Beauty's in the AI of the Beholder: How AI Anchors Subjective and Objective Predictions

Lauren Rhue
Wake Forest School of Business, larhue@gmail.com

Beauty is in the AI of the Beholder: How Artificial Intelligence Anchors Human Decisions on Subjective vs. Objective Measures

Completed Research Paper

Lauren Rhue

lrhue@rhsmith.umd.edu

Abstract

Researchers increasingly acknowledge that algorithms can exhibit bias, but artificial intelligence (AI) is increasingly integrated into the organizational decision-making process. How does biased AI shape human choices? We consider a sequential AI-human decision that mirrors organizational decisions; an automated system provides a score and then a human decides a score using their discretion. We conduct an AMT survey and ask participants to assign one of two types of scores: a subjective, context-dependent measure (Beauty) and objective, observer-independent measure (Age). Participants are either shown the AI score, shown the AI score and its error, or not shown the AI score. We find that participants without knowledge of the AI score do not exhibit bias; however, knowing the AI scores for the subjective measure induces bias in the participants' scores due to the anchoring effect. Although participants' scores do not display bias, participants who receive information about the AI error rates devalue the AI score and reduce their error.

This study makes several contributions to the information systems literature. First, this paper provides a novel way to discuss artificial intelligence bias by distinguishing between subjective and objective measures. Second, this paper highlights the potential spillover effects from algorithmic bias into human decisions. If biased artificial intelligence anchors human decisions, then it can induce bias into previously unbiased scores. Third, we examine a method to encourage participants to reduce their reliance on the artificial intelligence, reporting the error rate, and find evidence that it is effective for the objective measure.

Introduction

Artificial intelligence is used as a means to extend decision-support systems (Phillips-Wren & Jain, 2006) and is increasingly integrated into the organizational decision-making process (Shrestha, Ben-Menahem, & von Krogh, 2019). Artificial intelligence holds promise for firm processes such as human resources (Tambe, Cappelli, & Yakubovich, 2019). One example of artificial intelligence, video screening software for employment, demonstrates the potential AI-human hybrid decisions. In this software, candidates answer predefined questions in a recorded video, and facial recognition/video intelligence analyzes the potential applicant along several dimensions and then rates the candidates for the open position (Zetlin, 2018). This type of facial artificial intelligence enables organizations to screen more candidates with fewer HR managers, boosting productivity and efficiency of HR professionals. In general, proponents of artificial intelligence claim that incorporating AI in firms' decision-making processes can yield notable organizational benefits.

Despite the benefits of algorithms and artificial intelligence in decision making, researchers increasingly acknowledge that algorithms are not neutral (O'Neil, 2016). Artificial intelligence and "big data" can exhibit bias (Buolamwini & Gebru, 2018) and create disparities in treatment and outcomes (Barocas & Selbst, 2016; Crawford, 2013). If organizations encourage employees to use these biased AI scores in their decisions, then a feedback loop is created, similar to the feedback loop in which the models' predictions shape consumer preferences, which in turn shape future predictions (Adomavicius, Bockstedt, Curley, &

Zhang, 2013). A biased AI can induce biased business decisions, so organizations must consider fairness when they implement algorithms in their business processes.

As such, the use of algorithms for decisions had led to concern about the implementation of algorithms and their fairness. Some argue that the definition of fairness, in the context of algorithms, should be approached using a counterfactual perspective; algorithms may be biased but less biased than human discretion and thus their bias is acceptable (Cowgill & Tucker, 2017). Organizational experts argue that algorithms need to have oversight (Luca, Kleinberg, & Mullainathan, 2016) because algorithms excel at pattern-finding and not necessarily at the "soft" and non-quantifiable goals such as fairness. Can decision-makers to properly "manage" algorithms?

In particular, in order to manage algorithms, decision-makers must be capable of judging the artificial intelligence on outside measures. As artificial intelligence expands into new territory, being designed to answer subjective questions such as "job fit", people may struggle to judge its accuracy and/or errors. First, the context-dependent nature of the subjective scores may increase people's proclivity towards using "context clues", anchors, to guide their decisions. Second, it is more difficult to report an error rate or accuracy for the model with subjective measures because the model itself relies on the operationalization of the subjective measure. For example, a model to identify "funny" jokes needs a baseline subset of people who find particular jokes funny. Because humor is inherently subjective, the models will be highly dependent on the context and the process for labeling jokes as "funny" in the dataset. Previous research into algorithmic bias and decision-making has not distinguished between these two types of measures, and we expect that this distinction yields important implications for decision-making with biased systems.

This study addresses this gap by examining how people predict subjective and objective measures in conjunction with artificial intelligence. We consider a sequential AI-human decision that mirrors organizational decisions; an automated system provides a score and then a human makes a decision using their discretion. First, we use artificial intelligence to score pictures on a subjective measure, beauty, and an objective measure, age. Similar to previous research, we find evidence that the artificial intelligence is biased based on skin tone (Buolamwini & Gebru, 2018). Second, we conduct an Amazon Mechanical Turk (AMT) study and ask people to score the same pictures on beauty and on age. In the first condition, people rate the pictures without the AI's score. In the second condition, people rate the pictures after receiving the AI's score. In the third condition, people rate the pictures after receiving the AI's score and the average error of the AI scores across all pictures. We find that participants' scores without the AI do not exhibit bias; however, the introduction of the AI scores induces bias in the participants' scores due to the anchoring effect (Tversky & Kahneman, 1974) for the subjective measure. The participants' scores are not susceptible to the bias in the AI scores for the objective measure. Still, reporting the AI error rate to participants does encourage participants to devalue the AI score and reduces the participants' error.

This study makes several contributions to the information systems literature on algorithmic bias and the need to manage it. First, this paper underscores the importance of reducing algorithmic bias by highlighting the difficulty of "managing" algorithms. Because decision-makers anchor their decisions to the models' predictions, even when the AI is known to be inaccurate, it is imperative that we develop mechanisms to reduce bias in the automated systems and not solely rely on human oversight. Second, this paper finds that the facial recognition AI produces biased predictions and *induces bias* in the decision-makers. Again, if decision-makers anchor their results on biased predictions, this suggests that biased AI may increase bias in decision-makers, even those without bias prior to the introduction to the AI. Lastly, this paper introduces a novel way to classify the prediction goals of AI: subjective measures and objective measures. By using this classification method, we can assess the influence of AI on decision-making.

Background and Hypotheses

Subjectivity vs Objectivity

Research distinguishes between *subjective*, perspective-driven measures and quantifiable *objective* measures (Brookes, 1980). *Subjective* information is situational whereas *objective* information is observer-independent (Hjørland, 2007), and thus the objective information has a relatively indisputable truth

associated with it. Subjective information relies heavily on the operationalization, quantification and data collection process for the "truth" of the information.

Although artificial intelligence is particularly suited for decision-making with a predefined domain of knowledge (Brynjolfsson & McAfee, 2014), commercially available artificial intelligence platforms are expanding into intuitive, subjective domains like beauty and emotions. Attributes such as beauty or humor are highly situational and context-dependent. The first step in these models is to quantify subjective topics into numerical data, often using surveys or labels. Despite the methods used, subjective data are not objective, and this study distinguishes between subjective and objective data.

Bias, Algorithmic Fairness and Oversight

Artificial intelligence and big data can be used in manners that produce a disparate impact for minorities, women, and/or other traditionally disadvantaged, low "status" groups (Barocas & Selbst, 2016). For example, Boston's Street Bump mobile application automatically detected potholes and alerted the city; however, Street Bump detected more potholes in affluent neighborhoods because lower-income constituents were less likely to own smart phones and download Street Bump (Crawford, 2013). Seemingly innocuous algorithms, such as scheduling algorithms to respond to customer demand, can exacerbate existing power differences and harm workers (Barocas & Levy, 2016). Thus, AI and big data are not neutral, and groups with lower status can receive worse scores and outcomes. Because AI is trained on existing data, the resulting models reflect the societal biases around given attributes. This phenomenon is particularly acute for subjective, context-dependent measures such as beauty, humor, or credit-worthiness.

Given that artificial intelligence reflects societal bias, researchers identify two intertwined concerns. First, how can organizations ensure that the models and algorithms in place account for the "soft" organizational goals and targets? Organizations do not want biased artificial intelligence, such as algorithms that automatically assign lower scores to women. Some researchers argue that algorithms should have "managers" (Luca et al., 2016) to provide the necessary oversight to ensure that organizational goals are being met and to reduce liabilities. Second, how can organizations judge the "fairness" of the algorithms to ensure that these algorithms do not exacerbate any pre-existing bias? An extensive literature is emerging with methods to quantify the fairness of model results (Agarwal, Beygelzimer, Dudfk, Langford, & Hanna, 2018; Hardt, Price, & Srebro, 2016; Kleinberg, Mullainathan, & Raghavan, 2017) as a means to inform de-biasing techniques for artificial intelligence.

The literature in algorithmic bias agrees that artificial intelligence will likely reflect societal bias for sensitive topics and/or protected attributes like race and gender. Subjective measures, being context-dependent, are especially susceptible to societal bias. By contrast, objective measures are less susceptible to societal bias because of their external criteria.

H1. Artificial intelligence scores will reflect societal bias in the subjective measure but less societal bias in the objective measure.

One stream of literature argues that, although algorithms may predict disparities among groups, the bias and/or fairness of algorithms should be approached using a counterfactual perspective. The counterfactual for biased algorithmic decisions is not unbiased algorithmic decisions but biased human decisions (Cowgill & Tucker, 2017). A biased AI may still be less biased than a person because the person is more attune to social norms and pressures. Therefore, we expect that participants will generate biased scores for subjective measures. Because the objective measures are observer-independent, we expect participants' scores to exhibit less bias.

H2a. Participants' scores in the subjective measure will reflect societal bias.

H2b. Participants' scores in the objective measure will exhibit less bias.

Anchoring

A view of human-AI decisions suggests that organizations can use a sequential approach (Shrestha et al., 2019), such as the artificial intelligence providing guidance for the employee who acts as the final decision-maker. In the context of hiring software, artificial intelligence reviews the applicant videos to provide a score, and human resources professionals decide which applicants to move further along in the process based on these scores. Perhaps the biased artificial intelligence can still improve human decisions, leading to ultimately lower bias than without the artificial intelligence.

Behavioral economics yields some insights into how artificial intelligence may influence individual decisions. Decision-makers are sensitive to the initial starting point in their predictions, which is a phenomenon called *anchoring* (Tversky & Kahneman, 1974). Because decision-makers will not sufficiently adjust their predictions from the initial suggestion, different anchors will produce different predictions (Tversky & Kahneman, 1974).

There is evidence that consumer decisions are sensitive to suggestions of others, including suggestions by artificial intelligence. Consumers' preferences are shaped by the predictions of a recommender system, likely because those recommendations anchor the consumer preferences (Adomavicius et al., 2013).

Because decisions are malleable, the anchoring effect exacerbates the challenge of algorithmic bias. People may generate unbiased scores (i.e., scores without disparities on sensitive identities such as gender and ethnicity); however, when they are primed with a model that exhibits algorithmic bias, the biased score *anchors* their decisions and people mirror that bias. Because subjective scores are situation and context-dependent, we expect that subjective scores are more susceptible to the anchoring effect.

H3a. Artificial intelligence will anchor the participants' scores for subjective measure.

H3b. Artificial intelligence will anchor the participants' scores for objective measure.

Errors

Anchoring occurs partially because decision-makers do not make sufficient adjustments from the initial anchor (Tversky & Kahneman, 1974), so we examine how to encourage decision-makers to adjust from the initial starting point and "distrust" the anchor. People place enormous trust in machines, even when the machines lead them against their common sense and personal knowledge (Robinette, Li, Allen, Howard, & Wagner, 2016). People often trust automated systems to the point of following them into walls during an emergency; however, this "overtrust" could be dampened by indicating that the automated system was broken (Robinette et al., 2016). We expect that providing evidence that the AI is inaccurate will mitigate the anchoring effect.

Algorithms can be demonstrably inaccurate for objective measures unlike situationally-dependent subjective measures. Artificial intelligence that predicts an objective measure is easier to analyze for bias because the information is observer-independent. For example, gender is an example of an objective measure; facial recognition software predicts gender significantly worse for darker-skin faces than lighter-skin faces (Buolamwini & Gebru, 2018). Because gender is an observer-dependent characteristic, skin color biases in artificial intelligence performance are captured through accepted goodness-of-fit measures such as accuracy rates.

One challenge about predicting context-dependent measures is the difficulty in assessing the model's predictions. Because subjective measures inherently rely on the perspective of the observer, group-specific biases in artificial intelligence are difficult to assess. The hallmark of AI bias for subjective measures is the difficulty in answering the question: is the model biased or are the data biased?

If AI is biased or inaccurate, and this bias is known, then humans could devalue the guidance provided by the artificial intelligence in determining their own scores. Does providing the average error rate for the AI decrease the human reliance on the AI for decision-making? If so, this insight could provide a path forward

in our understanding about to incorporate flawed and/or biased algorithms into individual decision-making without hindering decisions.

H4. The anchoring effect will be lessened when participants receive information about the AI's error.

Summary

This study seeks to understand how the anchoring effect of AI predictions can exacerbate bias in human decisions. The research framework is summarized below in Figure 1.

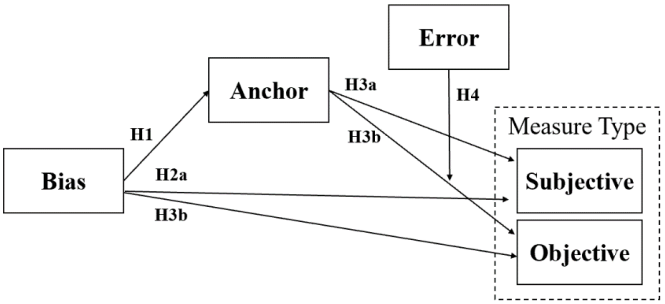

Figure 1. Research Framework

Because artificial intelligence is not neutral, it is expected that the predictions will be biased in a way that reflects societal preferences (**H1**). Societal expectations often seep into preferences, so people, and their predictions, are biased (**H2**). Given the seeming ubiquity of the anchoring effect, we expect that survey participants will predict values closer to the AI prediction when those values are available. We expect that the anchoring effect will influence human prediction for subjective data (**H3a**) and objective measures (**H3b**). By providing the average error rate, we expect to reduce the anchoring effect of the AI prediction (**H4**). In summary, this study provides insight in understanding how to incorporate flawed and/or biased algorithms into individual decision-making without hindering decisions.

Methodology

Data

This study uses the context of commercially available facial recognition software to examine how artificial intelligence anchors predictions and introduces bias. Facial recognition is increasingly embedded in society and organizations for security and screening purposes, such as software for company threat detection as well as video intelligence software for screening job candidates. In addition to its widespread adoption, facial recognition analyzes faces for subjective and objective measures, an important criterion for this study.

Commercially available facial recognition models analyze and predict an ever-increasing number of facial attributes. One commercially available facial recognition software, Face++, offers a number of APIs to analyze faces. These offerings range from objective measures like eye placement on the face to subjective measures of emotion and beauty.

In order to compare how participants make predictions for objective data versus subjective data, we use the beauty API and the age API. In the beauty score API, Face++ analyzes each individual face in the picture and scores the faces by their attractiveness to men and women on a scale of 0-100. Beauty is our subjective measure because it is widely considered to vary by the individual and there is no observer-independent measure for the beauty of an individual. We operationalize bias as skin-tone for beauty because prior research finds that evidence of colorism, i.e., privilege associated with lighter skin (M. Hunter, 2007). Societal norms of colorism elevate lighter skin-tone as the beauty ideal (Bond & Cash, 1992; M. L. Hunter,

2002; Jenq, Pan, & Theseira, 2015; Li, Min, & Belk, 2008), leading to higher status due to the additional social capital (M. L. Hunter, 2002).[1] In the age API, Face++ analyzes each individual face in the picture and predicts the age of the person in the picture. Age is our objective measure because each individual has an actual age, and the age is independent of the observer and other characteristics.

We construct these measures using press pictures from contestants in the 2018 Miss America pageant. We collect pictures from the top 15 contestants, and then other contestants are included to ensure a diverse set of 26 contestants. These contestants are smiling and wearing regular clothes with full makeup and a blowout hairstyle (i.e., hanging by their shoulders and not pulled back). We then categorize these contestants based on their skin color using the Fitzpatrick skin color categories (Buolamwini & Gebru, 2018). This image dataset controls for differences in picture quality and gender. Also, by using beauty pageant contestants, we are more likely to remove subjective preferences for hair length, facial expression, and make-up.

Survey Design

This study investigates biased AI predictions and its influence on human scores. We conduct an experiment on Amazon Mechanical Turk (AMT). We ask survey participants to predict the beauty score and the age score for the contestant faces. Table 1 shows an overview of the design.

This study uses an incomplete 3x2 matrix to compare the participants' predictions without guides to the predictions with guides. There are two distinct prediction measures: Beauty (subjective) and Age (objective) and three distinct anchoring conditions: *No Guide*, *Guide – AI*, *Guide – AI & Error Reported*.

	Subjective Measure (Beauty)	Objective Measure (Age)
No Guide	N=130	N=130
Guide – AI	N=130	N=130
Guide – AI (Error Reported)		N=130

Table 1. Experiment Design and Treatment Conditions

In the *No Guide* anchoring condition, the participant receives no additional information. In the *Guide – AI* anchoring condition, the participant receives the AI prediction but no information about the accuracy of the AI. In the *Guide – AI & Error Reported* condition, the participant receives both the AI prediction and the average error across all pictures. Each picture received 5 ratings per condition.

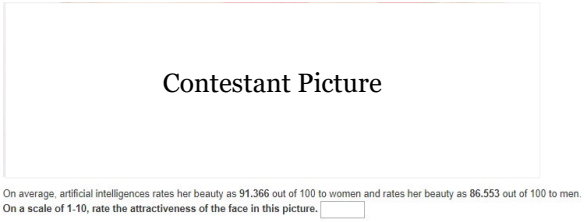

Figure 2. AMT Survey with the guide

[1] Colorism is damaging and disempowering (Li et al., 2008). The operationalization of bias as colorism is an acknowledgement of current societal struggles with colorism and algorithmic bias (Buolamwini & Gebru, 2018). We do not endorse colorism or its premise, i.e., privileging people by skin-tone.

Figure 2 shows an example of the AMT survey for the Guide condition on the Beauty measure. In this condition, the participant rates the attractiveness of the contestant on a scale of 1-10 with the guide of the Face++ rating.

Summary Statistics

Table 2 shows the summary statistics for this dataset. There are differences among the scores for each of the conditions, although the differences are not statistically significant.

	Beauty Scores			**Estimated Age**		
Condition	Mean	Median	Std. Dev	Mean	Median	Std. Dev
None	70.5	70	13.2	34.6	29.5	16.4
Guide	73.9	70	10.9	34.9	30	15.6
Error	N/A	N/A	N/A	40.6	31	22

Table 2. Summary Statistics for Participants' Scores

Overall, participants rated the contestants as attractive with a mean beauty score above 7 in both conditions. The participants' beauty scores are higher and the standard deviation is lower in the *Guide* condition. For Age, the participants consistently guessed that the contestants are in their early 30s. The participants' mean estimated age and the standard deviation are higher for the *Error* condition.

Ratings Comparison Across Objective and Subjective Measures

Objective measures have an inherent, observer-independent true value. Therefore, the model's error rate can be calculated and communicated to the algorithm's "manager". To quantify the objective bias, we use the mean absolute error (MAE):

$$MeanAbsError = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

MAE captures the average of absolute difference of the predicted value $\hat{y}_i$ from the true value y_i for all N records. MAE treats any deviation from the true value y_i as error.

We compare that to the mean error:

$$MeanError = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)$$

Mean error includes the sign for the deviation, so predictions that undershoot the true value offset predictions that exceed the true value. This measure for error allows us to know the direction of the error.

Table 3 reports the comparisons of the MAE and the Mean Error conditions for the predicted ratings. We also examine the difference between the Face++ scores and participants' scores across conditions.

Objective Measure: Estimated Age				
	Prediction	MAE	Mean Difference (Human - AI)	MAE (AI as "true value")
Face++	31.577 (1.305)	9.615 (1.300)		
No Guide	29.966 (0.529)	8.025 (0.525)	1.471 (0.723)	6.311 (0.452)
Guide – AI	30.917 (0.523)	8.950 (0.502)	0.860 (0.534)	3.868 (0.408)

Subjective Measure: Beauty Score				
	Prediction	MAE	Mean Difference (Human - AI)	MAE (AI as "true value")
Face++	80.052 (1.395)			
No Guide	70.462 (1.160)	N/A	9.590 (1.270)	13.791 (0.923)
Guide – AI	73.923 (0.955)	N/A	6.129 (0.991)	9.977 (0.709)

Table 3. Prediction Comparisons

For *Age*, both the AI and the participants predicted a surprisingly higher age than the contestants' actual age, nearly 9 years. Participants in the *Guide* condition were less accurate than participants in *No Guide*, since Face++ estimated age was 9.6 years older than the participants' true age. Participants in the Guide condition also mirrored the Face++ predictions within a year, on average, suggesting that the Face++ prediction did indeed anchor their predictions.

For *Beauty*, the participants rated the contestants lower on average than the AI. For a subjective measure, we could not quantify the error. On average, the difference between the participant's rating and the AI rating decreased in the *Guide* condition, again supporting the premise that the anchoring effect may influence the participants' ratings.

Empirical Results

H1-H2: Algorithmic Bias and Human Bias

H1 and H2 suggest that humans and artificial intelligence will both exhibit bias, so we analyze the participants' answers for bias based on skin-tone. We compare the Face++ estimated scores with the average responses from participants in the *No Guide* category. Our model uses the following specification:

$$E\left[\hat{Y}_i \middle| X_i\right] = \alpha + \beta_1 Bias_i + \epsilon_i \ (1)$$

Where the $\hat{Y}$ is the estimated prediction for either our objective measure (*Age*) or the subjective measure (*Beauty*). *Bias* is the skin-color rating on a range of 1 to 6 for human skin color according to the Fitzpatrick scale, so a significant coefficient on this variable is an indication of bias in the scores. We also include an individual fixed effect for each contestant's picture and control for some characteristics of the individual.

Dependent Variable =	**Estimated Age**			**Estimated Beauty Score**		
Variable	(1)	(2)	(3)	(4)	(5)	(6)
Bias	1.540 (0.889)+	0.4639 (0.3646)	1.7167 (1.1505)	-1.7490 (0.9427)+	-0.6350 (0.8229)	2.000 (2.804)
Contestant Fixed Effects	N/A	No	Yes	N/A	No	Yes
R^2	0.1111	0.01364	0.3726	0.1254	0.004631	0.1848
p-value	0.096	0.206	0.003	0.076	0.442	0.547
Group	AI	Humans	Humans	AI	Humans	Humans
N	26	130	130	26	130	130
*p-values: + <0.10; * <0.05; ** -- <0.01; *** -- <0.001*						

Table 4. Participants' Scores (No Guide)

Table 4 shows that results from this estimation. There is limited evidence that the model is biased in its predictions for both *Age* and *Beauty*. *Bias* is significant at p<=0.10 for the Face++ predictions, despite the

low sample size of N=26. *Bias* is not significant for the survey participants in their predictions for *Estimated Age* (columns 2 and 3) or their assessments for *Estimated Beauty* (columns 5 and 6). We find no evidence that survey participants' predictions were biased in the *No Guide* condition.

This directionally significant coefficient estimate on *Bias* suggests that Face++ is more sensitive to the beauty norms in its estimates than the participants are. There has been extensive research on the relationship between skin color and beauty standards as well as skin color and age, so this research is not attempting to delve into those waters. Obviously, algorithmic bias should be changed, and prior research offers a host of concrete suggestions (Agarwal et al., 2018; Barocas & Levy, 2016; Barocas & Selbst, 2016; Buolamwini & Gebru, 2018; Hardt et al., 2016; Kleinberg et al., 2017; Luca et al., 2016). We consider a different question: does the bias of Face++ influence the participants?

We investigate this question to understand how bias in artificial intelligence, however it originates, seeps into human decisions by anchoring participants' predictions. We consider how AI predictions anchor participants' scores because this experiment reflects the organizational process by which the AI prediction serves as a guide for decisions such as whether to extend a loan or to hire someone.

H3: The Anchoring Effect on Predictions

To understand the anchoring effect, we examine the differences in the rating between the *Guide* and *No Guide* anchoring conditions. We examine the following model:

$$E\left[\hat{Y}\middle|X\right] = \alpha + \beta_1 Guide_i + \beta_2 Bias_i + \beta_3 Guide_i * Bias_i + \epsilon_i \ (2)$$

Where the $\hat{Y}$ is the participant's prediction for either our objective measure (*Age*) or the subjective measure (*Beauty*), same as before. *Guide* is the binary variable indicating whether the participant received a guide with the Face++ predictions. *Bias* is the same as above. *Guide*Bias* is the interaction term between these two conditions. Because each contestant receives five ratings, we can include an individual fixed effect and control for some characteristics of the contestant.

Anchoring: Subjective Measures

Using the subjective measure, *Beauty*, we find evidence of the anchoring effect. Table 5 shows the coefficient estimates for the subjective measure, Beauty Score. The main effect for *Guide* is significant across all models (columns 1-4), suggesting that showing participants the Face++ estimated beauty score does indeed influence participants' ratings and providing support for the anchor effect. Interestingly, the coefficient estimate is positive across all models, indicating that facial recognition predictions encourage more positivity and pulls the predicted scores upward.

	Dependent Variable = Beauty Score			
Variable	(1)	(2)	(3)	(4)
Guide – AI	3.462 (1.503)*	3.4615 (1.4673)*	11.3501 (3.4827)**	11.3501 (3.4351)**
Bias		-1.9169 (0.5195)***	-0.6350 (0.7274)	0.6152 (1.7952)
Guide*Bias			-2.5638 (1.0287)*	-2.5638 (1.0147)*
Contestant Fixed Effects				Yes
R^2	0.0202	0.069	0.091	0.199
p-value	0.02205	9.628e-05	1.864e-05	0.001449
*p-values: + <0.10; * <0.05; ** -- <0.01; *** -- <0.001*				

Table 5. Participants' Scores for Subjective Measure

Bias is significant and negative for column 2, indicating that participants did exhibit bias. However, the interaction term between *Guide* and *Bias* suggests that the negative effect of *Bias* on estimated *Beauty* occurs in conjunction with the anchoring effect. In essence, the participants' scores may mirror the bias in the facial recognition software rather than their own intrinsic beauty preferences. This result still holds if we control for the individual differences in faces, as shown in model (4). This result provides evidence that bias is contagious for the subjective measure.

Anchoring: Objective Measures

We estimate the model for predictions on the objective measure, *Age*, and we do not find evidence of the anchoring effect. Table 6 shows the coefficient estimates for this model.

	Dependent Variable = Predicted Age			
Variable	(1)	(2)	(3)	(4)
Guide – AI	0.9510 (0.7436)	0.9854 (0.7347)	-0.3073 (1.7408)	-0.4873 (1.5038)
Bias		0.6719 (0.2564)**	0.4639 (0.3610)	1.9432 (0.8228)*
Guide*Bias			0.4204 (0.5132)	0.4244 (0.4430)
Contestant Fixed Effects				Yes
R^2	0.006825	0.03479	0.03753	0.358
p-value	0.2022	0.01505	0.02868	3.465e-10
*p-values: + <0.10; * <0.05; ** -- <0.01; *** -- <0.001*				

Table 6. Estimated Ratings for Objective Measures

Guide is not significant across any of the specifications, suggesting that participants are not necessarily influenced by the facial recognition's estimated age. *Bias* is significant and positive for columns 2 and 4, indicating that darker skin contestants are viewed as slightly older than lighter skin contestants across all individuals. However, the interaction term *Guide*Bias* is not significant, highlighting the limited influence of AI's predictions to anchor individual predictions.

The Anchoring Effect on Relative Ratings

We have evidence so far that the *Guide* condition anchors subjective measures but not objective measures. However, the above analysis only examines the anchoring effect on the absolute score. If the anchoring effect holds, then we would expect that the participants' scores to be nearer to the AI scores. We examine the differences between the Face++ rating and the participants' rating between the *Guide* and *No Guide* anchoring conditions. We examine the following model:

$$E[|\hat{y_i} - y_i||X] = \alpha + \beta_1 Guide_i + \beta_2 Bias_i + \beta_3 Guide_i * Bias_i + \epsilon_i \ (3)$$

Where $\hat{Y}$ is the estimated prediction for either our objective measure (*Age*) or the subjective measure (*Beauty*), same as before, and y_i is the Face++ estimated beauty. The other variables, *Guide, Bias*, and *Guide*Bias* are the same as above. Again, we can include an individual contestant fixed effect and control for some characteristics of the individual.

Subjective Relative Anchoring

The coefficient estimates for the subjective measure, *Beauty*, are shown in Table 7. *Guide* is significant and negative across all specifications, showing that giving participants the AI's score does indeed anchor

participants' predictions and reduce the gap between the participants' score and the AI score. *Bias* is not significant, indicating that bias does not influence whether the participants anchor their scores to the artificial intelligence guide. The coefficient estimate for *Guide*Bias* is significant with the fixed effects, suggesting that the guide may widen the absolute difference between the participants' score and the artificial intelligence score for darker skin contestants.

	Dependent Variable = Absolute Ratings Difference for Beauty				
Variable	(1)	(2)	(3)	(4)	(5)
Guide	-3.814 (1.164)**	-3.814 (1.166)**	-7.729 (2.79)**	-3.814 (1.081)***	-7.729 (2.581)**
Bias		0.000 (0.413)	-0.636 (0.582)	-1.2428 (1.299)	-1.8790 (1.3486)
Guide* Bias			1.272 (0.824)		1.272 (0.762)+
Fixed Effects	No	No	No	Yes	Yes
R^2	0.0176	0.0179	0.040	0.253	0.2616
p-value	0.033	0.098	0.005	0.0213	2.951e-06
*p-values: + <0.10; * <0.05; ** -- <0.01; *** -- <0.001*					

Table 7. Coefficient Estimates for Relative Anchoring Effect

Objective Relative Anchoring

The coefficient estimates for the objective measure, *Age*, are shown in Table 8. Across all models, *Guide* is not significant, so we find no evidence of anchoring in the objective measures. *Bias* remains significant with the fixed effects, suggesting that participants' score contestants with darker skin as older, but it does not appear influenced by the facial recognition scores.

	Dependent Variable = Ratings Difference for Age				
Variable	(1)	(2)	(3)	(4)	(5)
Guide	-0.6111 (0.8967)	-0.5689 (0.8855)	-0.81827 (0.63537)	0.672 (2.099)	0.487 (1.504)
Bias		0.8227 (0.3090)**	1.85312 (0.79465)*	1.022 (0.435)*	2.057 (0.823)*
Guide*Bias				-0.403 (0.619)	-0.424 (0.443)
Fixed Effects	No	No	Yes	No	Yes
R^2	0.001948	0.031	0.554	0.033	0.556
p-value	0.4962	0.024	< 2.2e-16	0.04909	< 2.2e-16
*p-values: + <0.10; * <0.05; ** -- <0.01; *** -- <0.001*					

Table 8. Coefficient Estimates for Objective Relative Anchoring Effect

Overall, these results indicate that the artificial intelligence estimates do indeed influence participants' choices, and the manner that they are altered is consistent with the anchoring effect. Anchoring is only present in the subjective measure, *Beauty*, and not in the objective measure, *Age*. This finding highlights the distinction between these types of measures.

H4: AI and Accuracy

The last hypothesis investigates whether information about the errors of the artificial intelligence could dampen the anchoring effect. Because there is no observer-independent truth for subjective measures, we cannot gauge whether the overall error of the artificial intelligence with an observer-independent true value. However, the truth exists for objective measures so we can examine how knowledge about the artificial intelligence error shapes participants' decisions.

Organizations ultimately want their employees to improve their decisions, and in the case of objective measures, accurately predict the measure. To understand how artificial intelligence influences participants' accuracy, we examine the mean absolute error between the participants' predicted age and the contestant's true age in the *Guide*, *No Guide,* and *No Guide & Error Reported* anchoring conditions. We examine the following model:

$$E\left[|\hat{Y} - \bar{y}_i|\middle|X\right] = \alpha + \beta_1 Guide_i + \beta_2 Bias_i + \beta_3 Guide_i * Bias_i + \beta_4 Error_i + \beta_5 Error_i * AI\_MAE_i + \epsilon_i \ (4)$$

Where $\bar{y}_i$ is the true age of the beauty contestant, and the variables *Guide* and *Bias* are the same as in equation (3). *Error* is a binary variable indicating whether the participant received information about the error of the artificial intelligence. *AI_MAE* is the Face++ mean average error for the predicted age.

Objective Relative Accuracy

Table 9 shows the estimation results. There is no evidence that the predicted *Guide* or *Bias* affects the MAE, supporting our prior finding that participants were impervious to the artificial intelligence anchors for the objective measure. Reporting the average error for the artificial intelligence predictions does indeed reduce the participants' MAE (β = -2.432, $p<0.05$). However, the interaction term for *ErrorReporting*AI_MAE* is positive and significant, suggesting that reporting larger MAEs for the models increases participants' MAE. We find evidence that reporting the error in the artificial intelligence can mitigate the AI errors. results and that the participants' anchor their scores to the AI predictions.

Dependent Variable =	*Predicted Age MAE*		
Variable	(1)	(2)	(3)
Guide	0.782 0.646	0.805 (0.640)	-0.450 (1.377)
Error reporting	-0.271 0.665	-2.558 (1.090)*	-2.432 (1.096)*
AI's MAE	-0.085 0.321	-0.148 (0.319)	-0.146 (0.319)
Error reporting* AI MAE		0.235 (0.089)**	0.221 (0.090)*
Bias	1.896 1.417	1.859 (1.404)	1.604 (1.426)
Guide*Bias			0.407 (0.395)
Fixed Effects	Yes	Yes	Yes
R^2	0.258	0.272	0.272
p-value	1.258e-14	1.72e-15	2.66e-15

Conditions	All	All	All
*p-values: + <0.10; * <0.05; ** -- <0.01; *** -- <0.001*			

Table 9. Coefficient Estimates for Objective Error Rates

Because objective measures are demonstrably true or false, the average error rates can be given to decision makers to aid their decisions (if available to consumers). Future work can consider the sensitivity of the type of error on participants' scores as well as how to quantify error for subjective measures.

Summary

Overall, the findings support most of the hypotheses. There is support for **H1**, that there is bias in the algorithm's ratings, although there is no evidence of bias in the human. Although this study only examines a single commercially available facial recognition software, Face++, this finding confirms previous research that finds bias in artificial intelligence in general (Barocas & Levy, 2016; Barocas & Selbst, 2016; O'Neil, 2016) and in facial recognition in particular (Buolamwini & Gebru, 2018).

Furthermore, the results support **H3a**, that the AI scores anchor the human scores in the subjective measure, but not for **H3b**, that the AI scores anchor the human scores in the objective measure. There is evidence that reporting the AI errors does reduce the errors in the participants' scores, supporting **H4**.

Discussion

In this study, we begin with the premise that both artificial intelligence and human participants display bias. For a scoring task, we expect that participants' scores with exhibit bias will reflect societal bias, especially for a subjective, context-dependent measure like beauty. This study confirms that expectation for artificial intelligence, but not for participants.

We then conduct a study on AMT to understand how artificial intelligence can shape participants' scores. Participants are divided along two dimensions: one dimension of Measure Type with two measures (Subjective and Objective), and one dimension of AI guidance with three conditions (*None*, *Guide - AI*, and *Guide - AI & Error Reporting*). Participants are asked to score beauty contestants with varying levels of guidance from artificial intelligence.

Using the theory of anchoring, we explain how artificial intelligence can induce bias. Participants, whose answers do not display bias in the No Guidance condition, demonstrate bias in their scores in the Guide – AI condition for the subjective measure, *Beauty*. This observation suggests that the AI guided participants towards biased scores for *Beauty*.

Beauty depends heavily on context and culture, so there are no externally valid criteria on which to judge the "error" of the AI model. We can tell if the artificial intelligence displays systematic preferences for one group over another, i.e., finding one group more beautiful than another group. In this study, we consider that bias but others would argue that the measure captures real differences in beauty among the groups, i.e., one group truly is more attractive than another. Subjective measures lack that observer-independent measure to act as the arbiter of these disagreements.

From a legal and public relations standpoint, organizations still cannot accept a decision process that generates biased decisions. If organizations want to guide participants into devaluing a biased model, there is not an available goodness-of-fit measure. Instead, artificial intelligence depends on the data collection mechanism for the labels, and this context-dependence is not necessarily communicated in the API. In the case of Face++, the Beauty Score API is communicated as "the beauty score" without an explanation for *whose* standards of beauty it captures.

We contrast our subjective measure, *Beauty*, with an objective measure, *Age*. For the objective measure, we find no evidence of artificial intelligence bias. Plus, AI does not induce bias, as participants are not guided by the artificial intelligence. Interestingly, according to the MAE, artificial intelligence demonstrates lower proficiency at estimating age than the participants, and this observation could explain why participants disregard the artificial intelligence guides in their scores. Even without reporting the error, participants appear able to disregard the inaccurate and possibly unhelpful advice of the AI. In the final condition, we

report the average error of the artificial intelligence. Participants do respond to this knowledge by devaluing the guidance of the AI and adjusting their score accordingly.

As artificial intelligence is increasingly embedded within organizations, it is unlikely that AI-based predictions will substitute entirely for human decisions. Instead, artificial intelligence is incorporated as a variety of potential AI-human decisions (Shrestha et al., 2019). AI may be used to support a final human decision, such as the score provided by the hiring intelligence software used by human resources professionals in the applicant screening process. Because human predictions often work in concert with artificial intelligence, it is important to understand the influence of AI, particularly biased AI, on human decisions. This study mirrors that process by retrieving artificial intelligence scores and then presenting those scores to participants to guide their decisions.

This study makes multiple contributions to the information systems literature on algorithmic bias. First, we introduce a novel framework to distinguish between types of artificial intelligence goals: subjective vs. objective. Given the emerging literature on algorithmic bias and de-biasing algorithms, the distinction between subjective and objective measures yields important implications for the literature on mitigating bias. Subjective measures increase the challenges in assessing the accuracy of artificial intelligence and determining its bias. Objective measures enable designers to identify bias because the model scores have an observer-independent, incontrovertible truth. This aspect of the artificial intelligence enables an easier and robust discussion among the stakeholders (organizations, designers, consumers, etc.) about its bias.

By introducing this framework for the measures, this research also highlights the difficulty in using legal means to regulate algorithmic bias. Prior attempts to regulate algorithms may run into national or local difficulties as countries have different perspectives towards artificial intelligence, innovation, and the measure itself. For example, the cultural norms in one country may dictate one standard of decorum, and another country may have another. An artificial intelligence to judge cultural fit will exhibit bias for the culture on whose data it was trained. Regulations would not necessarily succeed in changing the outputs.

Second, this research finds that artificial intelligence *induces bias* in the decision-makers. Prior literature has found evidence of algorithmic bias (Barocas & Selbst, 2016; Buolamwini & Gebru, 2018), but this research examines specifically how that bias can shape decisions. Many scholars assumed that human oversight and intervention could mitigate bias (Luca et al., 2016) or that artificial intelligence would lessen bias (Cowgill & Tucker, 2017). We find that biased artificial intelligence can stimulates bias. If people anchor their decisions on biased artificial intelligence, then this research suggests that those decisions will also be biased. We note that the measure type matters; perhaps artificial intelligence exacerbates bias for context-dependent measures but may reduce it for objective measures. As such, this paper underscores the need to reduce algorithmic bias. For a subjective measure, bias is contagious. Because decision-makers anchor their decisions to the models' predictions, even when the AI is known to be inaccurate, it is imperative that we develop mechanisms to reduce bias in the automated systems and not solely rely on human oversight to "manage" algorithms within organizations.

This study produces practical implications for organizations seeking to responsibly adopt artificial intelligence and limit algorithmic bias. First, this paper cautions against the widespread adoption of artificial intelligence to automate and/or provide guidance for all decisions. Biased artificial intelligence is contagious. Even if organizations restrict AI to guiding their employees rather than having it make final decisions, a biased tool still influences human decisions and could yield biased results. Also, as companies adopt AI for subjective measures such as "cultural fit", the concern about contagious algorithmic bias is particularly acute. Despite organizations' best efforts, AI could perpetuate bias among employee decisions with minimal definitive detection. This concern is less pressing for objective measures when the artificial intelligence can be judged on observer-independent criteria, but still raises the concern for organizations adopting AI.

Second, this study examines whether actually providing information about artificial intelligences' error reduces reliance on its guidance. We find provide evidence that participants are less likely to anchor their decisions on the artificial intelligence with information about its average error. Interestingly, this additional information does not necessarily increase participants' accuracy. This study confirms the importance of providing error information about artificial intelligence to aid employees in their judgements. In order to

assess the valuable in the artificial intelligence guidance, employees must have information about the relative accuracy of the system. Given that this type of information is only available for objective measures, organizations should proceed carefully with implementing artificial intelligence for subjective measures.

Also, participants may be susceptible to the information presentation. This study also shows the basic error rates for the artificial intelligence, but participants may find others forms of information presentation more compelling in their guidance. Future work could analyze how different presentations for error rates can disrupt the decision-makers use of bias AI. As a caution, AI produces notable benefits, so we want to discover a means of information by the bias AI is interpreted with error rates but an unbiased AI is not.

Furthermore, this work highlights the international nature of commercially available facial recognition and the context-dependent nature of some subjective measures. Beauty is often cultural, but these algorithms can extend their reach globally and influence the standards for users in other countries.

Limitations and Future Research

This study has multiple limitations. First, this research generalizes with subjectivity and objectivity with two measures, beauty and age respectively. This methodology naturally limits the appropriate datasets and the artificial intelligence that are suitable for the research, and this paper analyzes a fairly small sample size. Future research could expand to examine more nuanced measures instead of beauty, one of the most inherently subjective measures, and age, an objective true measure. Second, this study simplifies the AI-human decisions in order to focus on subjectivity vs. objectivity and the anchoring effect. Future research could embed this situation within an organizational context and test different situations of people as algorithmic "managers".

Conclusion

Artificial intelligence is emerging in subjective areas like emotional analysis, organizational "fit", and beauty. As famously quoted by Margaret Hungerford, "Beauty is in the eye of the beholder". Beauty is a quintessentially subjective measure, so perhaps society should exercise caution in using an artificial intelligence package to analyze it. Artificial intelligence is susceptible to bias, for a variety of potential reasons, and bias could spread bias from AI to human decision-makers by anchoring the humans' scores. This bias-contagion effect weakens the suggestion of human intervention as a means to mitigate bias and also highlights the importance of understanding AI in organizational decision-making. Future research must understand how individuals use AI to augment their decisions and the potential spillover effects for bias in the use of artificial intelligence.

References

Adomavicius, G., Bockstedt, J. C., Curley, S. P., & Zhang, J. (2013). Do Recommender Systems Manipulate Consumer Preferences? A Study of Anchoring Effects Do Recommender Systems Manipulate Consumer Preferences? A Study of Anchoring Effects. *Information Systems Research*, *24*(4), 956–975. https://doi.org/10.1287/isre.2013.0497

Agarwal, A., Beygelzimer, A., Dudfk, M., Langford, J., & Hanna, W. (2018). A reductions approach to fair classification. *35th International Conference on Machine Learning, ICML 2018*, *1*, 102–119.

Barocas, S., & Levy, K. (2016, August 31). What Customer Data Collection Could Mean for Workers. *Harvard Business Review*. Retrieved from https://hbr.org/2016/08/the-unintended-consequence-of-customer-data-collection

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review*, *104*, 671–732. http://dx.doi.org/10.15779/Z38BG31

Bond, S., & Cash, T. F. (1992). Black Beauty: Skin Color and Body Images among African-American College Women1. *Journal of Applied Social Psychology*, *22*(11), 874–888. https://doi.org/10.1111/j.1559-1816.1992.tb00930.x

Brookes, B. C. (1980). Measurement in Information Science: Objective and Subjective Metrical Space. *Journal of the American Society for Information Science*, *31*(4), 248–255. https://doi.org/10.1002/asi.4630310404

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. New York, NY: W.W. Norton & Company.

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research*, *81*, 1–15.

Cowgill, B., & Tucker, C. (2017). Algorithmic Bias: A Counterfactual Perspective. *Working Paper: NSFTrustworthy Algorithms*.

Crawford, K. (2013). The Hidden Biases in Big Data. *Harvard Business Review Blog Network*.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, (Nips), 3323–3331.

Hjørland, B. (2007). Information: Objective or subjective/situational? *Journal of the American Society for Information Science and Technology*, *58*(10), 1448–1456. https://doi.org/10.1002/asi.20620

Hunter, M. (2007). The Persistent Problem of Colorism: Skin Tone, Status, and Inequality. *Sociology Compass*, *1*(1), 237–254. https://doi.org/10.1111/j.1751-9020.2007.00006.x

Hunter, M. L. (2002). "If You're Light You're Alright": Light Skin Color as Social Capital for Women of Color. *Gender & Society*, *16*(2), 175–193. https://doi.org/10.1177/0891243O222104895

Jenq, C., Pan, J., & Theseira, W. (2015). Beauty, weight, and skin color in charitable giving. *Journal of Economic Behavior & Organization*, *119*, 234–253. https://doi.org/10.1016/j.jebo.2015.06.004

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. *Leibniz International Proceedings in Informatics, LIPIcs*, *67*, 1–23. https://doi.org/10.4230/LIPIcs.ITCS.2017.43

Li, E. P. H., Min, H. J., & Belk, R. W. (2008). Skin Lightening and Beauty in Four Asian Cultures. *ACR North American Advances*, *NA-35*. Retrieved from http://acrwebsite.org/volumes/13415/volumes/v35/NA-35

Luca, M., Kleinberg, J., & Mullainathan, S. (2016). Algorithms need managers, too. *Harvard Business Review*.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

Phillips-Wren, G., & Jain, L. (2006). Artificial Intelligence for Decision Making. In B. Gabrys, R. J. Howlett, & L. C. Jain (Eds.), *Knowledge-Based Intelligent Information and Engineering Systems* (pp. 531–536). Springer Berlin Heidelberg.

Robinette, P., Li, W., Allen, R., Howard, A. M., & Wagner, A. R. (2016). Overtrust of robots in emergency evacuation scenarios. *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 101–108. https://doi.org/10.1109/HRI.2016.7451740

Shrestha, Y. R., Ben-Menahem, S. M., & von Krogh, G. (2019). Organizational Decision-Making Structures in the Age of Artificial Intelligence. *California Management Review*, *61*(4), 66–83. https://doi.org/10.1177/0008125619862257

Tambe, P., Cappelli, P., & Yakubovich, V. (2019). Artificial Intelligence in Human Resources Management: Challenges and a Path Forward. *California Management Review*, *61*(4), 15–42. https://doi.org/10.1177/0008125619867910

Tversky, A., & Kahneman, D. (1974). *Judgment under Uncertainty: Heuristics and Biases* (pp. 1124–1131).

Zetlin, M. (2018, February 28). Got a Poker Face? Employers are Using AI to Analyze Candidates' Facial Expressions and Personalities. Retrieved September 4, 2019, from Inc.com website: https://www.inc.com/minda-zetlin/ai-is-now-analyzing-candidates-facial-expressions-during-video-job-interviews.html