Manuela Rodriguez

Faculty Advisor: Andrea Bonezzi

Equity vs. Equality in Automated Decision-Making Systems

Introduction

An automated decision-making system or "algorithm" is a system that takes data as an input and makes decisions based on statistical or mathematical models and without direct human intervention (Kyung, 2018). Algorithms have become a popular and controversial topic among the public and the media. Today, automated decision-making can be seen across a wide range of domains like the criminal justice system, where algorithms are used to compute a "risk score" that determines the likelihood of a person awaiting trial of committing a further crime; college admissions, where candidates are evaluated by machines based on a set of criteria like standardized test scores, high school grades, and predicted scores about the likelihood of dropping out or earning low grades; and HR departments, where resumes and applications are screened by a machine learning algorithm before getting reviewed by recruiters.

The increasingly widespread use of automated decision-making systems has raised serious concerns about the biases inherent in algorithms; more specifically, how these biases disproportionately affect historically underrepresented and marginalized groups and how they take human biases to a whole new level by not only reproducing them but also amplifying them on an unprecedented scale.

Algorithmic Fairness

Artificial intelligence algorithms learn from the data they are trained on, data about previous decisions made by humans, which is very likely to reflect prejudices and stereotypes that exist in human institutions. For example, an algorithm that is designed to process natural language will learn about cultural associations of the words, which can reflect stereotypes and be "offensive, objectionable or harmful." (Caliskan et al., 2017). A study conducted in 2017 by a team of researchers at Princeton University concluded that natural language processing algorithms inherit human biases. For example, researchers found that a widely used sentiment analysis technique, which classifies text as "positive," "negative," or "neutral," when applied to movie reviews gave a higher sentiment score to a sentence containing a European-American name than to the same sentence containing an African American name. Similarly, it found that when Google Translate converts Turkish genderless sentences to English sentences, it associates the word "doctor" more closely with the pronoun "he" than "she," and the word "nurse" more closely with the pronoun "she." These cases are clear examples of how algorithms can reproduce both racial and gender disparities. (Caliskan et al., 2017).

The biggest concern about algorithms inheriting human biases from the data they are trained on is the extent to which these biases can be amplified when enormous amounts of data are processes in a very short period and when there is so little transparency about the decisions being made by the algorithm at every step of the process. A study conducted by ProPublica (2016) concluded that risk scores are injecting bias into the U.S. court system by reproducing racial bias. The study found that an algorithm being used by multiple courts was more likely to flag Black defendants as future criminals at almost twice the rate as White defendants and mislabeled White defendants as "low risk" more often than Black defendants. Historically, the police are more likely to patrol low-income neighborhoods that have larger Black populations,

which means that people in these neighborhoods are more likely to get arrested. Therefore, based solely on race and regardless of criminal history, algorithms learn from the data they are being trained on that a Black person is more likely to get arrested than a White person.

The law protects certain features like race and gender by forbidding them from being used in decision-making to protect individuals from being discriminated against for belonging to a certain group. Therefore, algorithms cannot use these protected classes as predictors. For example, the Equal Credit Opportunity Act (ECOA) forbids race, religion, age, among other factors from being considered in credit decisions, and the Fair Housing Act (FHA) prohibits discrimination in real estate financing on the basis of race, religion, sex, and other factors (Gillis, 2019). The traditional anti-discrimination approach has been to limit the input that goes into the decision-making process, which is easier to do if the decision-maker is an artificial intelligence software rather than a human because in the case of many protected variables, if the decision-maker is a human, they can visually tell what group the individual belongs to. However, multiple studies have found that excluding these features from the predictors does not protect individuals against discrimination given that protected classes can be reconstructed from other features (Kleinberg, 2018). For example, race is strongly correlated with job status and income, which are not protected by law. Thus, despite race not being used as a predictor, the algorithm can still infer the race of the person from these and other correlated variables.

Previous studies suggest that including protected variables can help reduce bias and disparities in decision-making. A study conducted at the University of Chicago by Gillis and Spiess (2019) on automated credit pricing found that restricting inputs, that is, excluding protected variables like race and age from credit pricing does not guarantee equal pricing and can in fact lead to greater pricing disparities between protected groups. There are other non-protected

variables that reflect historical biases and can widen the gap between protected groups. For example, income can be a strong predictor of race and thus, can increase racial disparities when used to predict credit scores. When these biased variables are present, using forbidden variables as inputs can be a tool to correct for these biases, which raises the question of whether the algorithm should be made aware of protected classes to reduce bias and disparities.

The Principles of Equity and Equality

Algorithms that treat everyone equally are based on the principle of equality. For example, in the criminal justice system, such an algorithm would use only one threshold (e.g. a risk score of 6) to determine whether a defendant is "low risk" (if the risk score is smaller than or equal to 6) or "high risk" (if the risk score is greater than 6). But an algorithm can also be designed to use different thresholds for different groups of people. So, instead of using a risk score of 6 to classify every individual as low or high risk, the risk score threshold would be adjusted depending on the person's race. For example, a score of 8 could be used for Black people, who are more likely to receive a higher risk score regardless of their criminal history, as opposed to a score of 6 for White people. Such an algorithm would be based on the principle of equity, where the algorithm is made aware of protected classes (e.g. race and gender) upon which individuals are knowingly discriminated against to level the playing field among groups by evaluating them differently based on their characteristics and circumstances. The principle of equity can be applied to both human and automated decision-making. For example, consider a college admissions officer reviewing a candidate's application. The admissions officer is made aware of the fact that the candidate grew up in a low-income neighborhood and had no access to SAT preparation, so, intuitively, the officer knows that the candidate's SAT score is more likely to be lower than the one of a candidate who went to a private school and received top-tier SAT

preparation. Therefore, the admissions officer might adjust the SAT threshold for that candidate, requiring a score that is lower than the average. This raises the question of whether protected classes should be used as inputs and whether algorithms should be made aware of an individual's unique and relevant characteristics and circumstances for a more equitable outcome. Intuitively, one would think that excluding sensitive variables from an algorithm's analysis would yield a fairer outcome. Yet, research has shown that these anti-discrimination measures that limit the inputs do not lead to significant reductions in disparities between protected groups (Gillis, 2019) and that in order to be able to account for bias, the algorithm must know the protected group the individual belongs to (Kleinberg, 2018). However, the general public, and especially those belonging to minority groups that have been disproportionately affected by bias in decision-making across different domains, might not feel completely comfortable knowing that a machine is using these protected classes to make predictions and decisions that can be life-changing, which is why it is important to not only understand how the principle of equity can reduce bias and yield fairer outcomes from a technical perspective, but also understand how people perceive fairness of the results yielded by these systems and how much they trust them.

Fairness Perceptions

There is a need to understand how the public perceives fairness in automated decision-making systems as opposed to human decision-making systems. Although individual perceived fairness is very subjective and is different from the factual fairness of the algorithm, it is as important as the latter. Even if an algorithm is efficient and has been determined to be fair by certain standards, if people believe they are being treated unfairly, they can become distrustful of the institution that uses it (Marcinkowski et al., 2020). A study conducted in 2017 about why people voluntarily left tech jobs found that the single largest driver of employee turnover was a

perceived lack of fairness, which costs the industry $16 billion per year (Scott et al., 2017; Newman et al., 2020). Studies have found that people become distrustful of algorithms after finding out that they make mistakes. Researchers use the term "algorithmic aversion" to describe this phenomenon where people trust humans more than algorithms when they see algorithms making mistakes even when they yield more reliable and accurate results than humans (Wang et al., 2020). On the other hand, previous research has found that perceptions of injustice and disadvantage can lead people to take part in collective action against the entity believed to be yielding this injustice. A study about participation in collective action found that when people blame injustice and inequalities on society and institutions, they are likely to take collective action as a means of correcting these disadvantages (Corcoran, 2015; Bonezzi, 2021). With that in mind, enhancing the perceived fairness of automated decision-making systems is key to their broader acceptance across different domains that involve important societal institutions like court systems, universities, and corporations.

Previous research has focused on understanding the perceived fairness of automated decision-making in different domains like college admissions and the criminal justice system, and the consequences of perceived inequality and injustice, but there is still a need to understand how perceived fairness changes and how it differs from human decision-making when the decision-making process is based on equity. The present research addresses the question of whether automated decision-making based on equity rather than equality is perceived to be fairer or more unfair than human decision-making based on the same principle.

Hypothesis 1: Decisions made by algorithms based on equity are perceived as more unfair than decisions made by humans based on the same principle.

Hypothesis 2: Decisions made by algorithms based on equality are perceived as fairer than decisions made by humans based on the same principle.

Evaluating individuals differently based on their unique characteristics or circumstances requires a more holistic approach to decision-making than evaluating everyone equally. A recent study found that when people believe contextualization is important in the decision-making process, they tend to perceive decisions made by humans as fairer than those made by algorithms. This is due to the fact that people associate algorithms with reductionism and decontextualization, which "limits the use of accurate information on which fair procedures must rely," leading them to believe that humans make fairer decisions (Newman et al., 2020).

Hypothesis 3: Automated decision-making based on equity is perceived as more unfair than human decision-making because it is believed that algorithms decontextualize decision-making.

Previous research suggests that fairness perceptions of decision-making systems depend on the nature of the task that is being carried out. A study by a researcher at Carnegie Mellon University found that mechanical tasks performed by machines are perceived as equally fair to those performed by humans. However, human tasks that require soft skills are perceived to be fairer when performed by humans rather than machines because people believe that algorithms lack intuition and subjective judgment capabilities, so their judgments are less trustworthy (Lee, 2018). The study evaluated how people perceived fairness in four managerial decisions and found that "positive emotion from human decisions was attributed to social recognition, while negative emotion from algorithmic decisions was attributed to the dehumanizing experience of being evaluated by machines." Accordingly, if people perceive the task of evaluating individuals differently based on the group they belong to (e.g. race and gender) as more of a human than a

mechanical task and one that requires soft skills and intuition, then it is very likely that the public will perceive decisions made by algorithms based on the principle of equity as more unfair than decisions made by humans based on this same principle.

Hypothesis 4: People believe that algorithms lack empathy and intuition, therefore, they are less likely than humans to make a fair decision when applying different criteria across individuals based on the group they belong to and their individual characteristics.

Overview of the Studies

The study we conducted examines how people perceive fairness in decision-making systems; in particular, how perceived fairness changes depending on the decision-making principle (equity versus equality) and who the decision-maker is (human versus algorithm). The study tests the hypotheses that people believe algorithms yield fairer results than humans when decisions are made based on the principle of equality and more unfair results when decisions are made based on the principle of equity.

Methodology

Eight hundred seventy-four respondents (46.6% women; age: M = 40.7, SD=12.5) recruited from Amazon Mechanical Turk (MTurk) were randomly assigned to a 2 (decision-maker: human, algorithm) x 2 (decision-making principle: equality, equity) between-subjects design.

Respondents read a report about a university where admission decisions were based on an evaluation of applicants conducted by either an admission team (human) or an artificial intelligence software (algorithm). To ensure that respondents understood the information presented, they were asked to indicate who or what conducted the evaluation of the applicants

before moving on to the next screen. Next, respondents read that the decision was made based on either the principle of equality or equity. In the equality condition, participants were informed that the decision-maker applied the same standard across all candidates and were then presented with statistics about the university's acceptance rates broken down by ethnicity. The results showed a clear racial disparity; 47 out of 100 White applicants were accepted, while only 15 out of 100 Black applicants and 15 out of 100 Hispanic applicants were accepted. In the equity condition, participants were informed that the decision-maker applied different criteria across applicants based on their race to achieve comparable rates of acceptance across races and were then presented with statistics about the university's acceptance rates broken down by ethnicity. The results showed a consistent acceptance rate across ethnicities; 30 out of 100 White applicants were accepted, 32 out of 100 Black applicants were accepted and 32 out of 100 Hispanic applicants were accepted.

Respondents answered three items that measured the perceived fairness of the decisions made by either the admission team (human) or the artificial intelligence software (algorithm; 1 = very unfair/ very unacceptable/ very outrageous; 7 = very fair/ very unacceptable/ very outrageous). These items were averaged to create an index of perceived fairness (α=.97). Next, respondents answered three questions that measured the extent to which they agreed (1 = completely disagree; 7 = completely disagree) with the following statements: the admission team/ algorithm lacks the ability to consider an applicant's unique circumstances and characteristics, the admission team/ algorithm lacks empathy and the admission team/ algorithm can be biased. Finally, respondents reported their gender and age and were debriefed (see Appendix for details).

Results

A 2×2 ANOVA on perceived fairness of the decision revealed a significant main effect of decision-making principle ($F(1,809) = 26.594$, $p < .001$, $\eta_p^2 = .03$), a non-significant main effect of decision-maker ($F(1,809) = .115$), and a significant decision-making principle × decision-maker interaction ($F(1,809) = 4.754$, $p = .03$, $\eta_p^2 = .01$). As predicted, when the decision-making principle was equality, respondents considered the decision more fair when it was made by an algorithm ($M = 4.05$, $SD = 1.88$) rather than by a human ($M = 3.74$, $SD = 1.77$, $F(1,809) = 3.17$, $p = .075$, $\eta_p^2 = .004$). This effect was marginally significant. In contrast, when the decision-making principle was equity, respondents considered the decision less fair when it was made by an algorithm ($M = 4.43$, $SD = 1.79$) rather than by a human ($M = 4.66$, $SD = 1.74$, $F(1,809) = 1.697$, $p = .193$, $\eta_p^2 = .002$). Although this effect did not reach conventional levels of statistical significance, it was in the predicted direction.

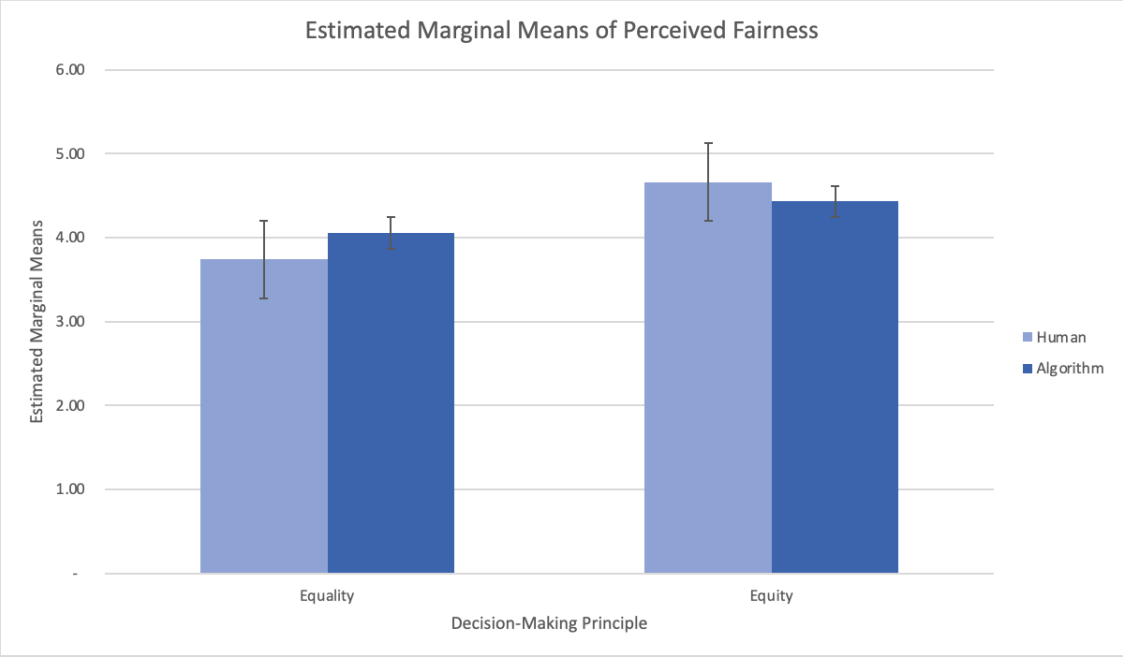

Table 1: Descriptive Statistics

Dependent Variable:

Decision-maker		Mean	Std. Deviation	N
	Equality	3.737	1.765	203
	Equity	4.660	1.742	204
Human	**Total**	4.200	1.811	407
	Equality	4.054	1.876	203
	Equity	4.429	1.786	203
Algorithm	**Total**	4.241	1.839	406
	Equality	3.896	1.826	406
	Equity	4.545	1.766	407
Total	**Total**	4.221	1.824	813

Table 2: Tests of Between-Subjects Effects

Dependent Variable:

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Powerb
Corrected Model	101.23	3	33.74	10.49	0.000	0.037	31.48	0.999
Intercept	14478.44	1	14478.44	4502.93	0.000	0.848	4502.93	1.000
Agent	0.37	1	0.37	0.12	0.734	0.000	0.12	0.063
Criterion	85.51	1	85.51	26.59	0.000	0.032	26.59	0.999
Agent * Criterion	15.29	1	15.29	4.75	0.030	0.006	4.75	0.586
Error	2601.21	809	3.22					
Total	17184.67	813						
Corrected Total	2702.44	812						

a. R Squared = .037 (Adjusted R Squared = .034)

b. Computed using alpha = .05

Table 3: Univariate Tests

Dependent Variable:

Criterion		Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Powera
	Contrast	10.194	1	10.194	3.1704	0.075357	0.003904	3.17044	0.428078
Equality	Error	2601.21	809	3.2153					
	Contrast	5.45577	1	5.4558	1.6968	0.193078	0.002093	1.6968	0.255535
Equity	Error	2601.21	809	3.2153					

Each F tests the simple effects of Agent within each level combination of the other effects shown. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.
a. Computed using alpha = .05

Table 4: Univariate Tests

Dependent Variable:

Agent		Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Powera
Human	Contrast	86.6563	1	86.656	26.951	2.64E-07	0.03224	26.9509	0.999371
	Error	2601.21	809	3.2153					
Algorithm	Contrast	14.2266	1	14.227	4.4246	0.035733	0.005439	4.4246	0.556095
	Error	2601.21	809	3.2153					

Each F tests the simple effects of Criterion within each level combination of the other effects shown. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.
a. Computed using alpha = .05

A 2×2 ANOVA on perceived bias of the decision-maker revealed a significant decision-making principle × decision-maker interaction ($F(1,809) = 89.377$, $p < .001$, $\eta_p^2 = .03$), a non-significant main effect of decision-maker ($F(1,809) = .757$), and a non-significant main effect of decision-making principle ($F(1,809) = 5.173$). As predicted, when the decision-making principle was equality, respondents considered the human ($M = 4.73$, $SD = 1.895$) more biased than the algorithm ($M = 4.01$, $SD = 2.28$; $F(1,809) = 12.394$, $p < .001$, $\eta_p^2 = .02$). This effect was statistically significant. In contrast, when the decision-making principle was equity, respondents considered the algorithm ($M = 4.83$, $SD = 2.049$) more biased than the human ($M = 4.23$, $SD =$ 2,046; $F(1,809) = 36.889$, $p < .003$, $\eta_p^2 = .01$). This effect was statistically significant.

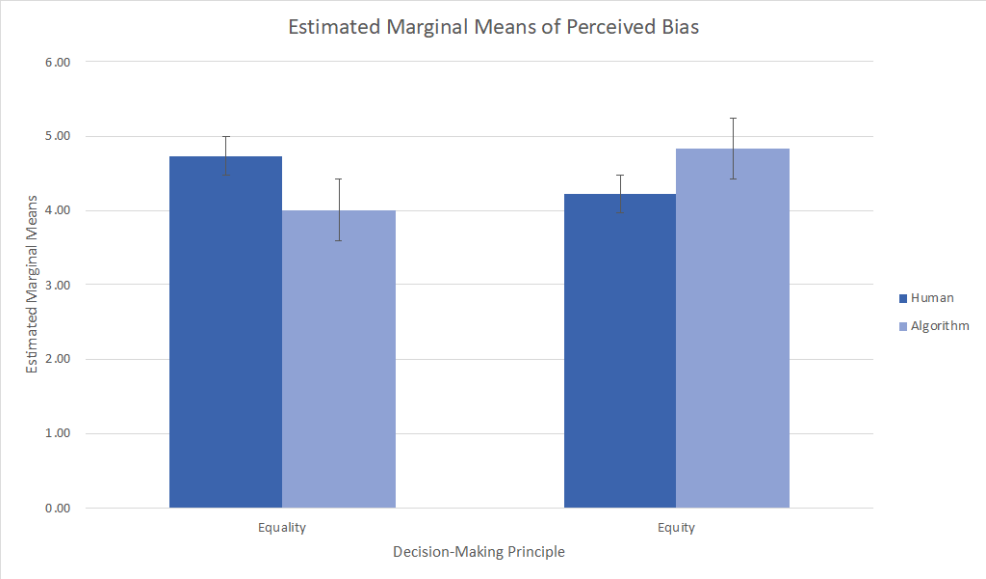

Table 5: Descriptive Statistics

Dependent Variable: Agent		Mean	Std. Deviation	N
	Equality	4.734	1.895	203
	Equity	4.230	2.046	204
Human	Total	4.482	1.986	407
	Equality	4.010	2.280	203
	Equity	4.833	2.049	203
Algorithm	Total	4.421	2.204	406
	Equality	4.372	2.125	406
	Equity	4.531	2.067	407
Total	Total	4.451	2.097	813

Table 6: Tests of Between-Subjects Effects

Dependent Variable:

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Powerb
Corrected Model	95.238	3	31.7460	7.393	0.000	0.027	22.178	0.985
Intercept	16111.566	1	16111.5664	3751.845	0.000	0.823	3751.845	1.000
Agent	0.757	1	0.7565	0.176	0.675	0.000	0.176	0.070
Criterion	5.173	1	5.1727	1.205	0.273	0.001	1.205	0.195
Agent * Criterion	89.377	1	89.3768	20.813	0.000	0.025	20.813	0.995

Error	3474.093	809	4.2943
Total	19679.000	813	
Corrected Total	3569.331	812	

a. R Squared = .027 (Adjusted R Squared = .023)
b. Computed using alpha = .05

Table 7: Univariate Tests

Dependent Variable: Can be biased by an applicant's race

Criterion		Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[a]
	Contrast	53.2241	1	53.224	12.394	0.0005	0.015	12.394	0.940
Equality	Error	3474.09	809	4.2943					
	Contrast	36.8891	1	36.889	8.5902	0.0035	0.011	8.590	0.833
Equity	Error	3474.09	809	4.2943					

Each F tests the simple effects of Agent within each level combination of the other effects shown. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.
a. Computed using alpha = .05

Table 8: Univariate Tests

Dependent Variable: Can be biased by an applicant's race

Agent		Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[a]
	Contrast	25.8048	1	25.805	6.0091	0.0144	0.0074	6.009	0.687
Human	Error	3474.09	809	4.2943					
	Contrast	68.6921	1	68.692	15.996	0.0001	0.0194	15.996	0.979
Algorithm	Error	3474.09	809	4.2943					

Each F tests the simple effects of Criterion within each level combination of the other effects shown. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.
a. Computed using alpha = .05

Discussion

Overall, the study shows that when decisions are made based on the principle of equality, the results yielded by algorithms are perceived to be fairer than the results yielded by humans,

while when decisions are made based on the principle of equity, the results yielded by algorithms are perceived to be more unfair than the results yielded by humans. Our findings suggest that perceptional bias of algorithms and humans changes based on the decision-making principle. In the equality principle, when the decision-maker applies the same criteria to all individuals, algorithms are perceived as less biased than humans. In contrast, in the equity principle, when the decision-maker applies different criteria across different groups of people, humans are perceived as less biased than algorithms.

General Discussion

The results of the study provide preliminary evidence that fairness perceptions of the decision-maker shift based on the decision-making principle. That is, when the decision-making principle is equality, people perceive decisions yielded by algorithms as fairer than those yielded by humans, but the opposite happens when the decision-making principle is equity—decisions yielded by humans are perceived as fairer. It should be noted that in the equity principle, the statistics presented to the respondents showed consistency across applicants of different ethnicities, and yet, people perceived the decisions made by the algorithm as more unfair than the decisions made by the admission team (human).

The results suggest that equity is a second-order equality, and because algorithms lack contextualization and flexibility, people tend to trust humans more than algorithms when decisions are made based on the principle of equity. To explain this idea further, in the equality principle, where every individual is subject to the same standards, people might perceive algorithms as fairer decision-makers because unlike humans, they are technically completely blind to ethnicity, gender or any other protected variable that is physically visible to human

decision-makers. For example, in a hiring process, even if the interviewer is not told what the gender, age or ethnicity of the interviewee is, they will make assumptions based solely on their physical appearance. Therefore, it is reasonable that people assume algorithms make more objective decisions and do not have the unconscious biases of humans. Our results corroborate this; when respondents were asked to rate the extent to which they agreed with the statement: "the algorithm/ human can be biased," in the equality principle, respondents agreed with the statement to a greater extent when the decision-maker was a human than when it was an algorithm. The opposite happened in the equity principle: respondents rated the algorithm as more likely to be biased than the human. In this principle, the algorithm is in fact actively discriminating (justifiably) since it is applying different criteria to different groups of applicants. However, it is not acknowledging the fact that there is heterogeneity that is not accounted for within the protected groups—evidently, not every individual of the same ethnicity is equal. To illustrate this, in the context of college admissions, SAT score thresholds would be higher for White applicants than other ethnic minorities because historically, the data has shown that White applicants have higher average scores and get accepted at disproportionally higher rates than Black or Hispanic applicants. However, that doesn't mean that all White applicants had access to top-tier preparatory resources while all Black or Hispanic applicants didn't. Yet, the algorithm will treat every White applicant equally without accounting for other possible factors that could have an impact on how well they perform on their SAT tests.

With that in mind, a human decision-maker has more flexibility to account for the heterogeneity within the protected groups. A human decision-maker is able to contextualize decision-making and consider other factors to make a fairer decision. The lack of flexibility and contextualization could be a potential explanation for why people believe humans yield fairer

results than algorithms when the decision is made based on the principle of equity, since the algorithm is technically discriminating against certain groups of individuals and is not able to account for the differences within the individuals belonging to that group. This is also corroborated by the results of the first part of the study, which show that respondents thought the results yielded by humans in the equity principle were fairer than the results yielded by the algorithm.

Limitation and Further Research

In the study, we asked respondents to rate to what extent they agreed with the following three statements: “The decision-maker lacks the ability to consider an applicant's unique circumstances and characteristics,” “The decision-maker lacks empathy,” and “The decision-maker can be biased by an applicant’s race,” where decision maker could be either the admission team (human) or an artificial intelligence software (algorithm). Although we didn’t get statistically significant results to support hypotheses 3 and 4 about why people perceive results yielded by humans as fairer than algorithms in the equity principle, the direction of our results suggests that one of the reasons why people trust humans more than algorithms in the equity principle could be that algorithms are viewed as lacking flexibility and contextualization capabilities. With that in mind, further research could focus on identifying and understanding the reasons why perceived fairness shifts when the decision-making principle changes and in particular, why algorithms are viewed as more likely to be biased and yield more unfair results than humans in the equity principle.

Lastly, in the future, we could test multiple domains where there is potential for discrimination like the criminal justice system, credit pricing and HR systems to increase the

robustness and validity of our conclusions, as well as further calibrate the stimuli to see how perceived fairness changes when in the equity principle, the statistics presented to respondents show a racial disparity, or in the equality principle, the statistics presented show consistency across ethnicities.

Works Cited

Lee, M. K. 2018. "Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management". Big Data & Society. https://doi.org/10.1177/2053951718756684

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. "Machine Bias." ProPublica, May 23. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Frank Marcinkowski, Kimon Kieslich, Christopher Starke, and Marco Lünich. 2020. "Implications of AI (un-)fairness in higher education admissions: the effects of perceived AI (un-)fairness on exit, voice, and organizational reputation." In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, 122–130. https://doi.org/10.1145/3351095.3372867

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018. "Algorithmic Fairness." *AEA Papers and Proceedings*, 108: 22-27. https://doi.org/10.1257/pandp.20181018

Wang, R., Harper, F. M., & Zhu, H. 2020. "Factors influencing perceived fairness in algorithmic decision-making." *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, https://doi.org/10.1145/3313831.3376813

Newman, David T., et al. 2020. “When Eliminating Bias Isn’t Fair: Algorithmic Reductionism and Procedural Justice in Human Resource Decisions.” *Organizational Behavior and Human Decision Processes*, vol. 160, Sept. 2020, pp. 149–67. *EBSCOhost*, https://doi-org.proxy.library.nyu.edu/10.1016/j.obhdp.2020.03.008.

Scott, A., et al. 2017. “Tech Leavers Study: A first-of-its-kind analysis of why people voluntarily left jobs in tech”. *Kapor Center for Social Impact*, https://www.kaporcenter.org/tech-leavers/.

Gillis, Talia B., and Jann L. Spiess. “Big Data and Discrimination.” *The University of Chicago Law Review*, vol. 86, no. 2, 2019, pp. 459–88, https://www.jstor.org/stable/26590562. Accessed 21 Apr. 2022.

Bonezzi, Andrea, and Massimiliano Ostinelli. “Can Algorithms Legitimize Discrimination?” *JOURNAL OF EXPERIMENTAL PSYCHOLOGY-APPLIED*, vol. 27, no. 2, June 2021, pp. 447–59. *EBSCOhost*, https://doi-org.proxy.library.nyu.edu/10.1037/xap0000294.

Corcoran, K. E. (. 1.)., et al. “Perceptions of Structural Injustice and Efficacy: Participation in Low/Moderate/High-Cost Forms of Collective Action.” *Sociological Inquiry*, vol. 85, no. 3, pp. 429–61. *EBSCOhost*, https://doi-org.proxy.library.nyu.edu/10.1111/soin.12082. Accessed 21 Apr. 2022.

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. "Semantics derived automatically from language corpora contain human-like biases." Science 356.6334 (2017): 183-186.

Appendix

Consent

ONLINE INFORMED CONSENT (IRB-FY2021-6171)

You have been invited to take part in a research study to learn more about interactions with technology. If you agree to be in this study, you will be asked to do the following: You will be presented with a decision-making scenario and asked some questions about it. There are no known risks associated with your participation in this research beyond those of everyday life. Although you will receive no direct benefits, this research may help improve our understanding of consumer interactions with technology. You will receive payment for completing the online survey; if you withdraw before the end of the study, no payment will be given. Confidentiality of your research records will be strictly maintained: Amazon Mechanical Turk Worker Identification will be collected for completion confirmation only and will afterwards be removed from the dataset. Then, the data set used for analysis will only contain anonymous ID numbers that cannot be connected with a particular person. Only aggregate statistics will be published. Information not containing identifiers may be used in future research or shared with other researchers without your additional consent. Data will be stored on password protected devices. Participation in this study is voluntary. Should you choose to participate in this study, you may refuse to participate or withdraw at any time without penalty. For interviews, questionnaires, or surveys, you have the right to skip or not answer any questions you prefer not to answer. If you agree to take part in the research and allow the researchers to use your responses for her research, please select the button below.

I CONSENT TO PARTICIPATE and consent to the use of the collected data for current and future research

I DO NOT CONSENT TO PARTICIPATE or consent to the use of the collected data for current and future research

Intro

Please read the information on the next pages carefully.
There are no right or wrong answers, we are simply interested in your opinion.

Human - Equality

A University recently published a document about its admission process, along with some statistics about the acceptance rate for different groups of applicants. The document shows that admission decisions are based on the results of an analysis of the applicants conducted by an admission team.

Who conducts the analysis of the applicants? _____

The documents shows that the admission team applies the same criteria across applicants. That is to say, to be admitted by the university all applicants are subject to the same standards.

The document also shows the acceptance rates (the rate at which applicants are accepted by the University) by different races:

- White applicants: 47% (47 out of 100 White applicants are accepted)
- Black applicants: 15% (15 out of 100 Black applicants are accepted)
- Hispanic applicants: 15% (15 out of 100 Hispanic applicants are accepted)

The documents shows that the admission team applies the same criteria across applicants. That is to say, to be admitted by the university all applicants are subject to the same standards.

The document also shows the acceptance rates (the rate at which applicants are accepted by the University) by different races:

- White applicants: 47% (47 out of 100 White applicants are accepted)
- Black applicants: 15% (15 out of 100 Back applicants are accepted)
- Hispanic applicants: 15% (15 out of 100 Hispanic applicants are accepted)

In your opinion, the decisions made by the admission team are:

1 = Very Unfair; 7 = Very Fair

In your opinion, the decisions made by the admission team are:

1 = Completely unacceptable, 7 = Completely acceptable

In your opinion, the decisions made by the admission team are:

1 = Very outrageous, 7 = Very commendable

To what extent do you agree with the following statements?

The admission team lacks the ability to consider an applicant's unique circumstances and characteristics.

1 = Completely disagree, 7 = Completely agree

The admission team lacks empathy.

1 = Completely disagree, 7 = Completely agree

The admission team can be biased by an applicant's race.

1 = Completely disagree, 7 = Completely agree

Human - Equity

A University recently published a document about its admission process, along with some statistics about the acceptance rate for different groups of applicants. The document shows that the admission decisions are based on the results of an analysis of the applicants conducted by an admission team.

Who conducts the analysis of the applicants? _____

The admission team applies different criteria across applicants based on the race of the applicant, to achieve comparable rates of acceptance across races. That is to say that to be admitted by the university applicants of some races are subject to higher standards than applicants of other races.

The document also shows the acceptance rates (the rate at which applicants are accepted by the University) by different races:

- White applicants: 30% (30 out of 100 White applicants are accepted)
- Black applicants: 32% (32 out of 100 Black applicants are accepted)
- Hispanic applicants: 32% (32 out of 100 Hispanic applicants are accepted)

The admission team applies different criteria across applicants based on the race of the applicant, to achieve comparable rates of acceptance across races. That is to say that to be

admitted by the university applicants of some races are subject to higher standards than applicants of other races.

The document also shows the acceptance rates (the rate at which applicants are accepted by the University) by different races:

- White applicants: 30% (30 out of 100 White applicants are accepted)
- Black applicants: 32% (32 out of 100 Black applicants are accepted)
- Hispanic applicants: 32% (32 out of 100 Hispanic applicants are accepted)

In your opinion, the decisions made by the admission team are:

1 = Very Unfair; 7 = Very Fair

In your opinion, the decisions made by the admission team are:

1 = Completely unacceptable, 7 = Completely acceptable

In your opinion, the decisions made by the admission team are:

1 = Very outrageous, 7 = Very commendable

To what extent do you agree with the following statements?

The admission team lacks the ability to consider an applicant's unique circumstances and characteristics.

1 = Completely disagree, 7 = Completely agree

The admission team lacks empathy.

1 = Completely disagree, 7 = Completely agree

The admission team can be biased by an applicant's race.

1 = Completely disagree, 7 = Completely agree

Algorithm - Equality

A University recently published a document about its admission process, along with some statistics about the acceptance rate for different groups of applicants. The document shows that the admission decisions are based on the results of an analysis of the applicants conducted by a software that uses an algorithm, that is, a set of mathematical equations.

What conducts the analysis of the applicants? _____

The document shows that the algorithm applies the same criteria across applicants. That is to say that to be admitted by the university all applicants are subject to the same standards.

The document also shows the acceptance rates (the rate at which applicants are accepted by the University) by different races:

- White applicants: 47% (47 out of 100 White applicants are accepted)
- Black applicants: 15% (15 out of 100 Back applicants are accepted)
- Hispanic applicants: 15% (15 out of 100 Hispanic applicants are accepted)

The document shows that the algorithm applies the same criteria across applicants. That is to say that to be admitted by the university all applicants are subject to the same standards.

The document also shows the acceptance rates (the rate at which applicants are accepted by the University) by different races:

- White applicants: 47% (47 out of 100 White applicants are accepted)
- Black applicants: 15% (15 out of 100 Back applicants are accepted)
- Hispanic applicants: 15% (15 out of 100 Hispanic applicants are accepted)

In your opinion, the decisions made by the algorithm are:

1 = Very Unfair; 7 = Very Fair

In your opinion, the decisions made by the algorithm team are:

1 = Completely unacceptable, 7 = Completely acceptable

In your opinion, the decisions made by the algorithm team are:

1 = Very outrageous, 7 = Very commendable

To what extent do you agree with the following statements?

The algorithm lacks the ability to consider an applicant's unique circumstances and characteristics.

1 = Completely disagree, 7 = Completely agree

The algorithm lacks empathy.

1 = Completely disagree, 7 = Completely agree

The algorithm can be biased by an applicant's race.

1 = Completely disagree, 7 = Completely agree

Algorithm - Equity

A University recently published a document about its admission process, along with some statistics about the acceptance rate for different groups of applicants. The document shows that the admission decisions are based on the results of an analysis of the applicants conducted by a software that uses an algorithm, that is, a set of mathematical equations.

What conducts the analysis of the applicants? _____

The algorithm applies different criteria across applicants, based on the race of the applicant, to achieve comparable rates of acceptance across races. That is to say that to be admitted by the university applicants of some races will be subject to higher standards than applicants of other races.

The document also shows the acceptance rates (the rate at which applicants are accepted by the University) by different races:

- White applicants: 30% (30 out of 100 White applicants are accepted)
- Black applicants: 32% (32 out of 100 Back applicants are accepted)
- Hispanic applicants: 32% (32 out of 100 Hispanic applicants are accepted)

The algorithm applies different criteria across applicants, based on the race of the applicant, to achieve comparable rates of acceptance across races. That is to say that to be admitted by the university applicants of some races will be subject to higher standards than applicants of other races.

The document also shows the acceptance rates (the rate at which applicants are accepted by the University) by different races:

- White applicants: 30% (30 out of 100 White applicants are accepted)
- Black applicants: 32% (32 out of 100 Back applicants are accepted)
- Hispanic applicants: 32% (32 out of 100 Hispanic applicants are accepted)

In your opinion, the decisions made by the algorithm are:

1 = Very Unfair; 7 = Very Fair

In your opinion, the decisions made by the algorithm team are:

1 = Completely unacceptable, 7 = Completely acceptable

In your opinion, the decisions made by the algorithm team are:

1 = Very outrageous, 7 = Very commendable

To what extent do you agree with the following statements?

The algorithm lacks the ability to consider an applicant's unique circumstances and characteristics.

1 = Completely disagree, 7 = Completely agree

The algorithm lacks empathy.

1 = Completely disagree, 7 = Completely agree

The algorithm can be biased by an applicant's race.

1 = Completely disagree, 7 = Completely agree

Demos

What is your gender?

Male

Female

Prefer not to say

What is your age? _____

Debriefing

Debriefing IRB-FY2022-6171

In this study we are interested in how fairness perceptions differ as a function of decision maker and decision principle. That is, to this end, some participants were presented with scenarios with a human as decision-maker while others were presented with the same scenarios with an algorithm as decision-maker. In order to receive participants' natural responses to the scenario without specifically directing their attention to the nature of the decision-maker, we withheld the full hypothesis. The presented scenarios are hypothetical as you have been informed prior to reading them.