qi, D. R. Nerenz, *JAMA Intern. Med.* **179**, 769–776 (2019).
42. K. Lum, W. Isaac, *Significance* **13**, 14–19 (2016).
43. I. Ajunwa, "The Paradox of Automation as Anti-Bias Intervention," available at SSRN (2016); https://ssrn.com/abstract=2746078.
44. S. DellaVigna, M. Gentzkow, "Uniform pricing in US retail chains" (National Bureau of Economic Research, 2017).
45. C. A. Gomez-Uribe, N. Hunt, *ACM Trans. Manag. Inf. Syst.* **6**, 13 (2016).

ACKNOWLEDGMENTS

We thank S. Lakhtakia, Z. Li, K. Lin, and R. Mahadeshwar for research assistance and D. Buefort and E. Maher for data science expertise. **Funding:** This work was supported by a grant from the National Institute for Health Care Management Foundation. **Author contributions:** Z.O. and S.M. designed the study, obtained funding, and conducted the analyses. All authors contributed to reviewing findings and writing the manuscript. **Competing interests:** The analysis was completely independent: None of the authors had any contact with the algorithm's manufacturer until after it was complete. No authors received compensation, in any form, from the manufacturer or have any commercial interests in the manufacturer or competing entities or products. There were no confidentiality agreements that limited reporting of the work or its results, no material transfer agreements, no oversight in the preparation of this article (besides ethical oversight from the approving IRB, which was based at a non-profit academic health system), and no formal relationship of any kind between any of the authors and the manufacturer. **Data and materials availability:** Because the data used in this analysis are protected health information, they cannot be made publicly available. We provide instead a synthetic dataset (using the R package synthpop) and all code necessary to reproduce our analyses at https://gitlab.com/labsysmed/dissecting-bias.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/366/6464/447/suppl/DC1
Materials and Methods
Figs. S1 to S5
Tables S1 to S4
References (*46*–*51*)

8 March 2019; accepted 4 October 2019
10.1126/science.aax2342

Causal Reasoning for Algorithmic Fairness

Joshua R. Loftus[1], Chris Russell[2,5], Matt J. Kusner[3,5], and Ricardo Silva[4,5]
[1]New York University [2]University of Surrey [3]University of Warwick
[4]University College London [5]Alan Turing Institute

Abstract

In this work, we argue for the importance of causal reasoning in creating fair algorithms for decision making. We give a review of existing approaches to fairness, describe work in causality necessary for the understanding of causal approaches, argue why causality is necessary for any approach that wishes to be fair, and give a detailed analysis of the many recent approaches to causality-based fairness.

1 Introduction

The success of machine learning algorithms has created a wave of excitement about the problems they could be used to solve. Already we have algorithms that match or outperform humans in non-trivial tasks such as image classification [18], the game of Go [37], and skin cancer classification [15]. This has spurred the use of machine learning algorithms in predictive policing [25], in loan lending [17], and to predict whether released people from jail will re-offend [9]. In these life-changing settings however, it has quickly become clear that machine learning algorithms can unwittingly perpetuate or create discriminatory decisions that are biased against certain individuals (for example, against a particular race, gender, sexual orientation, or other protected attributes). Specifically, such biases have already been demonstrated in natural language processing systems [5] (where algorithms associate men with technical occupations like 'computer programmer' and women with domestic occupations like 'homemaker'), and in online advertising [41] (where Google showed advertisements suggesting that a person had been arrested when that person had a name more often associated with black individuals).

As machine learning is deployed in an increasingly wide range of human scenarios, it is more important than ever to understand what biases are present in a decision making system, and what constraints we can put in place to guarantee that a learnt system never exhibits such biases. Research into these problems is referred to as *algorithmic fairness*. It is a particularly challenging area of research for two reasons: many different features are intrinsically linked to protected classes such as race or gender. For example, in many scenarios, knowledge

of someone's address makes it easy to predict their race with relatively high accuracy; while their choice of vocabulary might reveal much about their upbringing or gender. As such it is to easy to accidentally create algorithms that make decisions without knowledge of a persons race or gender, but still exhibit a racial or gender bias. The second issue is more challenging still, there is fundamental disagreement in the field as to what *algorithmic fairness* really means. Should algorithms be fair if they always make similar decisions for similar individuals? Should we instead call algorithms that make beneficial decisions for all genders at roughly the same rate fair? Or should we use a third different criteria? This question is of fundamental importance as many of these different criteria can not be satisfied at the same time [22].

In this work we argue that it is important to understand where these sources of bias come from in order to rectify them, and that causal reasoning is a powerful tool for doing this. We review existing notions of fairness in prediction problems; the tools of causal reasoning; and show how these can be combined together using techniques such as counterfactual fairness [23].

2 Current Work in Algorithmic Fairness

To discuss the existing measures of fairness, we use capital letters to refer to variables and lower case letters to refer to a value a variable takes. For example, we will always use A for a protected attribute such as gender, and a or a' to refer to the different values the attribute can take such as *man* or *woman*. We use Y to refer to the true state of a variable we wish to predict, for example the variable might denote whether a person defaults on a loan or if they will violate parole conditions. We will use $\hat{Y}$ to denote our prediction of the true variable Y. The majority of definitions of fairness in prediction problems are statements about probability of a particular prediction occurring given that some prior conditions hold. In what follows, we will use $P(\cdot \mid \cdot)$ to represent either conditional probability of events, probability mass functions or density functions, as required by the context.

2.0.1 Equalised Odds

Two definitions of fairness that have received much attention are equalised odds and calibration. Both were heavily used in the ProPublica investigation into Northpointe's COMPAS score, designed to gauge the propensity of a prisoner to re-offend upon release [16]. The first measure is equalised odds, which says that if a person truly has state y, the classifier will predict this at the same rate regardless of the value of their protected attribute. This can be written as an equation in the following form:

$$P(\hat{Y} = y \mid A = a, Y = y) = P(\hat{Y} = y \mid A = a', Y = y) \tag{1}$$

for all y, a, a'. Another way of stating this property is by saying that $\hat{Y}$ is independent of A given Y, which we will denote by $\hat{Y} \perp\!\!\!\perp A \mid Y$.

2.0.2 Calibration

The second condition is referred to as calibration (or 'test fairness' in [9]). This reverses the previous condition of equalised odds, and says that if the classifier predicts that a person has state y, their probability of actually having state y should be the same for all choices of attribute.

$$P(Y = y \mid A = a, \hat{Y} = y) = P(Y = y \mid A = a', \hat{Y} = y) \tag{2}$$

for all choices of y, a, and a', that is, $Y \perp\!\!\!\perp A \mid \hat{Y}$.

Although the two measures sound very similar, they are fundamentally incompatible. These two measures achieved some notoriety when Propublica showed that Northpointe's COMPAS score violated equalised odds, accusing them of racial discrimination. In response, Northpointe claimed that their COMPAS score satisfied calibration and that they did not discriminate. Kleinberg et al. [22] and Chouldechova [8] showed that both conditions cannot be satisfied at the same time except in special cases such as zero prediction error or if $Y \perp\!\!\!\perp A$.

The use of calibration and equalised odds has another major limitation. If $Y \not\perp\!\!\!\perp A$, the true scores Y typically have some inherent bias. This happens, for example, if the police are more likely to unfairly decide that minorities are violating their parole. The definitions of calibration or equalised odds do not explicitly forbid the classifier from preserving an existing bias.

2.0.3 Demographic Parity/Disparate Impact

Perhaps the most common non-causal notion of fairness is *demographic parity*, defined as follows:

$$P(\hat{Y} = y \mid A = a) = P(\hat{Y} = y \mid A = a'), \tag{3}$$

for all y, a, a', that is, $\hat{Y} \perp\!\!\!\perp A$. If unsatisfied, this notion is also referred to as *disparate impact*. Demographic parity has been used, for several purposes, in the following works: [14, 19, 20, 24, 42, 43].

Satisfying demographic parity can often require positive discrimination, where certain individuals who are otherwise very similar are treated differently due to having different protected attributes. Such *disparate treatment* can violate other intuitive notions of fairness or equality, contradict equalised odds or calibration, and in some cases is prohibited by law.

2.0.4 Individual Fairness

Dwork et al. [12] proposed the concept of *individual fairness* as follows.

$$P(\hat{Y}^{(i)} = y \mid X^{(i)}, A^{(i)}) \approx P(\hat{Y}^{(j)} = y \mid X^{(j)}, A^{(j)}), \text{ if } d(i, j) \approx 0, \tag{4}$$

where i, j refer to two different individuals and the superscripts $(i), (j)$ are their associated data. The function $d(\cdot, \cdot)$ is a 'task-specific' metric that describes

how any pair of individuals should be treated similarly in a fair world. The work suggests that this metric could be defined by 'a regulatory body, or ... a civil rights organization'. While this notion mitigates the issues with individual predictions that arose from demographic parity, it replaces the problem of defining fairness with defining a fair metric $d(\cdot, \cdot)$. As we observed in the introduction, many variables vary along with protected attributes such as race or gender, making it challenging to find a distance measure that will not allow some implicit discrimination.

2.0.5 Causal Notions of Fairness

A number of recent works use causal approaches to address fairness [1, 7, 21, 23, 35, 44], which we review in more detail in Section 5. We describe selected background on causal reasoning in Section 3. These works depart from the previous approaches in that they are not wholly data-driven but require additional knowledge of the structure of the world, in the form of a causal model. This additional knowledge is particularly valuable as it informs us how changes in variables propagate in a system, be it natural, engineered or social. Explicit causal assumptions remove ambiguity from methods that just depend upon statistical correlations. For instance, causal methods provide a recipe to express assumptions on how to recover from sampling biases in the data (Section 4) or how to describe mixed scenarios where we may believe that certain forms of discrimination should be allowed while others should not (e.g., how gender influences one's field of study in college, as in Section 5).

3 Causal Models

We now review causality in sufficient detail for our analysis of causal fairness in Section 5. It is challenging to give a self-contained definition of causality, as many working definitions reveal circularities on close inspection. For two random variables X and Y, informally we say that X *causes* Y when there exist at least two different *interventions* on X that result in two different probability distributions of Y. This does not mean we will be able to define what an "intervention" is without using causal concepts, hence circularities appear.

Nevertheless, it is possible to formally express causal assumptions and to compute the consequences of such assumptions if one is willing to treat some concepts, such as interventions, as primitives. This is just an instance of the traditional axiomatic framework of mathematical modelling, dating back to Euclid. In particular, in this paper we will make use primarily of the *structural causal model* (SCM) framework advocated by [29], which shares much in common with the approaches by [33] and [39].

3.1 Structural Causal Models

We define a causal model as a triplet (U, V, F) of sets such that:

is set to intervention levels a and a'. A joint distribution for $Y(a)$ and $Y(a')$ is implied by the model. Conditional distributions, such as $P(Y(a) = y_a, Y(a') = y_{a'} \mid A = a, Y = y, Z = z)$ are also defined. Figure 1(c) shows the case for interventions on Y. It is not difficult to show, as Y is not an ancestor of A in the graph, that $A(y, u) = A(y', u) = A(u)$ for all u, y, y'. This captures the notion the Y does not cause A.

3.3 Counterfactuals Require Untestable Assumptions

Unless structural equations depend on observed variables only, they cannot be tested for correctness (unless other untestable assumptions are imposed). We can illustrate this problem by noting that a conditional density function $P(V_j \mid V_i = v)$ can be written as an equation $V_j = f_1(v, U) \equiv F^{-1}_{V_i=v}(U) = F^{-1}_{V_i=v}(g^{-1}(g(U))) \equiv f_2(v, U')$, where $F^{-1}_{V_i=V}(\cdot)$ is the inverse cumulative distribution function corresponding to $P(V_j \mid V_i = v)$, U is an uniformly distributed random variable on $[0, 1]$, $g(\cdot)$ is some arbitrary invertible function on $[0, 1]$, and $U' \equiv g(U)$. While this is not fundamental for effects of causes, which depend solely on predictive distributions that at least in theory can be estimated from RCTs, different structural equations with the same interventional distributions will imply different joint distributions over the counterfactuals.

The traditional approach for causal inference in statistics tries to avoid any estimand that cannot be expressed by the marginal distributions of the counterfactuals (i.e., all estimands in which marginals $P(Y(a) = y_a)$ and $P(Y(a') = y_{a'})$ would provide enough information, such as the *average causal effect* $\mathsf{E}[Y(a) - Y(a')] = \mathsf{E}[Y \mid do(A = a)] - \mathsf{E}[Y \mid do(A = a')]$). Models that follow this approach and specify solely the univariate marginals of a counterfactual joint distribution are sometimes called *single-world* models [32]. However, as we will see, *cross-world* models seem a natural fit to algorithmic fairness. In particular, they are required for non-trivial statements that concern fairness at an individual level as opposed to fairness measures averaged over groups of individuals.

4 Why Causality is Critical For Fairness

Ethicists and social choice theorists recognise the importance of causality in defining and reasoning about fairness. Terminology varies, but many of their central questions and ideas, such as the role of agency in justice, responsibility-sensitive egalitarianism, and luck egalitarianism [10, 13, 31] involve causal reasoning. Intuitively, it is unfair for individuals to experience different outcomes caused by factors outside of their control. Empirical studies of attitudes about distributive justice [6, 26] have found that most participants prefer redistribution to create fairer outcomes, and do so in ways that depend on how much control individuals have on their outcomes. Hence, when choosing policies and designing systems that will impact people, we should minimise or eliminate the causal dependence on factors outside an individual's control, such as their perceived race or where they were born. Since such factors have influences on other

aspects of peoples' lives that may also be considered relevant for determining what is fair, applying this intuitive notion of fairness requires careful causal modelling as we describe here.

Is it necessary that models attempting to remove such factors be causal? Many other notions of algorithmic fairness have also attempted to control or adjust for covariates. While it is possible to produce identical predictions or decisions with a model that is equivalent mathematically but without overt causal assumptions or interpretations, the design decisions underlying a covariate adjustment are often based on implicit causal reasoning. There is a fundamental benefit from an explicit statement of these assumptions. To illustrate this, we consider a classic example of bias in graduate admissions.

4.1 Revisiting Gender Bias In Berkeley Admissions

The Berkeley admissions example [3] is often used to explain Simpson's paradox [38] and highlight the importance of adjusting for covariates. In the fall of 1973, about 34.6% of women and 44.3% of men who applied to graduate studies at Berkeley were admitted. However, this was not evidence that the admissions decisions were biased against women. Decisions were made on a departmental basis, and each department admitted proportions of men and women at approximately the same rate. However, a greater proportion of women applied to the most selective departments, resulting in a lower overall acceptance rate for women.

While the overall outcome is seemingly unfair, after controlling for choice of department it appears to be fair, at least in some sense. In fact, while the presentation of this example to illustrate Simpson's paradox often ends there, the authors in [3] conclude, "Women are shunted by their socialisation and education toward fields of graduate study that are generally more crowded, less productive of completed degrees, and less well funded, and that frequently offer poorer professional employment prospects." The outcome can still be judged to be unfair, not due to biased admissions decisions, but rather to the causes of differences in choice of department, such as socialisation. Achieving or defining fairness requires addressing those root causes and applying value judgements. Applicants certainly have some agency over which department they apply to, but that decision is not made free of outside influences. They had no control over what kind of society they had been born into, what sort of gender norms that society had during their lifetime, or the scarcity of professional role models, and so on.

The quote above suggests that the authors in [3] were reasoning about causes even if they did not make explicit use of causal modelling. Indeed, conditioning on the choice of the department only makes sense because we understand it has a causal relationship with the outcome of interest and is not just a spurious correlation. Pearl [29] provides a detailed account of the causal basis of Simpson's paradox.

4.2 Selection Bias and Causality

Unfairness can also arise from bias in how data is collected or sampled. For instance, if the police stop individuals on the street to check for the possession of illegal drugs, and if the stopping protocol is the result of discrimination that targets individuals of a particular race, this can create a feedback loop that justifies discriminatory practice. Namely, if data gathered by the police suggests that $P(Drugs = yes \mid Race = a) > P(Drugs = yes \mid Race = a')$, this can be exploited to justify an unbalanced stopping process when police resources are limited. How then can we assess its fairness? It is possible to postulate structures analogous to the Berkeley example, where a mechanism such as $Race \rightarrow Economic\ status \rightarrow Drugs$ explains the pathway. Debate would focus on the level of agency of an individual on finding himself or herself at an economic level that leads to increased drug consumption.

But selection bias cuts deeper than that, and more recently causal knowledge has been formally brought in to understand the role of such biases [2, 40]. This is achieved by representing a selection variable $Selected$ as part of our model, and carrying out inference by acknowledging that, in the data, all individuals are such that "$Selected = true$". The association between race and drug is expressed as $P(Drugs = yes \mid Race = a, Selected = true) > P(Drugs = yes \mid Race = a', Selected = true)$, which may or may not be representative of the hypothetical population in which everyone has been examined. The data cannot directly tell whether $P(Drugs = yes \mid Race = a, do(Selected = true)) > P(Drugs = yes \mid Race = a', do(Selected = true))$. As an example, it is possible to postulate the two following causal structures that cannot be distinguished on the basis of data already contaminated with selection bias: (i) the structure $Race \rightarrow Drugs$, with $Selected$ being a disconnected vertex; (ii) the structure $Race \rightarrow Selected \leftarrow H \rightarrow Drugs$, where H represents hidden variables not formally logged in police records.

In the latter case, we can check that drugs and race and unrelated. However, $P(Drugs = yes \mid Race = a, Selected = true) \neq P(Drugs = yes \mid Race = a', Selected = true)$, as conditioning on $Selected$ means that both of its causes $Race$ and H "compete" to explain the selection. This induces an association between $Race$ and H, which carries over the association between $Race$ and $Drugs$. At the same time, $P(Drugs = yes \mid Race = a, do(Selected = true)) = P(Drugs = yes \mid Race = a', do(Selected = true))$, a conclusion that cannot be reached without knowledge of the causal graph or a controlled experiment making use of interventions. Moreover, if the actual structure is a combination of (i) and (ii), standard statistical adjustments that remove the association between $Race$ and $Drugs$ cannot disentangle effects due to selection bias from those due to the causal link $Race \rightarrow Drugs$, harming any arguments that can be constructed around the agency behind the direct link.

4.3 Fairness Requires Intervention

Approaches to algorithmic fairness usually involve imposing some kind of constraints on the algorithm (such as those formula given by Section 2). We can view this as an intervention on the predicted outcome $\hat{Y}$. And, as argued in [1], we can also try to understand the causal implications for the system we are intervening on. That is, we can use an SCM to model the causal relationships between variables in the data, between those and the predictor $\hat{Y}$ that we are intervening on, and between $\hat{Y}$ and other aspects of the system that will be impacted by decisions made based on the output of the algorithm.

To say that fairness is an intervention is not a strong statement considering that any decision can be considered to be an intervention. Collecting data, using models and algorithms with that data to predict some outcome variable, and making decisions based on those predictions are all intentional acts motivated by a causal hypothesis about the consequences of that course of action. In particular, *not* imposing fairness can also be a deliberate intervention, albeit one of inaction.

We should be clear that prediction problems do not tell the whole story. Breaking the causal links between A and a prediction $\hat{Y}$ is a way of avoiding some unfairness in the world, but it is only one aspect of the problem. Ideally, we would like that no paths from A to Y existed, and the provision of fair predictions is predicated on the belief that it will be a contributing factor for the eventual change in the generation of Y. We are not, however, making any formal claims of modelling how predictive algorithmic fairness will lead to this ideal stage where causal paths from A to Y themselves disappear.

5 Causal Notions of Fairness

In this section we discuss some of the emerging notions of fairness formulated in terms of SCMs, focusing in particular on a notion introduced by us in [23], *counterfactual fairness*. We explain how counterfactual fairness relates to some of the more well-known notions of statistical fairness and in which ways a causal perspective contributes to their interpretation. The remainder of the section will discuss alternative causal notions of fairness and how they relate to counterfactual fairness.

5.1 Counterfactual Fairness

A predictor $\hat{Y}$ is said to satisfy *counterfactual fairness* if

$$P(\hat{Y}(a, U) = y \mid X = x, A = a) = P(\hat{Y}(a', U) = y \mid X = x, A = a), \quad (8)$$

for all y, x, a, a' in the domain of the respective variables [23]. The randomness here is on U (recall that background variables U can be thought of as describing a particular individual person at some point in time). In practice, this means we can build $\hat{Y}$ from any variable Z in the system which is not caused by A

Counterfactual Fairness

Matt Kusner *
The Alan Turing Institute and
University of Warwick
mkusner@turing.ac.uk

Joshua Loftus *
New York University
loftus@nyu.edu

Chris Russell *
The Alan Turing Institute and
University of Surrey
crussell@turing.ac.uk

Ricardo Silva
The Alan Turing Institute and
University College London
ricardo@stats.ucl.ac.uk

Abstract

Machine learning can impact people with legal or ethical consequences when it is used to automate decisions in areas such as insurance, lending, hiring, and predictive policing. In many of these scenarios, previous decisions have been made that are unfairly biased against certain subpopulations, for example those of a particular race, gender, or sexual orientation. Since this past data may be biased, machine learning predictors must account for this to avoid perpetuating or creating discriminatory practices. In this paper, we develop a framework for modeling fairness using tools from causal inference. Our definition of *counterfactual fairness* captures the intuition that a decision is fair towards an individual if it is the same in (a) the actual world and (b) a counterfactual world where the individual belonged to a different demographic group. We demonstrate our framework on a real-world problem of fair prediction of success in law school.

1 Contribution

Machine learning has spread to fields as diverse as credit scoring [20], crime prediction [5], and loan assessment [25]. Decisions in these areas may have ethical or legal implications, so it is necessary for the modeler to think beyond the objective of maximizing prediction accuracy and consider the societal impact of their work. For many of these applications, it is crucial to ask if the predictions of a model are *fair*. Training data can contain unfairness for reasons having to do with historical prejudices or other factors outside an individual's control. In 2016, the Obama administration released a report[2] which urged data scientists to analyze "how technologies can deliberately or inadvertently perpetuate, exacerbate, or mask discrimination."

There has been much recent interest in designing algorithms that make fair predictions [4, 6, 10, 12, 14, 16–19, 22, 24, 36–39]. In large part, the literature has focused on formalizing fairness into quantitative definitions and using them to solve a discrimination problem in a certain dataset. Unfortunately, for a practitioner, law-maker, judge, or anyone else who is interested in implementing algorithms that control for discrimination, it can be difficult to decide *which* definition of fairness to choose for the task at hand. Indeed, we demonstrate that depending on the relationship between a protected attribute and the data, certain definitions of fairness can actually *increase discrimination*.

*Equal contribution. This work was done while JL was a Research Fellow at the Alan Turing Institute.

[2]https://obamawhitehouse.archives.gov/blog/2016/05/04/big-risks-big-opportunities-intersection-big-data-and-civil-rights

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

In this paper, we introduce the first explicitly causal approach to address fairness. Specifically, we leverage the causal framework of Pearl [30] to model the relationship between protected attributes and data. We describe how techniques from causal inference can be effective tools for designing fair algorithms and argue, as in DeDeo [9], that it is essential to properly address causality in fairness. In perhaps the most closely related prior work, Johnson et al. [15] make similar arguments but from a non-causal perspective. An alternative use of causal modeling in the context of fairness is introduced independently by [21].

In Section 2, we provide a summary of basic concepts in fairness and causal modeling. In Section 3, we provide the formal definition of *counterfactual fairness*, which enforces that a distribution over possible predictions for an individual should remain unchanged in a world where an individual's protected attributes had been different in a causal sense. In Section 4, we describe an algorithm to implement this definition, while distinguishing it from existing approaches. In Section 5, we illustrate the algorithm with a case of fair assessment of law school success.

2 Background

This section provides a basic account of two separate areas of research in machine learning, which are formally unified in this paper. We suggest Berk et al. [1] and Pearl et al. [29] as references. Throughout this paper, we will use the following notation. Let A denote the set of *protected attributes* of an individual, variables that must not be discriminated against in a formal sense defined differently by each notion of fairness discussed. The decision of whether an attribute is protected or not is taken as a primitive in any given problem, regardless of the definition of fairness adopted. Moreover, let X denote the other observable attributes of any particular individual, U the set of relevant latent attributes which are not observed, and let Y denote the outcome to be predicted, which itself might be contaminated with historical biases. Finally, $\hat{Y}$ is the *predictor*, a random variable that depends on A, X and U, and which is produced by a machine learning algorithm as a prediction of Y.

2.1 Fairness

There has been much recent work on fair algorithms. These include fairness through unawareness [12], individual fairness [10, 16, 24, 38], demographic parity/disparate impact [36], and equality of opportunity [14, 37]. For simplicity we often assume A is encoded as a binary attribute, but this can be generalized.

Definition 1 (Fairness Through Unawareness (FTU))**.** *An algorithm is fair so long as any protected attributes A are not explicitly used in the decision-making process.*

Any mapping $\hat{Y} : X \to Y$ that excludes A satisfies this. Initially proposed as a baseline, the approach has found favor recently with more general approaches such as Grgic-Hlaca et al. [12]. Despite its compelling simplicity, FTU has a clear shortcoming as elements of X can contain discriminatory information analogous to A that may not be obvious at first. The need for expert knowledge in assessing the relationship between A and X was highlighted in the work on individual fairness:

Definition 2 (Individual Fairness (IF))**.** *An algorithm is fair if it gives similar predictions to similar individuals. Formally, given a metric $d(\cdot, \cdot)$, if individuals i and j are similar under this metric (i.e., $d(i, j)$ is small) then their predictions should be similar: $\hat{Y}(X^{(i)}, A^{(i)}) \approx \hat{Y}(X^{(j)}, A^{(j)})$.*

As described in [10], the metric $d(\cdot, \cdot)$ must be carefully chosen, requiring an understanding of the domain at hand beyond black-box statistical modeling. This can also be contrasted against population level criteria such as

Definition 3 (Demographic Parity (DP))**.** *A predictor $\hat{Y}$ satisfies demographic parity if $P(\hat{Y}|A = 0) = P(\hat{Y}|A = 1)$.*

Definition 4 (Equality of Opportunity (EO))**.** *A predictor $\hat{Y}$ satisfies equality of opportunity if $P(\hat{Y} = 1|A = 0, Y = 1) = P(\hat{Y} = 1|A = 1, Y = 1)$.*

These criteria can be incompatible in general, as discussed in [1, 7, 22]. Following the motivation of IF and [15], we propose that knowledge about relationships between all attributes should be taken into consideration, even if strong assumptions are necessary. Moreover, it is not immediately clear

for any of these approaches in which ways historical biases can be tackled. We approach such issues from an explicit causal modeling perspective.

2.2 Causal Models and Counterfactuals

We follow Pearl [28], and define a causal model as a triple (U, V, F) of sets such that

- U is a set of latent **background** variables,which are factors not caused by any variable in the set V of **observable** variables;
- F is a set of functions $\{f_1, \ldots, f_n\}$, one for each $V_i \in V$, such that $V_i = f_i(pa_i, U_{pa_i})$, $pa_i \subseteq V \backslash \{V_i\}$ and $U_{pa_i} \subseteq U$. Such equations are also known as **structural equations** [2].

The notation "pa_i" refers to the "parents" of V_i and is motivated by the assumption that the model factorizes as a directed graph, here assumed to be a directed acyclic graph (DAG). The model is causal in that, given a distribution $P(U)$ over the background variables U, we can derive the distribution of a subset $Z \subseteq V$ following an **intervention** on $V \setminus Z$. An intervention on variable V_i is the substitution of equation $V_i = f_i(pa_i, U_{pa_i})$ with the equation $V_i = v$ for some v. This captures the idea of an agent, external to the system, modifying it by forcefully assigning value v to V_i, for example as in a randomized experiment.

The specification of F is a strong assumption but allows for the calculation of **counterfactual** quantities. In brief, consider the following counterfactual statement, "the value of Y if Z had taken value z", for two observable variables Z and Y. By assumption, the state of any observable variable is fully determined by the background variables and structural equations. The counterfactual is modeled as the solution for Y for a given $U = u$ where the equations for Z are replaced with $Z = z$. We denote it by $Y_{Z \leftarrow z}(u)$ [28], and sometimes as Y_z if the context of the notation is clear.

Counterfactual inference, as specified by a causal model (U, V, F) given evidence W, is the computation of probabilities $P(Y_{Z \leftarrow z}(U) \mid W = w)$, where W, Z and Y are subsets of V. Inference proceeds in three steps, as explained in more detail in Chapter 4 of Pearl et al. [29]: 1. **Abduction**: for a given prior on U, compute the posterior distribution of U given the evidence $W = w$; 2. **Action**: substitute the equations for Z with the interventional values z, resulting in the modified set of equations F_z; 3. **Prediction**: compute the implied distribution on the remaining elements of V using F_z and the posterior $P(U \mid W = w)$.

3 Counterfactual Fairness

Given a predictive problem with fairness considerations, where A, X and Y represent the protected attributes, remaining attributes, and output of interest respectively, let us assume that we are given a causal model (U, V, F), where $V \equiv A \cup X$. We postulate the following criterion for predictors of Y.

Definition 5 (Counterfactual fairness)**.** *Predictor $\hat{Y}$ is* **counterfactually fair** *if under any context $X = x$ and $A = a$,*

$$P(\hat{Y}_{A \leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X = x, A = a), \tag{1}$$

for all y and for any value a' attainable by A.

This notion is closely related to **actual causes** [13], or token causality in the sense that, to be fair, A should not be a cause of $\hat{Y}$ in any individual instance. In other words, changing A while holding things which are not causally dependent on A constant will not change the distribution of $\hat{Y}$. We also emphasize that counterfactual fairness is an individual-level definition. This is substantially different from comparing different individuals that happen to share the same "treatment" $A = a$ and coincide on the values of X, as discussed in Section 4.3.1 of [29] and the Supplementary Material. Differences between X_a and $X_{a'}$ must be caused by variations on A only. Notice also that this definition is agnostic with respect to how good a predictor $\hat{Y}$ is, which we discuss in Section 4.

Relation to individual fairness. IF is agnostic with respect to its notion of similarity metric, which is both a strength (generality) and a weakness (no unified way of defining similarity). Counterfactuals and similarities are related, as in the classical notion of distances between "worlds" corresponding to different counterfactuals [23]. If $\hat{Y}$ is a deterministic function of $W \subset A \cup X \cup U$, as in several of

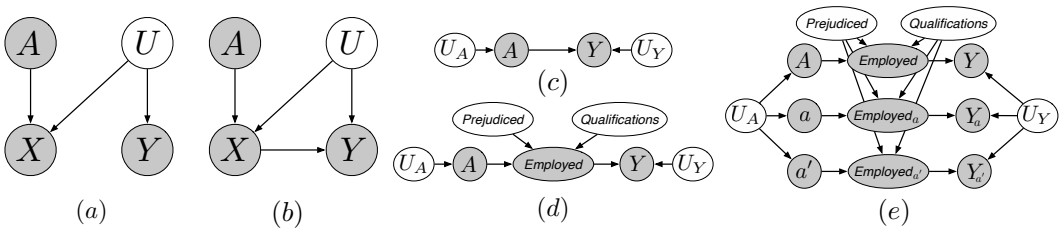


Figure 1: (a), (b) Two causal models for different real-world fair prediction scenarios. See Section 3.1 for discussion. (c) The graph corresponding to a causal model with A being the protected attribute and Y some outcome of interest, with background variables assumed to be independent. (d) Expanding the model to include an intermediate variable indicating whether the individual is employed with two (latent) background variables **Prejudiced** (if the person offering the job is prejudiced) and **Qualifications** (a measure of the individual's qualifications). (e) A twin network representation of this system [28] under two different counterfactual levels for A. This is created by copying nodes descending from A, which inherit unaffected parents from the factual world.

our examples to follow, then IF can be defined by treating equally two individuals with the same W in a way that is also counterfactually fair.

Relation to Pearl et al. [29]. In Example 4.4.4 of [29], the authors condition instead on X, A, and the observed realization of $\hat{Y}$, and calculate the probability of the counterfactual realization $\hat{Y}_{A \leftarrow a'}$ differing from the factual. This example conflates the predictor $\hat{Y}$ with the outcome Y, of which we remain agnostic in our definition but which is used in the construction of $\hat{Y}$ as in Section 4. Our framing makes the connection to machine learning more explicit.

3.1 Examples

To provide an intuition for counterfactual fairness, we will consider two real-world fair prediction scenarios: **insurance pricing** and **crime prediction**. Each of these correspond to one of the two causal graphs in Figure 1(a),(b). The Supplementary Material provides a more mathematical discussion of these examples with more detailed insights.

Scenario 1: The Red Car. A car insurance company wishes to price insurance for car owners by predicting their accident rate Y. They assume there is an unobserved factor corresponding to aggressive driving U, that (a) causes drivers to be more likely have an accident, and (b) causes individuals to prefer red cars (the observed variable X). Moreover, individuals belonging to a certain race A are more likely to drive red cars. However, these individuals are no more likely to be aggressive or to get in accidents than any one else. We show this in Figure 1(a). Thus, using the red car feature X to predict accident rate Y would seem to be an unfair prediction because it may charge individuals of a certain race more than others, even though no race is more likely to have an accident. Counterfactual fairness agrees with this notion: changing A while holding U fixed will also change X and, consequently, $\hat{Y}$. Interestingly, we can show (Supplementary Material) that in a linear model, regressing Y on A and X is equivalent to regressing on U, so off-the-shelf regression here is counterfactually fair. Regressing Y on X alone obeys the FTU criterion but is not counterfactually fair, so *omitting A (FTU) may introduce unfairness into an otherwise fair world.*

Scenario 2: High Crime Regions. A city government wants to estimate crime rates by neighborhood to allocate policing resources. Its analyst constructed training data by merging (1) a registry of residents containing their neighborhood X and race A, with (2) police records of arrests, giving each resident a binary label with $Y = 1$ indicating a criminal arrest record. Due to historically segregated housing, the location X depends on A. Locations X with more police resources have larger numbers of arrests Y. And finally, U represents the totality of socioeconomic factors and policing practices that both influence where an individual may live and how likely they are to be arrested and charged. This can all be seen in Figure 1(b).

In this example, higher observed arrest rates in some neighborhoods are due to greater policing there, not because people of different races are any more or less likely to break the law. The label $Y = 0$

does not mean someone has never committed a crime, but rather that they have not been caught. *If individuals in the training data have not already had equal opportunity, algorithms enforcing EO will not remedy such unfairness*. In contrast, a counterfactually fair approach would model differential enforcement rates using U and base predictions on this information rather than on X directly.

In general, we need a multistage procedure in which we first derive latent variables U, and then based on them we minimize some loss with respect to Y. This is the core of the algorithm discussed next.

3.2 Implications

One simple but important implication of the definition of counterfactual fairness is the following:

Lemma 1. *Let $\mathcal{G}$ be the causal graph of the given model (U, V, F). Then $\hat{Y}$ will be counterfactually fair if it is a function of the non-descendants of A.*

Proof. Let W be any non-descendant of A in $\mathcal{G}$. Then $W_{A\leftarrow a}(U)$ and $W_{A\leftarrow a'}(U)$ have the same distribution by the three inferential steps in Section 2.2. Hence, the distribution of any function $\hat{Y}$ of the non-descendants of A is invariant with respect to the counterfactual values of A. □

This does not exclude using a descendant W of A as a possible input to $\hat{Y}$. However, this will only be possible in the case where the overall dependence of $\hat{Y}$ on A disappears, which will not happen in general. Hence, Lemma 1 provides the most straightforward way to achieve counterfactual fairness. In some scenarios, it is desirable to define path-specific variations of counterfactual fairness that allow for the inclusion of some descendants of A, as discussed by [21, 27] and the Supplementary Material.

Ancestral closure of protected attributes. Suppose that a parent of a member of A is not in A. Counterfactual fairness allows for the use of it in the definition of $\hat{Y}$. If this seems counterintuitive, then we argue that the fault should be at the postulated set of protected attributes rather than with the definition of counterfactual fairness, and that typically we should expect set A to be closed under ancestral relationships given by the causal graph. For instance, if *Race* is a protected attribute, and *Mother's race* is a parent of *Race*, then it should also be in A.

Dealing with historical biases and an existing fairness paradox. The explicit difference between $\hat{Y}$ and Y allows us to tackle historical biases. For instance, let Y be an indicator of whether a client defaults on a loan, while $\hat{Y}$ is the actual decision of giving the loan. Consider the DAG $A \rightarrow Y$, shown in Figure 1(c) with the explicit inclusion of set U of independent background variables. Y is the objectively ideal measure for decision making, the binary indicator of the event that the individual defaults on a loan. If A is postulated to be a protected attribute, then the predictor $\hat{Y} = Y = f_Y(A, U)$ is not counterfactually fair, with the arrow $A \rightarrow Y$ being (for instance) the result of a world that punishes individuals in a way that is out of their control. Figure 1(d) shows a finer-grained model, where the path is mediated by a measure of whether the person is employed, which is itself caused by two background factors: one representing whether the person hiring is prejudiced, and the other the employee's qualifications. In this world, A is a cause of defaulting, even if mediated by other variables[3]. The counterfactual fairness principle however forbids us from using Y: using the twin network [4] of Pearl [28], we see in Figure 1(e) that Y_a and $Y_{a'}$ need not be identically distributed given the background variables.

In contrast, any function of variables not descendants of A can be used a basis for fair decision making. This means that any variable $\hat{Y}$ defined by $\hat{Y} = g(U)$ will be counterfactually fair for any function $g(\cdot)$. Hence, given a causal model, the functional defined by the function $g(\cdot)$ minimizing some predictive error for Y will satisfy the criterion, as proposed in Section 4.1. We are essentially learning a projection of Y into the space of fair decisions, removing historical biases as a by-product.

Counterfactual fairness also provides an answer to some problems on the incompatibility of fairness criteria. In particular, consider the following problem raised independently by different authors (e.g.,

[3] For example, if the function determining employment $f_E(A, P, Q) \equiv I_{(Q>0, P=0 \text{ or } A\neq a)}$ then an individual with sufficient qualifications and prejudiced potential employer may have a different counterfactual employment value for $A = a$ compared to $A = a'$, and a different chance of default.

[4] In a nutshell, this is a graph that simultaneously depicts "multiple worlds" parallel to the factual realizations. In this graph, all multiple worlds share the same background variables, but with different consequences in the remaining variables depending on which counterfactual assignments are provided.

[7, 22]), illustrated below for the binary case: ideally, we would like our predictors to obey both Equality of Opportunity and the *predictive parity* criterion defined by satisfying

$$P(Y = 1 \mid \hat{Y} = 1, A = 1) = P(Y = 1 \mid \hat{Y} = 1, A = 0),$$

as well as the corresponding equation for $\hat{Y} = 0$. It has been shown that if Y and A are marginally associated (e.g., recidivism and race are associated) and Y is not a deterministic function of $\hat{Y}$, then the two criteria cannot be reconciled. Counterfactual fairness throws a light in this scenario, suggesting that both EO and predictive parity may be insufficient if Y and A are associated: assuming that A and Y are unconfounded (as expected for demographic attributes), this is the result of A being a cause of Y. By counterfactual fairness, we should *not* want to use Y as a basis for our decisions, instead aiming at some function $Y_{\perp_A}$ of variables which are not caused by A but are predictive of Y. $\hat{Y}$ is defined in such a way that is an estimate of the "closest" $Y_{\perp_A}$ to Y according to some preferred risk function. *This makes the incompatibility between EO and predictive parity irrelevant*, as A and $Y_{\perp_A}$ will be independent by construction given the model assumptions.

4 Implementing Counterfactual Fairness

As discussed in the previous Section, we need to relate $\hat{Y}$ to Y if the predictor is to be useful, and we restrict $\hat{Y}$ to be a (parameterized) function of the non-descendants of A in the causal graph following Lemma 1. We next introduce an algorithm, then discuss assumptions that can be used to express counterfactuals.

4.1 Algorithm

Let $\hat{Y} \equiv g_\theta(U, X_{\not\succ A})$ be a predictor parameterized by θ, such as a logistic regression or a neural network, and where $X_{\not\succ A} \subseteq X$ are non-descendants of A. Given a loss function $l(\cdot, \cdot)$ such as squared loss or log-likelihood, and training data $\mathcal{D} \equiv \{(A^{(i)}, X^{(i)}, Y^{(i)})\}$ for $i = 1, 2, \ldots, n$, we define $L(\theta) \equiv \sum_{i=1}^{n} \mathbb{E}[l(y^{(i)}, g_\theta(U^{(i)}, x^{(i)}_{\not\succ A})) \mid x^{(i)}, a^{(i)}]/n$ as the empirical loss to be minimized with respect to θ. Each expectation is with respect to random variable $U^{(i)} \sim P_{\mathcal{M}}(U \mid x^{(i)}, a^{(i)})$ where $P_{\mathcal{M}}(U \mid x, a)$ is the conditional distribution of the background variables as given by a causal model $\mathcal{M}$ that is available by assumption. If this expectation cannot be calculated analytically, Markov chain Monte Carlo (MCMC) can be used to approximate it as in the following algorithm.

1: **procedure** FAIRLEARNING($\mathcal{D}, \mathcal{M}$) ▷ Learned parameters $\hat{\theta}$
2: For each data point $i \in \mathcal{D}$, sample m MCMC samples $U^{(i)}_1, \ldots, U^{(i)}_m \sim P_{\mathcal{M}}(U \mid x^{(i)}, a^{(i)})$.
3: Let $\mathcal{D}'$ be the augmented dataset where each point $(a^{(i)}, x^{(i)}, y^{(i)})$ in $\mathcal{D}$ is replaced with the corresponding m points $\{(a^{(i)}, x^{(i)}, y^{(i)}, u^{(i)}_j)\}$.
4: $\hat{\theta} \leftarrow \text{argmin}_\theta \sum_{i' \in \mathcal{D}'} l(y^{(i')}, g_\theta(U^{(i')}, x^{(i')}_{\not\succ A}))$.
5: **end procedure**

At prediction time, we report $\tilde{Y} \equiv \mathbb{E}[\hat{Y}(U^\star, x^\star_{\not\succ A}) \mid x^\star, a^\star]$ for a new data point $(a^\star, x^\star)$.

Deconvolution perspective. The algorithm can be understood as a deconvolution approach that, given observables $A \cup X$, extracts its latent sources and pipelines them into a predictive model. We advocate that *counterfactual assumptions must underlie all approaches that claim to extract the sources of variation of the data as "fair" latent components*. As an example, Louizos et al. [24] start from the DAG $A \rightarrow X \leftarrow U$ to extract $P(U \mid X, A)$. As U and A are not independent given X in this representation, a type of penalization is enforced to create a posterior $P_{fair}(U \mid A, X)$ that is close to the model posterior $P(U \mid A, X)$ while satisfying $P_{fair}(U \mid A = a, X) \approx P_{fair}(U \mid A = a', X)$. But *this is neither necessary nor sufficient for counterfactual fairness*. The model for X given A and U must be justified by a causal mechanism, and that being the case, $P(U \mid A, X)$ requires no postprocessing. As a matter of fact, model $\mathcal{M}$ can be learned by penalizing empirical dependence measures between U and pa_i for a given V_i (e.g. Mooij et al. [26]), but this concerns $\mathcal{M}$ and not $\hat{Y}$, and is motivated by explicit assumptions about structural equations, as described next.

Week 4: Fairness and Friends (to be added)