

What You See Is What You Get?

The Impact of Representation Criteria on Human Bias in Hiring

Andi Peng, Besmira Nushi, Emre Kıcıman, Kori Inkpen, Siddharth Suri, Ece Kamar

Microsoft Research
Redmond, WA, USA

{andipeng, benushi, emrek, kori, suri, eckamar}@microsoft.com

Abstract

Although systematic biases in decision-making are widely documented, the ways in which they emerge from different sources is less understood. We present a controlled experimental platform to study gender bias in hiring by decoupling the effect of world distribution (the gender breakdown of candidates in a specific profession) from bias in human decision-making. We explore the effectiveness of *representation criteria*, fixed proportional display of candidates, as an intervention strategy for mitigation of gender bias by conducting experiments measuring human decision-makers' rankings for who they would recommend as potential hires. Experiments across professions with varying gender proportions show that balancing gender representation in candidate slates can correct biases for some professions where the world distribution is skewed, although doing so has no impact on other professions where human persistent preferences are at play. We show that the gender of the decision-maker, complexity of the decision-making task and over- and under-representation of genders in the candidate slate can all impact the final decision. By decoupling sources of bias, we can better isolate strategies for bias mitigation in human-in-the-loop systems.

Introduction

Machine learning can aid decision-making and is used in recommendation systems that play increasingly prevalent roles in the world. We now deploy systems to help hire candidates (HireVue 2018), determine who to police more (Veale, Van Kleek, and Binns 2018), and assess the likelihood of an individual to recidivate on a crime (Angwin et al. 2016). Because these systems are trained on real world data, they often produce biased decision outcomes in a manner that is discriminatory against underrepresented groups. Systems have been found to unfairly discriminate against defendants of color in assessing bail (Angwin et al. 2016), incorrectly classify minority groups in facial recognition tasks (Raji and Buolamwini 2019), and engage in wage theft for honest workers (McInnis et al. 2016). While much of the algorithmic fairness literature has focused on understanding bias from algorithms in isolation (Dwork and Ilvento 2018),

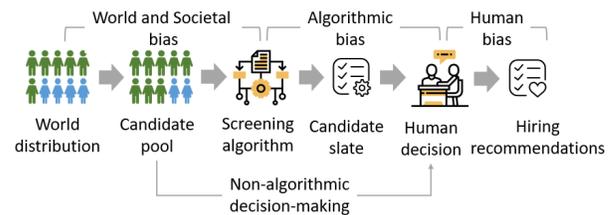


Figure 1: A high-level schematic for a hybrid system for hiring. A biased decision can be impacted by world, algorithmic, and human bias. Representation criteria is an intervention deployed when the candidate slate is generated.

little is known about how human decisions, when interacting with these systems, are impacted (Green and Chen 2019).

In this paper, we study algorithmic decision-making in hiring. While there exists a long history of studying hiring discrimination in fields like economics and social psychology, there has been little work done on the interplay between these demonstrated biases and the influence of algorithmic systems, especially across a wide variation of different professions and study participants. Here, we conduct a large-scale experiment studying *screening* or *recommendation systems* that choose a person (or set of people) from a pool of candidates (Kleinberg et al. 2019). Figure 1 shows the process flow for these hybrid systems, also known as human-in-the-loop or algorithm-in-the-loop systems. Data is gathered from the existing hiring pools in the world, often in larger quantities than what a human can feasibly assess, and fed into an algorithm which then screens for a candidate slate. The human decision-maker utilizes this filtered list to produce the final decision on whether or not a candidate should be recommended for hire. Biased gender hiring recommendations can be due to many different sources, including from (1) the world distribution (gender breakdown of candidates in a specific professions) (2) algorithmic bias in what's displayed to the human and (3) human decision-making itself.

Approach This study seeks to understand the impact of different sources of bias on hiring, specifically what properties of the candidate slate impact human decision-making deci-

sions. We define a biased decision to mean any difference in the system outcome such that one gender is favored over another in a manner that does not correspond to the gender distribution of the candidate slate as fed into the system. For example, imagine that two systems, A and B, both input the same distribution (50-50 M-F candidates). If System A produces outcomes that result in a 50-50 distribution of M-F recommended for hire but System B does not, then System B would be a biased system. We utilize the same reasoning for humans in assessing output decisions with respect to the input distribution that may come from the world (in traditional hiring practices) or from an algorithm (in hybrid systems). This view ascribes no notion of fairness or justice to the decision outcome.

We conduct experiments using Amazon Mechanical Turk to study how participants recommend candidates for different professions given candidate slates where we control factors such as education and experience, and artificially assign distributions of female and male candidates (see Figure 2). We generate profiles where we replace the names and pronouns displayed in the text and hold all other factors constant. Using representation criteria (i.e., fixed proportional display of gender distribution on the candidate slate), we randomly assign gender and ordering of profiles on each individual task to study how hiring decisions vary across different professions and gender distributions. This design affords us the following advantage: for any candidate slate, we can ask “how would the decision outcome be different if the particular gender of the candidates were changed?”. We ask participants to rank, out of 8 total, their top 4 candidates to recommend to a friend. We vary the proportion of M-F candidates in each profession and observe the impact of representation criteria on hiring outcomes. By displaying the specific gender distributions displayed to a human decision-maker as fixed controls, we can attribute any observed disparity in hiring outcomes as bias linked to human decision-making, rather than algorithmic or world distribution biases. We compare these to baseline outcomes of each profession conducted on the current world distribution of each profession and an AI model trained on word embeddings, both which represent systems that do not take into account representation criteria. We ask the following questions:

1. Does balancing the gender distribution in candidate slates mitigate bias? How does this effect vary across different professions?
2. In professions where this intervention is not enough, does over-representation help?
3. How do personal features of the decision-maker, such as gender, impact the outcome?

Takeaways Our results suggest three key takeaways regarding the relationship between world distribution bias and human decision-making. First, when genders are balanced in the candidate slate, gender bias in many professions with skewed world distributions can be mitigated. However, there are some professions in which this intervention is not enough to completely correct bias. Second, for professions like nannies and OBGYNs, no adjustment of representation

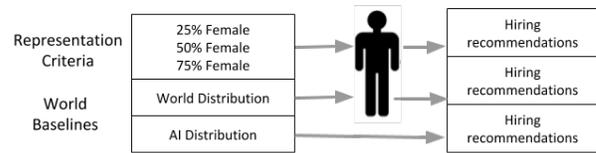


Figure 2: Our experimental design, highlighting the different candidate slates that can be displayed to the decision-maker. Representation criteria represent the intervention tested and world baselines highlight examples of existing processes.

criteria, even to the point of extreme over-representation, can fully correct for biased outcomes, suggesting that there are *persistent preferences* at play with respect to which genders people prefer for specific jobs that are independent of how candidates are displayed. Third, even across the same profession, there are personal features of the decision-maker, such as gender, that impact both the direction and strength of decision bias. As we seek to understand more regarding how algorithms can be deployed safely and *fairly* in the real-world, we must also study how bias from algorithms impacts decision-making by understanding the effect of different sources of bias on decision outcomes.

Related Work: Sources of Bias

Human Decision-Making Bias The opaqueness and inconsistencies of human decision-making have long been documented (Cunningham 2013; Glascher et al. 2011; Kahneman 2003). Although frameworks for evaluating risk (Rangel 2009), particularly with limited information under uncertainty (Platt and Huettel 2008), have been evaluated, their applicability to real-world tasks has often not been studied. Cognitive biases—unconscious, automatic influences on human decision-making that reliably produce reasoning errors—have been understood as heuristics, or mental shortcuts, that humans take when evaluating large amounts of uncertain information in a messy world (Tversky and Kahneman 1974; Thaler and Sunstein 2008). These heuristics range from availability bias (Tversky and Kahneman 1973), our willingness to judge the frequency of events in the world by the ease with which examples come to mind, to hindsight bias (Fischhoff and Beyth 1975), our tendency to believe falsely, after the outcome is known, that we’d have accurately predicted it. While manifestations of cognitive bias have been detailed, consistent theories on bias mitigation—prevention and reduction of these biases—are still limited.

Gender Bias in Hiring Hiring has long been a discriminatory practice, with socially salient features such as race and gender identified as being impactful towards hiring outcomes (Isaac, Lee, and Carnes 2009; Steinpreis, Anders, and Ritzke 1999). We focus on the role of gender in the hiring process due to its long-standing history of inequality and current usage in real-world systems, including the latest release from the World Economic Forum which extrapolates that it will take another 118 years to close the world gender gap (Schwab et al. 2017). This can be due to many factors, including on average women are more likely to work

part-time, be employed in low-paid jobs, and not take on management positions (Tobin 2017). As seen in Figure 3, the world distributions for M-F candidates in different professions vary greatly, presenting a data representation problem. Furthermore, there is also evidence that gender inequalities in the workplace stem, at least in part, from bias directed against women from those who hold internal sexist attitudes or preferences for a particular gender in different professions. Studies have documented personal discrimination against women by decision-makers (Koch, D’Mello, and Sackett 2015). For instance, a study found that the higher the participants scored on a hostile sexism test, the more likely they were to recommend a male candidate rather than female for a managerial position (Masser and Abrams 2004). These findings demonstrate that there exist human biases that must be decoupled from world distribution problems, particularly when placed within an algorithmic process.

Moreover, while these biases have been studied directly through classic resume screening experiments (Bertrand and Mullainathan 2003) as well as indirectly through questioning techniques (Hoffmann and Musch 2019), due to the difficulty in conducting studies assessing candidate preferences in the real world, much of this work has largely consisted of field studies focusing on isolated instances of a single profession or job call and its impact on callback rates (Kang et al. 2017). Algorithmic hiring and the increased usage of automated systems now motivates the need to understand this process in its entirety through decomposition of component biases, as the ability to deploy algorithms in screening processes has already been shown to have alarmingly detrimental and widespread effects (The Guardian 2018). Furthermore, while hiring processes were often unique to the specific company or recruiter soliciting candidates, algorithmic screening now impacts individuals on a scale that was impossible to see prior. Today, platforms such as LinkedIn Recruiter or Google Hire provide a singular large-scale, centralized screening database for many professions and companies, further motivating the need for a comprehensive study of the breakdown of decision biases across a wide array of professions and decision-makers. In a centralized platform where a recruiter can utilize the same tool to screen for thousands of candidates across many different professions, how does the input distribution and representational display impact the final decision? We build on this wide expanse of work in designing experiments to better understand the integration of these factors: how do real world gender distributions across different professions integrate with algorithmic interventions to impact hiring bias in hybrid systems?

Algorithmic Bias Although efforts have been made to study and correct bias in algorithmic hiring (Raub 2018), replicating bias has been challenging (Thebault-Spieker et al. 2017). Many attempts have sought to collect information on the current state of the world and, post-decision, statistically demonstrate a reduction of disparate outcomes. For example, researchers have attempted to debias a model trained on word embeddings after they found undesirable statistical links between certain jobs (such as programmer) to specific genders (such as male) (Bolukbasi et al. 2016; De-Arteaga et al. 2019). The rise of online platforms has also resulted in

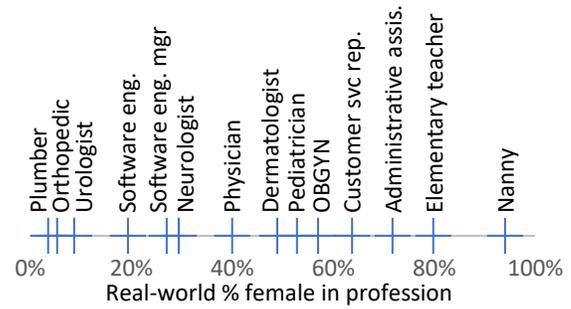


Figure 3: Real-world distributions across professions.

discriminatory rankings of candidates, (Hannak et al. 2017; Chen et al. 2018; Celis, Straszak, and Vishnoi 2018). For example, an empirical study found that STEM ads displayed on Facebook were targeted more at men than women (Lambrecht and Tucker 2017). Despite explicit attempts made by hiring platforms to improve algorithmic fairness (Geyik, Ambler, and Kenthapadi 2019), we still do not understand the impact of different sources of bias and how they impact final recommendations when used by a human in the real world. Does this statistical correlation mean that people hold an internal preference for men as programmers or do the current skewed world distributions, of which AI systems only serve to accentuate (Zhao et al. 2017), create biased decision outcomes?

Problem Formulation: Hiring

Definitions We recognize that the study of bias and fairness spans many fields, including psychology, computer science, and the social sciences. We offer the following terms as a reference guide for how we employ them, understanding that they are utilized differently in other contexts.

Bias: Any observed difference in recommendation decisions such that one gender is favored over another in a manner that does not correspond to the distribution input in to the algorithm or human. This can be impacted by algorithms, (*algorithmic bias*), humans (*decision-making* or *human bias*), or skewed world distributions (*world bias*).

Decision-making: The human process of taking input data and outputting a decision.

Representation Criteria: Fixed proportional display of candidates by gender. Slates can be 50-50 M-F, 75-25, etc.

World Distribution: The gender breakdown of candidates in a specific profession. We use this as a metric for understanding the labor pools available for hire.

Candidate Slate: The list of candidates generated from the world distribution by the screening algorithm that is displayed to a human decision-maker. Representation criteria is an intervention that alters the candidate slate.

World Distribution Problem Labor pools highlight a data representation problem, as gender breakdowns in different professions vary greatly (see Figure 3). Skews in distribution range from plumbers, where 3.5% of the workforce is female to nannies, where 94% are female (Bureau of Labor Statistics 2019). Even within licensed MDs in the United

States, specialties can range from OBGYNs, where 57% are female to orthopedic surgeons, where 5.3% are female (Association of American Medical Colleges 2018). This poses a ripe opportunity to isolate gender bias in world distributions from human bias and explore how hiring decisions are impacted by unrepresentative data as opposed to human decision-making processes.

Algorithmic Hiring Tools The rise of online profiles and job platforms has led to a shift in how jobs are filled: it is now easier than ever for recruiters to find candidates and for candidates to apply for jobs without ever engaging with a human. Many processes are now, if not completely automated, supported by automated screening tools such as Google's Hire (Lunden 2019).

Recent failures of algorithmic-aided hiring tools, such as Amazon's recruiting tool (The Guardian 2018), have spurred action to create less-biased, more representative systems. One such example is LinkedIn Recruiter: a service where a human has access to a portal where they can query and filter potential candidates by job related criteria. To ensure that the top ranked candidates displayed for a search reflect the underlying talent pool with respect to gender distribution, LinkedIn deployed a "representational ranking" algorithm (Geyik, Ambler, and Kenthapadi 2019). The intervention works as follows: after ML algorithms have calculated a "best fit" score for each candidate for a particular search, "representation ranking" post-processes the display of candidates so that, on each page, the proportion of female candidates shown is the same as the corresponding proportion of LinkedIn profiles matching that query. Note, this does not mean a 50-50 split is shown—rather, the goal is to reduce any gender skew that the AI system may have generated. However, the resulting effectiveness of such approaches has not been studied.

Research Question 1: Does balancing the gender distribution in candidate slates mitigate bias? How does this effect vary across different professions?

Research Question 2: For professions where this intervention is not enough, does over-representation help?

Research Question 3: How do personal features of the decision-maker, such as gender, impact human decision-making in hiring recommendations?

Human-in-the-Loop Methodology

We conduct large-scale experiments to study the effect of representation criteria on gender bias. We generate textual profiles for different professions and randomize the assignment of gender to individual candidates. We then ask over 4,900 unique study participants to select their top candidates to recommend for hire. Across trials, we vary representation criteria: the proportion of female and male candidates displayed. We compare these results against the expected distribution based on two world baselines: the current world representation for that profession and an AI score generated by projecting each profession across the gender subspace in a word embedding (Mikolov, Yih, and Zweig 2013).

Profile Generation We study the following professions: doctors (dermatologists, neurologists, OBGYNs, orthope-

Dr. **Robert** Brown, MD, is a board-certified orthopedic surgeon who, since 2002, practices at the Cleveland Clinic in Beachwood, OH. **He** is a graduate of the Johns Hopkins School of Medicine and completed **his** residency in Cleveland. **He** spends much of **his** time educating medical students at the Cleveland Clinic Lerner College of Medicine of Case Western Reserve University, where **he** serves as a Orthopedics Advisor and as Course Director for rotations that integrate bone fracture prevention and healthy living. **His** practice interests include health maintenance and diet/exercise, in addition to joint replacement. In **his** free time, **Robert** enjoys biking and exploring the outdoors.

Dr. **Mary** Brown, MD, is a board-certified orthopedic surgeon who, since 2002, practices at the Cleveland Clinic in Beachwood, OH. **She** is a graduate of the Johns Hopkins School of Medicine and completed **her** residency in Cleveland. **She** spends much of **her** time educating medical students at the Cleveland Clinic Lerner College of Medicine of Case Western Reserve University, where **she** serves as a Orthopedics Advisor and as Course Director for rotations that integrate bone fracture prevention and healthy living. **Her** practice interests include health maintenance and diet/exercise, in addition to joint replacement. In **her** free time, **Mary** enjoys biking and exploring the outdoors.

Figure 4: Example of a male and female candidate assignment for an orthopedic surgeon profile.

dic surgeons, pediatricians, physicians, and urologists), nannies, plumbers, administrative assistants, software engineers, software engineering managers, customer service representatives, and elementary teachers. For each profession, eight fictional candidates are generated using online biographies as inspiration. Our experimental data underwent ethics, privacy, and IRB review. All profiles follow roughly the same setup and include: name, degree, specialty or focus within the profession, experience length (shown explicitly through years of work or implicitly via graduation year), and information on personal life. This was done to induce low variance and control for possible confounding factors such as education and expertise. To further ensure profile equivalency, we run pilot studies asking participants to differentiate between candidates and remove profiles that were ranked too many or few times. We then randomize the gender assignment of each candidate with respect to the slate tested. For each profile, the number of times the name and gendered pronouns appear is kept constant. We generate our profiles using the most common female and male first names according to the 2018 Census, keeping last names constant. Textio analysis is run on each profile to remove other gendered words (Gannes 2015). All profiles are consistently 300-400 words long.

HIT Generation Using these profiles, we generate HIT tasks where each participant is shown a unique candidate slate and asked to recommend candidates to *a friend who is looking for a specific profession*. This is done with the intention that a recommendations to a third-party would induce less personal bias. Each HIT has the same layout, where participants are shown 8 candidates and asked to rank, from 1 to 4 (with 1 as their top choice), their top candidates out of 8. We enforce representation criteria by randomly assigning gender to each profile according to the distribution studied (for example, a 25F slate would randomly assign 2 female and 6 male candidates, then randomly order their display). All gender assignments vary across individual HITs in a condition, with no participant ever shown the same slate. This is done to remove possible confounding factors such as rank ordering preferences, recency bias, and variance across profiles to isolate the gender variable. Afterwards, participants are asked to fill out a survey regarding their demographic information and to qualitatively describe which factors they

believe to be important in their decision. Each HIT requires 5-10 minutes to complete.

Ranking Task We deploy our experiments using a crowd-sourced pool on Amazon Mechanical Turk and compensate workers at a wage of \$15 per hour. All participants are preliminarily screened according to the following qualifications: hold above a 90% approval rating, unique study participant, and based in the United States. For each profession, 300 unique study participants are recruited and each assigned a candidate slate of 8 profiles. The experiment is deployed in 3 batches for each profession: 100 HITs for gender slates of 25% female-75% male (25F), 50% female-50% male (50F), and 75% female-25% male (75F). We exclude duplicate or missing ranks. All tasks were completed between January 18 and April 19, 2019.

Statistical Testing A hypergeometric distribution, which represents unbiased hiring decisions, is generated for each representation criteria (25F, 50F, 75F) to model the expected number of female candidates selected in the top 4 (out of 8). Such a distribution models the discrete probability distribution of binary draws *without* replacement from a finite population. This is relevant to our task as the probability of each candidate being ranked changes on each draw. Note that, modeling the screening decision-making process with a hypergeometric random variable $X \sim \text{hypergeometric}(K, N, n)$ depends on the total number of candidates (N), the number of candidates from a gender group (K), and the number of candidates to be selected (n). We intentionally remove other factors (such as expertise), to study the effect of representation criteria in isolation. Here, $\Pr(X = k)$ denotes the probability that exactly k candidates from the candidate slate made it onto the list of n recommended candidates. An unbiased decision process that does not favor any one gender should follow (as closely as possible) a hypergeometric distribution with the following probability mass function:

$$\Pr(X = k) = \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}}$$

If the decision process does not follow such a distribution, then the decision-making entity is either (i) biased towards a specific group ; e.g., towards female candidates if the mass of the observational data distribution is concentrated above the hypergeometric mean (i.e., $\frac{nK}{N}$), or (ii) there exist other factors that govern the decision process leading to decision bias. Since we randomly assign gender to profiles for each HIT and control for equal expertise in the profile text, deviations of the observational distribution from the corresponding hypergeometric can only be attributed to gender bias.

We conduct Chi-square goodness-of-fit tests for our observed distributions against the hypergeometric to test if selections of female candidates deviate significantly. Note: the hypergeometric captures the number of female candidates selected, but not the ranking order. We study ranking by reformulating the problem as recommending 1 candidate out of 8, 2 out of 8, etc and observing the distribution of female candidates selected. The closer the observed distribution of rankings is to the expected, the higher the correlation value

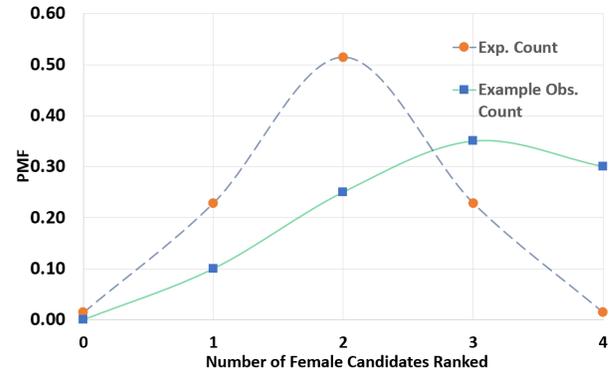


Figure 5: Example probability mass function for hypergeometric ($K = 400$, $N = 800$, $n = 400$) where, for 100 ranking slates where each slate displays 4/8 candidates as women (50F), the expected number of total female candidates ranked out of 4 follows this distribution. The dashed line represents the ideal and the solid line a biased recommendation distribution in favor of women.

and consequently lower the observed gender bias is.

World Baselines To determine the cross-sectional impact of representation criteria on hiring outcomes, we create two baselines: current world representation and an AI score generated from word embeddings. They are given as additional representation criteria to model potential hiring decisions that are currently utilized today by decision-makers. We recognize that these do not fully represent real-world AI hiring systems, but are useful for understanding the potential impact of real world data on algorithmic systems.

Current World Representation To address the discretization challenge of displaying world distributions that were not multiples of 8 (for example, it is impossible to display 3% of plumbing candidates out of 8 as female for one slate), we run the required number of candidates to create the appropriate proportion across all HITs. For example, we ran 24 HITs of 12.5F for plumbers to create 24 female candidates out of the 800 total candidates required for $p_F = 0.03$ and $n = 100$. We then test our observed rankings against a weighted hypergeometric distribution of the same proportions.

AI Score Hiring platforms do not publicly share their ranking algorithms, so to better model the impact of world distributions on algorithmic systems, we create example AI decision scores by projecting each profession across a gender direction as represented by word embeddings trained on a Google News text-corpora. A word embedding is a representation of a single word (or common phrase) w as a d -dimensional word vector $\vec{w} \in R^d$ (Mikolov, Yih, and Zweig 2013). Word embeddings, trained on word co-occurrence in a mass text-corpora, can serve as a dictionary of sorts to analyze word meaning for words with similar meaning tend to have vectors that are closer together. To understand the gender biases present across professions (which professions are closer to *she* than to *he*), we utilize the w2vNEWS embedding, a $d = 300$ -dimensional word2vec pre-trained on the Google News corpus, and project the profession of interest

across the gender direction v_g (Bolukbasi et al. 2016). The cosine similarity between each word vector w^d and the gender direction v_g is represented by the following equation:

$$\cos(w^d, v_g) = \frac{w^d \cdot v_g}{\|w^d\| \|v_g\|}$$

We then convert the similarity into a gender score (Bolukbasi et al. 2016) by normalizing the resulting correlation against the same $\vec{she} - \vec{he}$ vector to obtain a score from 0–100, where 0 represents a perfect *she* match, 50 a gender-neutral word, and 100 a perfect *he* match.

Results

1. Balancing gender slates can serve as an effective intervention for mitigating bias across many professions, even those that have extremely skewed world distributions.
2. For professions where this is not effective, the human decision-making bias is so strongly in favor of a specific gender that it dominates any over-representation.
3. Personal features of the decision-maker, such as gender, can also impact the hiring outcome.

Bias Induced by the World

We show findings from experiments where we present a gender balanced candidate slate for all professions to test the effectiveness of this as an intervention strategy. We discuss the impact of candidate gender in observed outcomes by comparing the observed ranking of female candidates to their expected distributions to decouple the impact of human decision-making bias from world bias.

Gender of Candidate Hiring decisions do not always reflect the world distribution. When we control the gender representation of candidate slates to be equal (50F), many outcomes can be seen as unbiased and follow the expected hypergeometric distribution, not the input representation.

Result 1: Balancing candidate slates to equally represent gender can serve as an effective intervention for bias mitigation across many professions.

Despite many professions where the world distribution is radically skewed towards one gender, we see no bias if gender is represented equally across a candidate slate. For example (see Table 1), although only 5.3% of orthopedic surgeons in the world are female, we observe no human bias in the gender balanced ranking tasks (%F Rank4 and %F Rank1). This was also seen in plumbers, software engineers, software engineering managers, neurologists, physicians, pediatricians, customer service representatives, administrative assistants, and elementary teachers. This finding suggests that some unrepresentative professions may be entirely due to the lack of candidates in the hiring pool and not from inherent human biases, and creating a balanced slate can serve as an effective form of bias mitigation.

Observation 1a: Slates that are not gender balanced can create biased decisions. Male participants exhibit bias against female candidates in ranking dermatologists ($\chi = 6.13, 0.047$) and urologists ($\chi = 9.39, 0.009$) when candidate slates are 75F on the 4-rank task. This suggests a desire to

over-represent male candidates when the distribution does not favor them. On the other hand, male participants exhibit bias against male candidates in ranking OBGYNs ($\chi = 11.05, 0.004$) and physicians ($\chi = 6.76, 0.034$) when candidate slates are 25F on the 4-rank task, meaning that they wish to over-represent female candidates when few are present. Female participants also exhibit bias against female candidates in ranking urologists ($\chi = 10.67, 0.005$) when candidate slates are 25F on the 4-rank task.

These results suggest that for some professions, balancing the genders might not illicit biased decisions but unbalanced representations in favor of one gender does. For example, male participants may hold an internal preference for wishing to see male dermatologists and urologists, professions where there could exist a sense of relatability in understanding concerns, but that preference only emerges when the candidate slates are skewed against male candidates.

Observation 1b: Some professions never create biased decisions, regardless of the representation criteria. We observe several professions such as neurologists, pediatricians, plumbers, administrative assistants, software engineers, software engineering managers, and customer service representatives, where 25F and 75F skewed slates show no bias in hiring decisions, i.e., the resulting distributions reflect the expected distributions for these slates (see Figures 6e-g). We postulate this is because in these professions, there is no clear reason why one gender would be better suited to succeed at the job, and the lack of bias generated suggests that representation criteria remain a valid intervention strategy for skewed world distributions.

Bias Induced by Humans

For professions where balancing the candidate slate was not effective in mitigating bias, we test over-representation of a specific gender. We show findings from experiments where we present gender imbalanced slates for all professions to study the strength of human decision-making bias. For each profession, we test candidate slates of 25F, 50F, and 75F and real-world baselines (representation criteria generated from world distributions and example AI scores).

Result 2: Despite over-representation of candidates, some professions consistently produce biased decisions, suggesting there are persistent human biases at play as to which gender is preferred for certain professions.

Some professions demonstrate persistent human biases that are not impacted by representation criteria at all. For all candidate slates ranging from 25F to 75F, female decision-makers demonstrate a clear and persistent preference for female candidates and over-represent them on all ranking tasks and representation distributions (see Figures 6a-d).

For example (see Table 1), the number of female OBGYNs ranked out of 8 was far more than expected, with 16 participants ranking all 4 selected candidates vs. the expected 1.4 (4F Selected). This was also seen in urologists, dermatologists, and nannies, suggesting that persistent preferences for who humans prefer for certain jobs may be ingrained in some professions, possibly due to societal stereotypes of what gender is best suited for specific jobs or per-

Profession	0F Selected	1F Selected	2F Selected	3F Selected	4F Selected	%F World	AI Score (0-100)	%F Rank4 (p)	%F Rank1 (p)
<i>Exp. count (n=100)</i>	1.4	22.9	51.4	22.9	1.4	50.0	50.0	50	50
Plumber	2	22	55	18	3	3.5	40.0	50(0.513)	46(0.424)
Ortho. surgeon	2	26	57	12	3	5.3	31.6	47(0.086)	54(0.428)
Urologist	6	23	48	22	1	8.7	38.0	47(0.005)	53(0.549)
Software eng.	0	20	53	24	3	19.3	36.5	53(0.460)	47(0.549)
Software eng. man.	1	24	57	17	1	27.0	40.6	48(0.659)	48(0.689)
Neurologist	2	22	56	17	3	29.4	42.9	49(0.420)	51(0.842)
Physician	1	23	48	26	2	40.0	43.1	51(0.907)	51(0.841)
Dermatologist	5	28	49	18	0	48.9	65.0	45(0.013)	45(0.317)
Pediatrician	2	27	40	28	3	52.8	66.2	51(0.171)	56(0.230)
OBGYN	1	16	40	27	16	57.0	74.2	60(<0.000)	59(0.072)
Customer serv. rep.	0	29	51	20	0	63.7	65.9	48(0.301)	50(1.000)
Admin. assistant	1	17	49	31	2	71.7	59.0	54(0.301)	59(0.072)
Elem. teacher	1	15	56	27	1	79.8	67.9	5(0.391)	54(0.424)
Nanny	0	18	40	35	7	94.0	78.2	58(<0.000)	53(0.549)

Table 1: Observed number of female candidates ranked in the top 4 by profession for the 50F distribution. The top row details the expected hypergeometric values. The percentage of female candidates recommended for hire for the top 4 and top 1 ranking tasks are compared to the current world representation and normalized AI scores calculated from word embeddings. The p -value represents the Chi-square test of the distribution of observed female candidates selected against the expected hypergeometric.

sonal comfort related to the individual. These could not be mitigated through any representation or display criteria. This motivates future work in discovering other important factors that take into account cultural and societal expectations.

Gender of Decision-Maker We decompose our data across the identity of the decision-maker. Figures 7-8 detail how human bias can be separated by decision-making gender.

Result 3: There are personal features of the decision-maker, such as gender, that impact hiring outcomes. Sometimes a biased outcome is driven primarily by one gender and other times an unbiased outcome is created due to opposing effects from each gender that cancel out.

Observation 3a: Biased hiring decisions are at times driven by strong effects from one decision-making gender. For a subset of the professions where we observe bias, the aggregate effect is primarily driven by one decision-making gender. Figure 7 shows that strong bias in professions like OBGYNs and nannies is mostly attributed to female decision-makers. Here, the weaker or nonexistent effect by male participants is subsumed by the stronger effect demonstrated by female participants. This suggests that for certain jobs, there exist personal preferences that are ingrained within the identity of the individual decision-maker.

Observation 3b: Aggregate unbiased decisions are sometimes hidden by opposite biases from decision-making genders. For a subset of the professions where no aggregate hiring bias is seen, we find differences when we split the data by the gender of the decision-maker. Figure 8 shows how opposite effects from male and female participants create an aggregate unbiased decision for some professions. This suggests that bias decomposition into factors related to the decision-maker’s identity is important, as signals can be lost when looking at aggregate outcomes alone.

Bias Induced by the Task

We now detail additional observations related to how the complexity and display of the task may have impacted the outcomes presented. We show findings from differences in hiring outcomes across participants’ top 1, 2, and 4-ranked candidates and qualitative feedback collected.

Observation 4a: Decision-makers rarely demonstrate bias for the top candidate, but exhibit bias in later rankings. As seen in Table 2, in professions where aggregate biased decisions is observed in top 4 rankings (Human Select 4), the effect decreases for the top 2 ranking (Human Select 2) and disappears entirely for the top 1 ranking (Human Select 1). This may be due to task fatigue, where humans, when asked to rank one candidate, focus on more objective factors related to competency such as education and experience. However, when the task becomes more complex, markers for the best candidates disappear and cognitive biases take over. This suggests that interventions can be tailored to specific tasks depending on the complexity and amount of information that decision-makers are exposed to.

Observation 4b: Participants are more willing to express their choice of a female candidate as being important to their ranking decision. For OBGYNs, the strongest bias seen in any profession tested, both male and female participants overwhelmingly state that they expected a friend to prefer a female doctor for addressing feminine needs. However, when we test urologists, the male counterpart, we do not observe the same willingness to comment on a preference for male doctors. We postulate this is due to an understanding of societal norms surrounding when it is acceptable to express preference, and that overt preference for a female candidate in under-represented fields is considered more *fair*.

Observation 4c: Participants often give justifications for decisions that do not correlate to the effect observed, particularly when representations are unbalanced. For exam-

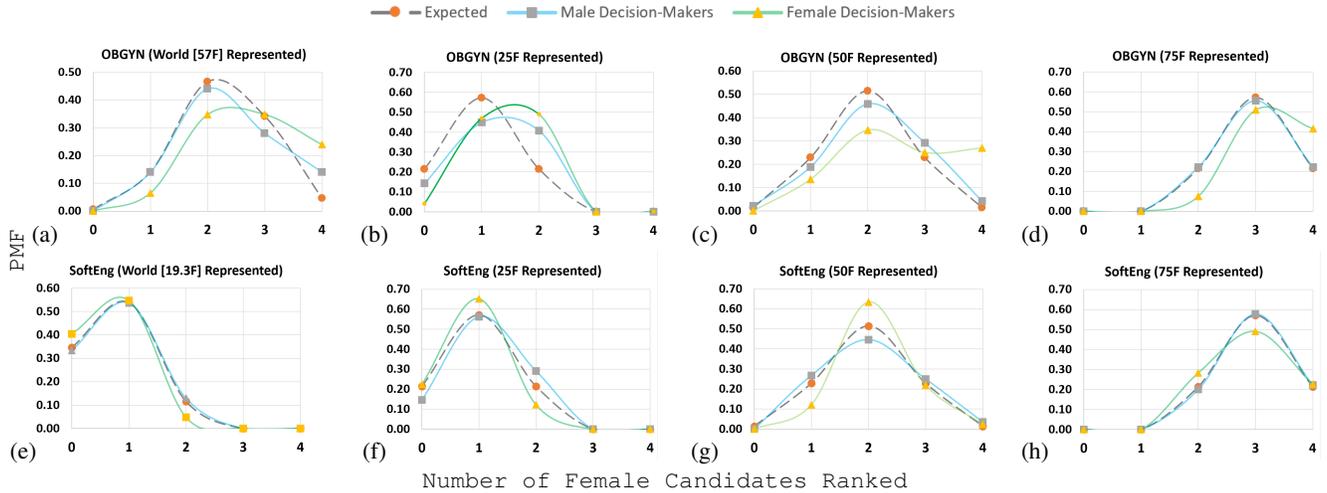


Figure 6: Hypergeometric distributions of the expected and observed rankings for OBGYNs and software engineers by gender of the decision-maker. We observe near-universal bias in favor of female OBGYNs but none for software engineers.

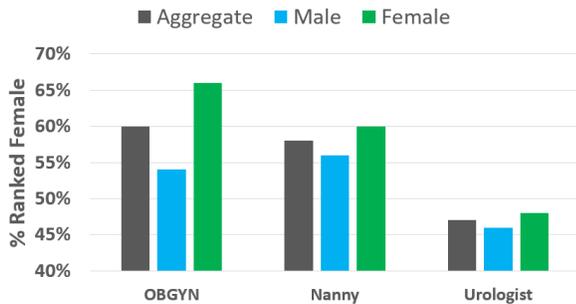


Figure 7: Professions where female decision-makers primarily drove the overall biased outcomes.

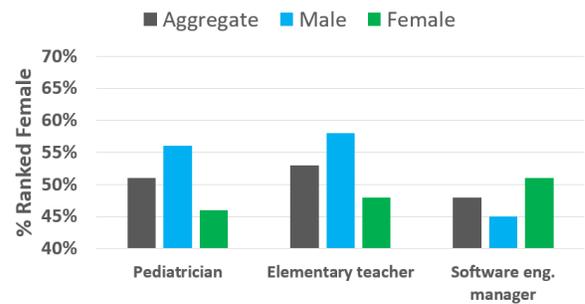


Figure 8: Professions where unbiased outcomes are cancelled out by opposing decision-making genders.

ple, while we observe significant male bias for male urologists and dermatologists on the 75F task, almost none of the justifications indicate that gender is influential on the decision. Instead, factors such as “experience” and “education” are stated, even though these are controlled variables and random assignment of gender and order to each profile for each HIT eliminates the possibility that biases can be linked to factors other than gender. This may be due to the phenomenon that humans often have a difficult time justifying their decisions (or were not aware of their own biases) and instead look retroactively for explanations to rationalize their choices (Kleinberg et al. 2019).

Discussion

This project seeks to understand how different types of bias in hybrid systems impact final decision outcomes. While bias is defined in a pragmatic statistical manner in this paper, we emphasize the need to integrate social and political notions into broader discussions on algorithmic fairness.

Experimental Validity As with all mTurk experiments, there are concerns of experimental validity. First, hiring is a

complex task. Although we detail a breakdown of the workflow governing how a system for screening candidates might work, we undoubtedly failed to capture other important variables, including other forms of bias (e.g., insider referrals of privileged candidates). Second, while we pre-screen for high-quality mTurk workers and exclude results where participants spend too little time, there is still error from participants who did not afford the task their undivided attention. Third, we do not study a real AI system that algorithmically recommends candidates. Future work that studies bias in real world decisions and investigates other confounding factors should be explored to complement these results (e.g. how much more expertise must a female software engineer have to be selected equally frequently in a screening process?). It is also important to recognize that some of our interventions would not be possible in the real world because of extreme availability skews of certain genders in the hiring pool. For example, we generate candidate slates that are 50% female for plumbers even though the world distribution is 3% female. This remains a pipeline problem. Future work that studies more nuanced and realistic representation criteria is needed.

Profession	Human Select 4	Human Select 2	Human Select 1
Urologist	$\chi^2 = 15.02$ (0.005)	$\chi^2 = 0.78$ (0.678)	$\chi^2 = 0.36$ (0.549)
Dermatologist	$\chi^2 = 12.66$ (0.013)	$\chi^2 = 0.87$ (0.647)	$\chi^2 = 1.00$ (0.317)
OBGYN	$\chi^2 = 154.10$ ($p < 0.000$)	$\chi^2 = 20.66$ ($p < 0.000$)	$\chi^2 = 3.24$ (0.072)
Nanny	$\chi^2 = 33.18$ ($p < 0.000$)	$\chi^2 = 18.51$ ($p < 0.000$)	$\chi^2 = 0.36$ (0.549)

Table 2: Chi-Square test statistics and p-values for top 1, 2, and 4-ranked candidates.

Data Segregation We do not study the impact of representation criteria at the individual level but rather as aggregate groups. It is important to know how collective groups behave, but this does not give us a granular level understanding of how representation criteria impacts individual decision-making. As seen in Observations 3a and 3b, important signals can be lost when studying aggregate outcomes. We slice the data by the gender of the decision-maker, but recognize there are many other personal features that can be analyzed such as age, race, and occupation. Future work is needed to study other confounding factors. Informed segregation of data is important if different forms of intervention will be needed for different groups of decision-makers or even individuals who might exhibit unique biases.

Choosing for Self vs. Others We ask participants to recommend candidates for a friend with the intention of creating a separation between the personal preferences of the human vs. review of the qualifications of each candidate. However, the processes that govern if an individual were to select for themselves may differ and should be separately assessed.

Binary Gender We only study male and female candidates. The challenges for analyses of non-binary gender are two-fold. First, in studying the personal features of the decision-maker, we were unable to achieve the requisite number of study participants that would identify as non-binary. Second, it is not clear how to assign a non-binary gender to profiles that are textual, for the signal of “they” may be interpreted in multiple ways, such as a data anonymization strategy. We recognize the troubling trend present in the algorithmic fairness literature on the lack of available work studying gender as a non-binary variable and hope that future work integrates this consideration.

Conclusion

Our study begins to decompose biases in decision-making by separating effects of world bias from the human decision-making process itself. Experiments across professions with varying gender proportions show that balancing the candidate list can correct for some professions where the world distribution is extremely skewed, although it has no impact on other professions where persistent preferences are at play. We show that the gender of the decision-maker, complexity of the decision-making task and over- and under-representation of genders in the candidate list can have con-

founding effects on final decisions. We believe that such a method of study, where bias mitigation is not merely examined from an algorithmic perspective in isolation but also from a hybrid system perspective, is of utmost importance as we continue to integrate automated decision systems within existing human processes. We hope this work serves as a foundation for a more nuanced understanding of bias and how to build systems for effective mitigation.

Acknowledgements

We would like to thank Adam Kalai for his wisdom on word embeddings, Krishnaram Kenthapadi for feedback on algorithmic hiring tools, Mary Gray and the MSR ethics and privacy team for IRB and standards support, and Jacki O’Neill, Luke Stark, and Ben Fish for helpful discussions.

References

- Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias: there’s software used across the country to predict future criminals, and it’s biased against blacks. *ProPublica*.
- Association of American Medical Colleges, N. 2018. Physician specialty report.
- Bertrand, M., and Mullainathan, S. 2003. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *The National Bureau of Economic Research*.
- Bolukbasi, T.; Chang, K.-W.; Zou, J.; Saligrama, V.; and Kalai, A. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 2016 Conference on Neural Information Processing Systems (NIPS 2016)*. NIPS.
- Bureau of Labor Statistics, N. 2019. Labor force statistics from the current population survey.
- Celis, L. E.; Straszak, D.; and Vishnoi, N. 2018. Ranking with fairness constraints. In *Proceedings of the 2018 International Colloquium on Automata, Languages and Programming (ICALP 2018)*. LIPIcs.
- Chen, L.; Ma, R.; Hannak, A.; and Wilson, C. 2018. Investigating the impact of gender on rank in resume search engines. In *Proceedings of the 2018 Conference on Human Factors in Computing Systems (CHI 2018)*. ACM.
- Cunningham, T. 2013. Biases and implicit knowledge. *Munich Personal RePEc Archive*.
- De-Arteaga, M.; Romanov, A.; Wallach, H.; Chayes, J.; Borgs, C.; Chouldechova, A.; Geyik, S.; Kenthapadi, K.; and Kalai, A. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (FAT* 2019)*. ACM.
- Dwork, C., and Ilvento, C. 2018. Group fairness under composition. In *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency (FAT* 2018)*. ACM.
- Fischhoff, B., and Beyth, R. 1975. I knew it would happen: remembered probabilities of once-future things. *Organizational Behavior and Human Performance* 13:1–16.

- Gannes, L. 2015. Textio spell checks for gender bias.
- Geyik, S.; Ambler, S.; and Kenthapadi, K. 2019. Fairness-aware ranking in search & recommendation systems with application to LinkedIn talent search. In *Proceedings of the 2019 Conference on Knowledge Discovery and Data Mining (KDD 2019)*. ACM.
- Glascher, J.; Daw, N.; Dayan, P.; and O’Doherty, J. 2011. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *66:585–595*.
- Green, B., and Chen, Y. 2019. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (FAT* 2019)*, 90–99. ACM.
- Hannak, A.; Wagner, C.; Garcia, D.; Mislove, A.; Strohmaier, M.; and Wilson, C. 2017. Bias in on-line freelance marketplaces: evidence from taskrabbit and fiverr. In *Proceedings of the 2017 Conference on Computer-Supported Cooperative Work and Social Computing (CSCW 2017)*. ACM.
- HireVue. 2018. Video interview software for recruiting & hiring.
- Hoffmann, A., and Musch, J. 2019. Prejudice against women leaders: insights from an indirect questioning approach. *Sex Roles*.
- Isaac, C.; Lee, B.; and Carnes, M. 2009. Interventions that affect gender bias in hiring: a systematic review. *U.S. National Library of Medicine* 84:1440–1446.
- Kahneman, D. 2003. A perspective on judgment and choice. *American Psychologist* 58:697–720.
- Kang, S.; DeCelles, K.; Tilcsik, A.; and Jun, S. 2017. Whitened resumes: race and self-presentation in the labor market. *Administrative Science Quarterly*.
- Kleinberg, J.; Ludwig, J.; Mullainathan, S.; and Sunstein, C. R. 2019. Discrimination in the age of algorithms. *SSRN*.
- Koch, A.; D’Mello, S.; and Sackett, P. 2015. A meta-analysis of gender stereotypes and bias in experimental simulations of employment decision making. *Applied Psychology* 100:128–161.
- Lambrecht, A., and Tucker, C. 2017. Algorithmic bias? an empirical study into apparent gender-based discrimination in the display of stem career ads. *Social Science Research Network*.
- Lunden, I. 2019. Google takes hire, its g suite recruitment platform, to its first global markets, uk and canada.
- Masser, B., and Abrams, D. 2004. Reinforcing the glass ceiling: the consequences of hostile sexism for female managerial candidates. *Sex Roles* 51:609–615.
- McInnis, B.; Cosley, D.; Chaebong, N.; and Leshed, G. 2016. Taking a hit: designing around rejection, mistrust, risk, and workers’ experiences in amazon mechanical turk. In *Proceedings of the 2016 Conference on Human Factors in Computing Systems (CHI 2016)*. ACM.
- Mikolov, T.; Yih, S. W.-T.; and Zweig, G. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*. ACL.
- Platt, M., and Huettel, S. 2008. Risky business: the neuroeconomics of decision making under uncertainty. *Nature Neuroscience* 11:398–403.
- Raji, I., and Buolamwini, J. 2019. Actionable auditing: investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 Conference on Artificial Intelligence, Ethics, and Society (AIES 2019)*. AAAI/ACM.
- Rangel, A. 2009. The neuroeconomics of simple goal-directed choices. *Frontiers in Computational Neuroscience* 3.
- Raub, M. 2018. Bots, bias and big data: artificial intelligence, algorithmic bias and disparate impact liability in hiring practices. *Arkansas Law Review* 71.
- Schwab, K.; Samans, R.; Zahidi, S.; Leopold, T. A.; Ratcheva, V.; Hausmann, R.; and Tyson, L. D. 2017. The global gender gap report.
- Steinpreis, R.; Anders, K.; and Ritzke, D. 1999. The impact of gender on the review of the curricula vitae of job applicants and tenure candidates: a national empirical study. *Sex Roles*.
- Thaler, R., and Sunstein, C. 2008. *Nudge: improving decisions about health, wealth, and happiness*. Penguin Books.
- The Guardian, N. 2018. Amazon ditched ai recruiting tool that favored men for technical jobs.
- Thebault-Spieker, J.; Kluver, D.; Klein, M.; Halfaker, A.; Hecht, B.; Terveen, L.; and Konstan, J. 2017. Simulation experiments on (the absence of) ratings bias in reputation systems. In *Proceedings of the 2017 Conference on Computer-Supported Cooperative Work and Social Computing (CSCW 2017)*. ACM.
- Tobin, S. 2017. World employment social outlook: trends for women 2017.
- Tversky, A., and Kahneman, D. 1973. Availability: a heuristic for judging frequency and probability. *Cognitive Psychology* 5:207–232.
- Tversky, A., and Kahneman, D. 1974. Judgment under uncertainty: heuristics and biases. *Science* 185:1124–1131.
- Veale, M.; Van Kleek, M.; and Binns, R. 2018. Fairness and accountability design needs for algorithmic support in high stakes public sector decision-making. In *Proceedings of the 2018 Conference on Human Factors in Computing Systems (CHI 2018)*. ACM.
- Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2017. Men also like shopping: reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*. ACL.