Revisiting the managerial impact of AI algorithmic fairness on business decisions: A replication of Cowgill et al. 2020*

Yingying Zhou

23/04/2021

Abstract

This paper reproduces Cowgill, Dell'Acqua, and Matz (2020) to reinvestigate the managerial impact of algorithm fairness using a RCT field experiment under two business decision-making scenarios. This paper identifies demographic traits explaining the fundamental diversity in firm manager's attitude towards AI and further examines the effect of interventions on AI adoption through activism arguments on algorithmic fairness. This paper finds that counterfactual advocacy arguments on algorithmic bias are more effective in promoting AI adoption for business use. Besides argument manipulations, race, gender, and knowledge on the status quo fundamentally impact manager's judgement on AI adoption.

Contents

*Code and data are available at: https://github.com/StephaininZ/algorithm_fairness

1 Introduction

Artificial Intelligence algorithmic bias in business decision-making inflicts ethical consequences which mostly materialize to company image damage and revenue loss. While AI technology has expedited the decision-making process in business operations, liberating manual efforts and replacing rule-based models in the age of big data for firms, concerns have been aroused on AI adoption by business decision-makers.

Cowgill, Dell'Acqua, and Matz (2020) examines the managerial effects of argument intervention through two Randomized Controlled Trial field experiment series in two business cases on hiring and lending decisions via investigations on effect originating from opinion polarity and from scientific veneer. In the first study, Cowgill, Dell'Acqua, and Matz (2020) randomly assigns subjects to one of three op-ed conditions (fatalistic, counterfactual or no op-ed) to assess participant's adoption decisions and relationship of the argument to the status quo. In the second study, Cowgill, Dell'Acqua, and Matz (2020) measures the impact of adding scientific authority to arguments on AI ethics using a 2×2 factor design.

Cowgill, Dell'Acqua, and Matz (2020) finds that fatalism op-ed discourages AI adoption, while the counterfactual encourages it. As for the belief on the fixability of the algorithmic bias, participants tend to have more faith in correcting the fairness problems under fatalistic op-ed conditions. On the other hand, scientific veneer on arguments would reinforce the persuasive effect of the viewpoint on an approximately equal scale for either direction (positive or negative) and thus affect manager's decision to adopt AI.

Using the replication dataset and code provided by the original authors, I re-implement the first study to further the managerial impact of activism arguments through the working channel of opinion polarity in altering business decision-maker's perception and adoption decision on AI. This paper identifies the heterogenous effects from individual demographic characteristics on their consistent choice set of attitudes towards AI adoption (excluding treatment effects) via a multiple linear regression. Subsequently, it analyses the marginal effect of exerting engineering effort on the belief of AI bias fixability through a panel OLS regression with individual fixed effects.

My findings about the managerial effects by argument polarity are consistent with Cowgill, Dell'Acqua, and Matz (2020) that arguments stressing the inevitability of algorithmic bias lead managers to abandon AI and op-eds claiming the superiority of AI models relative to human in producing lower bias would encourage AI adoption. In addition, the counterfactual op-ed precondition would sway manager's belief in being able to improve AI bias after the engineering effort (fixability outcomes). For the ordinary linear regression

inference about demographic traits and status quo condition, this paper discovers that algorithm models are significantly less favored by female, African Americans, and politically liberal managers compared to other genders, race and political affiliations throughout the study.

Artificial Intelligence technology has a promising prospect in accelerating business decision-making process by delegating data-heavy insight mining and predicting optimal business plans provided that its ethical fairness problem can be mitigated in the future. Once the algorithmic bias can be addressed, AI would be one of the workhorses in the era of big data, especially from the perspective of business operations. Therefore, algorithmic fairness activism should promote the adoption of AI, allow some time for technology refinement despite its current defect in ethical bias. This paper builds on these incentives to investigate which activism intervention influences manager's adoption decision on the AI technology.

The remainder of the paper is constructed as follows. Section 2 describes the dataset, experiment design, and exploratory data analysis on feature visualization. Section 3 outlines the reproduced experiment models, which is designed to discover relationships between features and the target variable. Section 4 summarizes the model results according to evaluation criteria. Finally, Section 5 discusses our research findings and provides directions for future research.

R statistical programming language (R Core Team 2020) is used to replicate the experiment. To be specific, `tidyvers` package is for data preprocessing (Wickham et al. 2019), `kableExtra` package is applied to generate tables (Zhu 2020), `ggplot2` is used to draw diagrams (Wickham 2016), `ggthemes` is for diagram theme changing (Arnold 2021), and `lfe` is used to conduct panel linear regression with fixed effects (Gaure 2013).

2 Data

2.1 Dataset features

The paper intends to replicate the experiment report for Study 1 by Cowgill, Dell'Acqua, and Matz (2020) using the survey data provided. The panel dataset (op-ed) records 498 subject observations with 50 variables for two business cases, including demographic information about the subject (i.e. dummy for gender,[1] race[2]), education,[3] and AI-related experience,[4] indicator variables flagging experiment treatments,[5] and five survey answers at different experiment stages[6] in numeric scale.

Table 1: Variable names for raw data

hiringfirst	sqshw	sqopt	sqhid	consrv_std
cf	fat	sqseen_pre_first	sqseen_pre_second	posneg_pre_first
posneg_pre_second	posneg_post_first	posneg_post_second	posneg_growth_first	posneg_growth_second
lawsuitspr_pre_first	lawsuitspr_pre_second	lawsuitspr_post_first	lawsuitspr_post_second	lawsuitspr_growth_first
lawsuitspr_growth_second	damaging_pre_first	damaging_pre_second	damaging_post_first	damaging_post_second
damaging_growth_first	damaging_growth_second	recscale_pre_first	recscale_pre_second	recscale_post_first
recscale_post_second	recscale_growth_first	recscale_growth_second	recyesno_pre_first	recyesno_pre_second
recyesno_post_first	recyesno_post_second	recyesno_growth_first	recyesno_growth_second	male
female	othergend	hispanic	white	black
asian	otherethn	workdecisions	eduprepared	knowsml

The dependent variables of interest listed below are answers to five adoption-related survey questions given by subjects at different stages during the experiment, which proxy respondent's adoption decisions.

- `posneg`: Positive impact level the subject considers AI adoption (1-6)

[1] male, female, and other gender
[2] white, black, asian, hispanic, and other ethnicity
[3] eduprepared: whether feel educationally prepared for these decisions
[4] workdeicisions: whether make AI-related decisions on their jobs; knowsml: whether knows machine learning
[5] op-ed treatment: cf (counterfactual), fat (fatalistic); status quo condition: sqshw (shown), sqopt (optional), sqhid (hidden)
[6] pre or post "algorithm fix" condition; first or second business case

- `recscale`: Recommendation scale on AI adoption (1-7)
- `recyesno`: Whether recommend adopting AI technology (0 or 1)
- `lawsuitspr`: How likely are AI lawsuits or PR problems (1-7)
- `damaging`: Magnitude of damage by AI bias (1-7)

The following graphs explore some preliminary relationships between the participant demographic traits and their response to AI adoption in business decision-making experiments.

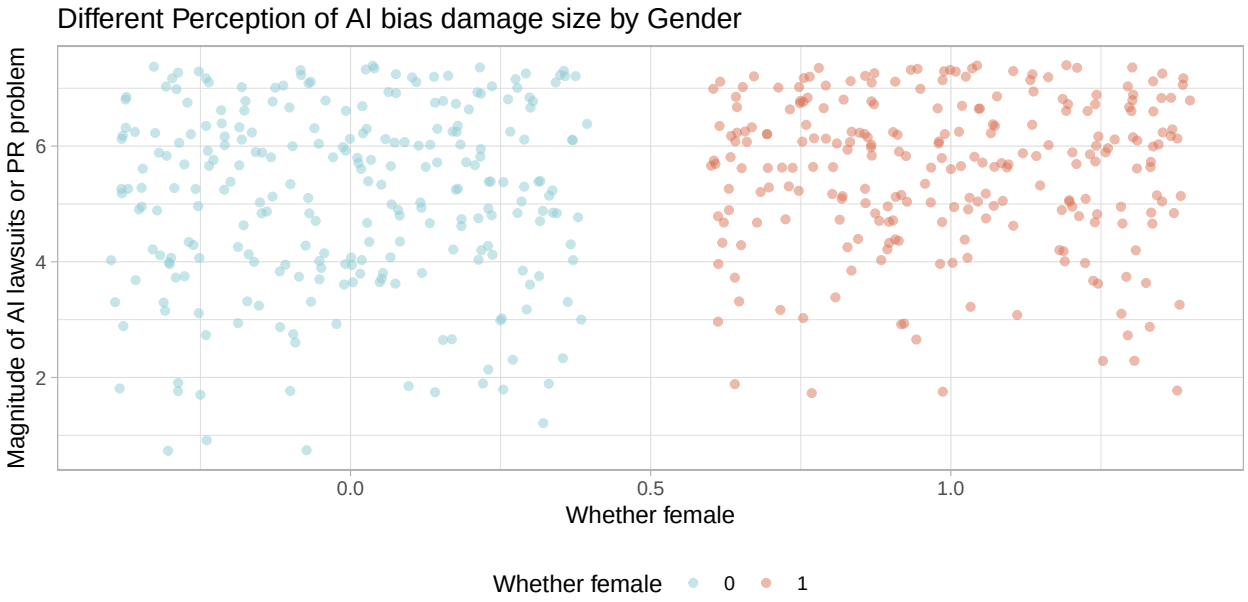

Figure 1: Different Perception of AI bias damage size by Gender

Gender vs. Damage size: We can see that female holds a more negative presumption on the damage size caused by AI bias; most responses concentrate around the more severe level around 6 to 7 and none of them rate the damage at its lowest level of 1.

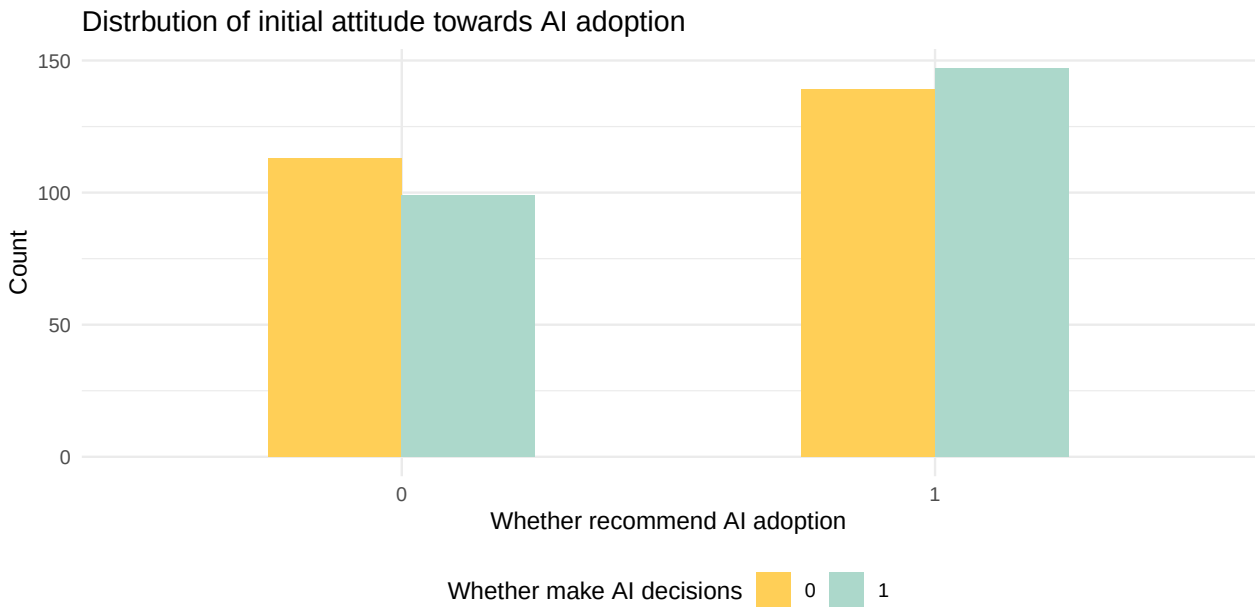

Figure 2: Distrbution of initial attitude towards AI adoption by AI experience

AI experience vs. whether recommend AI adoption: For managers with experience in making AI-related decisions, they tend to favor AI adoption more than their peers with no relevant AI experience.

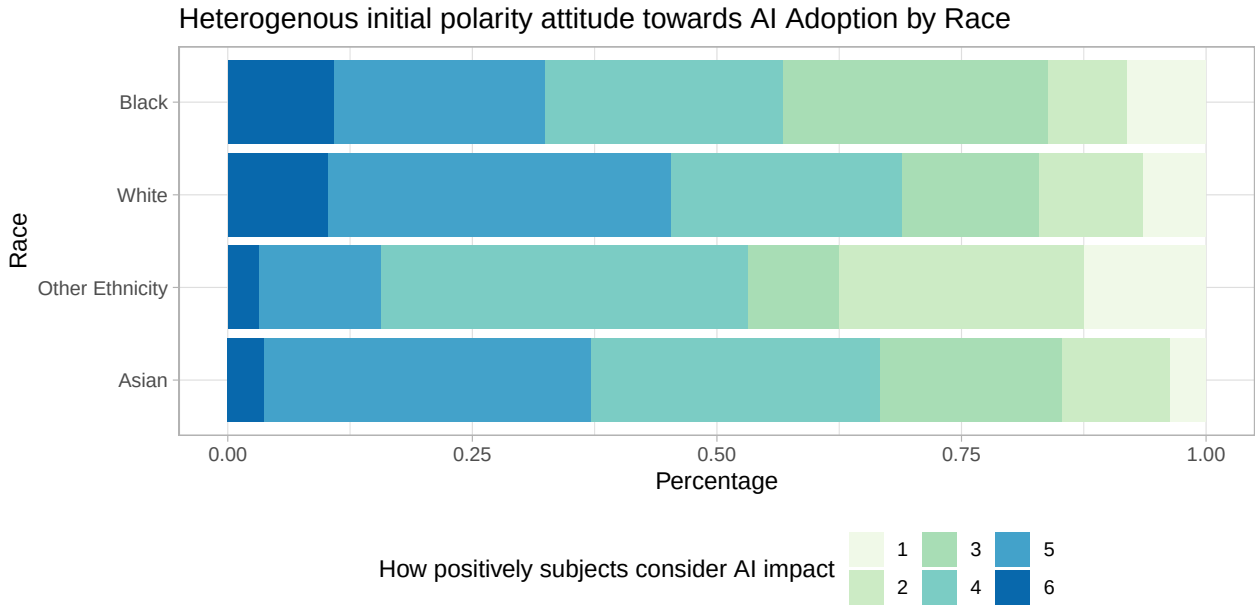

Figure 3: Heterogenous initial polarity attitude towards AI Adoption by Racer

Race vs. positive opinion on AI impact: People of different race have dissimilar opinions on AI impact. Besides other ethnicity, black people are the most pessimistic (1 to 2) towards AI impact, while white people are the most optimistic (5 to 6), followed by Asians, where most of them hold a positive vision on AI impact.

2.2 Population, frame, and sampling

The survey was conducted in January 2020, U.S. using the online participant recruitment platform called Prolific Academic. Target population is U.S. adults with management experience. Table 2[7] lists the detailed sample demographic statistics for gender, race, education and AI related experience traits in frequency. Apart from an equal selection in gender categories of male versus female, the majority race group is white, with latinx, black, Asian, and other ethnicity being the minority race groups, which resembles the population distribution characteristics in the U.S. Although Cowgill, Dell'Acqua, and Matz (2020) did not state their sampling methods in the paper, based on the selected subject distributions in demographics, it is probable that stratified random sampling (Jha (2017)) is used in participant recruitment to arrive at this representative subset of U.S. population. Half of the participants make AI-related decisions on their jobs. Around half of them know machine learning, and most of them feel educationally prepared for these decisions. There is no missing data in the dataset; respondents have answered every question encountered in the survey, i.e., non-response problem is non-existent in this study.

2.3 Experiment, methodology & intervention

For Study 1, Cowgill, Dell'Acqua, and Matz (2020) designed two business cases on recruiting or financial lending where subjects need to complete sequentially, with the same set of five AI adoption-related questions[8] to answer in each scenario. Each case puts the subject in the role of a manager making an AI adoption decision. After reading the case and before answering questions, subjects are randomized to receive **status quo** information on AI algorithmic imbalance levels (status quo shown, optional view, none). After these five questions, subsequently, the subject is informed with a new condition, **algorithm engineering fix**,.[9] They are then asked to re-answer the five questions, incorporating beliefs about the algorithm fixability given the sequential scenario of a six-month engineering effort to correct AI bias.

The **op-ed** treatment is placed in the second business case; subjects will be randomized to read an additional op-ed after viewing the case before receiving a randomized status quo information. The op-ed is manipulated to be a persuasive argument about algorithmic fairness randomized to be **"counterfactual"**,[10] **"fatalistic"**,[11] or no op-ed. Details about experiment flow design is in Appendix A.

Table 2: Descriptive Statistics: Subject Demographics

	Study 1: Op-Ed	Study 2: Veneer
Male	0.49	0.52
Female	0.49	0.45
Other Gender	0.02	0.02
Latinx	0.10	0.07
White	0.81	0.85
Black	0.07	0.05
Asian	0.05	0.04
Other Ethnicity	0.06	0.06
AI Decisions	0.49	0.40
Prepared by Educ.	0.75	0.71
Knows ML	0.46	0.45

[7] Table 2 uses the "dplyr" package Wickham et al. (2021) for data cleaning

[8] see Appendix

[9] The subject is told to imagine the company has dedicated 6 months of additional engineering effort to fix AI bias at the end of the case.

[10] A pro-AI view stressing the usefulness of AI once its bias is reduced below the level of "counterfactual" human decision-making

[11] An anti-AI perspective claiming that bias of AI cannot be fully cleansed and thus should be avoided

2.4 Data structure transformation and data preprocessing

The two sequential business case studies along with repeated process in question-answering make the observations to be in the format of longitudinal panel data, with time periods[12] to be `pre-` algorithm fix condition in the `first` case, `post-` algorithm fix condition in the `first` case, `pre-` algorithm fix condition in the `second` case, and `post-` algorithm fix condition in the `second` case. For the ease of further panel data regression analysis, the raw dataframe is reshaped into a long format, expanding the time dimension of four experiment stages for each subject.

In addition, the five dependent variables are standardized to be of Normal (0,1) distribution to unify the measurement scale for the ease of comparison in the regression results across five models. Interaction terms between sequential treatments are generated. The final cleaned dataset has 1,992 rows with 48 columns.

A change in participant's decision responses towards AI adoption over the experiment can be visualized below in the trend plot (Figure 4. We can see that on average, subjects perceive the AI adoption as more positive and less damaging after each algorithm engineering fix condition (pre vs. post), and their baseline preconceived idea on AI (pre_first vs. pre_second) seem to improve as well.

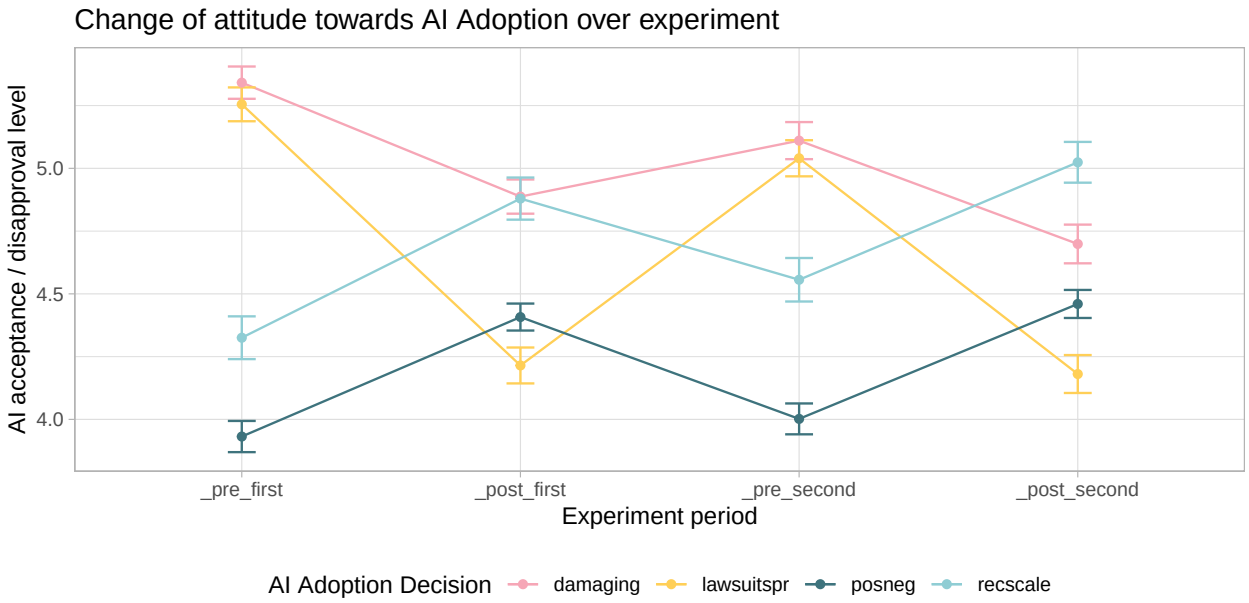

Figure 4: Change of attitude towards AI Adoption over experiment

3 Model

Cowgill, Dell'Acqua, and Matz (2020) used a multiple panel linear regression with fixed effects to examine the effects of activism treatment conditions on manager's AI adoption decisions through the channel of belief about algorithm fixability and persuasive arguments about AI fairness. The fixed effect absorbs all the individual-level response difference due to heterogeneity in individual characteristics. Based on Cowgill's model, a supplementary multiple linear regression is conducted by this paper to further investigate the impact of endogenous individual demographics heterogeneity on manager's decisions, which was masked by fixed effects in the previous model.

To measure subject's decision on AI adoption from the aspects of recommendation and perceived damage separate regressions are run on

[12] pre_first, post_first, pre_second, post_second

- `Positive`: how positive do they consider AI algorithm
- `Rec(Y/N)`: whether they recommend AI adoption
- `Rec(Scale)`: recommendation scale
- `Lawsuit`: how likely is the occurrence of lawsuits and PR problems
- `Damage Size`: the damage size due to AI bias conditional on it ever happening

Positive coefficients imply a positive effect on decision response for the first three dependent variable, and the opposite relationship holds for the last two Y variables.

3.1 Panel linear regression with individual Fixed Effect & Clustered SE

I replicate the STATA panel linear regression model which uses areg to absorb multi-level categorical factors with the `felm` linear model in the `lfe` package in RStudio. Due to the sequential and repeated procedure nature of the experiment, the survey data is collected over the dimensions of same individuals i (cross-sectional) and over time t (longitudinal). Therefore, the observed change in response variable might result from treatment effects after controlling for time or caused by heterogeneity in attributes across individuals but is time t invariant. Such entity difference might come from gender, race, education, and work experience and thus should be captured as individual fixed effect as FE_i in the regression model (Gaure 2013). It justifies the original author's rationale for using panel regression with fixed effects.

In contrast to the homoskedasticity assumption for linear regression, chances are that the variance of residuals for panel data are not constant for all data points due to autocorrelations across time within each individual and their dissimilar attributes. To account for heteroskedasticity, clustered robust standard errors are used instead of standard errors to get the unbiased standard errors of regression coefficients and to obtain correct p-value for statistical inference (Cameron and Miller 2015).

The equation for the panel regression model is shown below:

$$Y_{i,t,fix} = \beta[1(t=1) + 1(fix=1) + Hiring + FE_i + CF_{i,t} + FAT_{i.t} +$$

$$CF_{i,t} \times 1(fix=1) + FAT_{i,t} \times 1(fix=1)] + \epsilon$$

The treatment variables of `engineering fix on the algorithm` (fix) condition and the `counterfactual` (CF) or `fatalistic` (FAT) op-ed condition and their interaction terms are included in the model 1 panel regression. The order of the business case t and its content ($hiring$ or lending) are included as control variables.

3.2 Multiple linear regression with interaction terms

The equation for the multiple linear regression is specified as follows:

$$Y_{i,t,fix} = \beta[1(t=1) + 1(fix=1) + Hiring + SqShown_{i,t} + SqOpt_{i,t} + SqShown_{i,t} \times 1(fix=1) +$$

$$SqOpt_{i,t} \times 1(fix=1) + SqSeen_{i,t} \times 1(t=1) + Female + Black + Asian + Hispanic + OtherEthn +$$

$$PoliticConsrv + AIDecision + EducPrepared + KnowsML] + \epsilon$$

The individual-level fixed effect in the first panel regression model absorbs the time t-invariant effects such as individual attributes (i.e. gender, race, education, work experience, etc.) and the randomized status quo viewing conditions. The status quo conditions are irrelevant with case t ($t \in \{1, 2\}$) since they are present in both cases and the randomization make the two conditions independent from each other.

To study the effects generated by individual attributes and status quo conditions within the fixed effect, the factors of interest are included in this supplementary multiple linear regression as my second model.

To study within the individual subject heterogeneity, demographic information of gender, race, education, and political conservatism and AI- related experiences in work and knowledge are specified as regressors

on subject's response to the survey question. In addition to the treatment variable of `algorithm fix`, the randomized status quo conditions and their interaction terms with the treatment are included to serve as control variables.

To account for the subject heterogeneity and its resulting heteroskedasticity of regression errors, standard errors are clustered by subject `response id` for accurate standard errors. Accurate standard errors would in turn deliver unbiased p-values for regression coefficients which are a critical basis for statistical inference on the variable significance.

4 Results

4.1 Model 1: Effects of "Counterfactual" and "Fatalism" Activism

As shown in the summary table, Table 3[13], for panel linear regression, the model goodness of fit is rather high, at R-squared between 0.6 to 0.7. The constructed model can explain 60 ~ 70% of the variations in subject's response on AI adoption decisions.

Reading the counterfactual op-ed leaves the managers with a more positive impression on the algorithm and encourages the adoption of AI. It increases the perceived positive impact of the technology by 0.24 standard deviations and significantly increases recommendation by 0.14 standard deviations. The opposite effect is observed when managers read the fatalistic op-ed. Managers deduct marks on their impression towards AI by 0.17 standard deviations and discourages adoption significantly (-0.28). Furthermore, their pessimistic vision on AI bias leads them to worry more about the likelihood of lawsuits and PR problems (+0.19). The control group for op-ed treatment is having no op-ed.

As for the algorithm `fix` treatment, it can significantly improve people's attitude on their perceived algorithm positivity (+0.37), scale of recommendation (+0.27), and damage incidents (-0.6) and size (-0.28), but with limited efficacy on altering people's binary decision choice of yes or no (0.02, p-value insignificant).

When we consider the joint treatment effects from algorithm fix and op-ed, it is observed that the counterfactual op-ed significantly deters managers from recommending the adoption (-0.28) but deepen their concern on the damage incidents (0.16), compared to the no op-ed situation. This may partly be the result of having set initial expectations relatively high in the counterfactual condition. The effects of the fatalism op-ed about algorithm fix are mixed; subjects are holding ambivalent attitudes towards technology perceived positivity and recommend scale, and only the impact on perceived damage incident seems significant but insubstantial (+0.1). It is probably because that the vision promised by the engineering effort to fix AI bias and the AI fatalism viewpoint are contradicting themselves, which causes confusion for the readers and thus affect their decision-making inconsistently.

Note that although my replication panel regression model delivers the same regression coefficients as the original paper, there are some disparities in the magnitude of the clustered standard errors. The standard errors from my replication are slightly smaller (by around 0.005), as a result, the detected significance level is slightly biased upwards, but no material effect on the threshold of 5%. It turns out the STATA uses a different clustering technique from R, called finite sample correction (Masterov 2018).

[13]The 'modelsummary' package Arel-Bundock (2021) is applied to create the table.

Table 3: Effects of Activism: Study 1 ("Counterfactual" and "Fatalism" Activism)

	Positive	Rec(Scale)	Rec(Y/N)	Lawsuit	Damage Size
Hiring	-0.088	-0.055	0.000	-0.040	-0.285***
	(0.075)	(0.087)	(0.000)	(0.081)	(0.078)
Algorithm Fix	0.367***	0.268***	0.019	-0.613***	-0.279***
	(0.030)	(0.023)	(0.041)	(0.035)	(0.023)
Counterfactual Op-ed	0.239**	0.188*	0.139***	-0.124	-0.142
	(0.099)	(0.099)	(0.053)	(0.109)	(0.108)
AI Fatalism Op-ed	-0.165*	-0.284**	-0.079	0.193**	0.061
	(0.096)	(0.110)	(0.081)	(0.093)	(0.078)
Fix x Counterfactual	-0.042	-0.006	-0.277***	0.159***	0.041
	(0.052)	(0.032)	(0.107)	(0.058)	(0.056)
Fix x AI Fatalism	-0.044	0.012	0.158	0.100**	0.012
	(0.050)	(0.036)	(0.162)	(0.044)	(0.029)
Num.Obs.	1992	1992	1992	1992	1992
R2	0.678	0.722	0.755	0.627	0.734
Fixed Effects	Subject	Subject	Subject	Subject	Subject

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

4.2 Model 2: Effects of Demographics and Status Quo Conditions

Consistent patterns can be observed for people in demographic subgroups under different status quo and algorithm fix conditions in Table 4. The linear regression model using demographics traits, status quo and algorithm fix as regressors can explain about 10% of the variations in subject's response.

Subjects who have viewed the status quo about AI fairness (either by being shown or choosing to click on the "view" option) assess the algorithm more favorably in their response to the five questions on a 5% significance level. Managers with AI work experience or someone who feel educationally-prepared for these decisions also hold an approving opinion. On the other hand, women, African Americans, and politically liberal managers tend to avoid this idea of AI adoption.

In addition, significant evidence has been found that engineering effort to fix the algorithm can brighten up people's belief on the algorithm bias fixability. A six-month focused engineering effort can effectively increase manager's evaluation on algorithm positivity and recommendation scale by 0.47 and 0.45 standard deviations. It also eases their concern on lawsuits and damage size by 0.7 and 0.34 standard deviations respectively.

Table 4: Status Quo Conditions and Demographics: Study 1 ("Counterfactual" and "Fatalism" Activism)

	Positive	Rec(Scale)	Rec(Y/N)	Lawsuit	Damage Size
Algorithm Fix	0.471***	0.453***	0.000	-0.699***	-0.338***
	(0.057)	(0.051)	(0.000)	(0.061)	(0.044)
Status Quo Shown	0.522***	0.651***	0.571***	-0.430***	-0.442***
	(0.103)	(0.100)	(0.094)	(0.095)	(0.099)
Status Quo Shown (only if clicked)	0.459***	0.635***	0.543***	-0.381***	-0.425***
	(0.106)	(0.099)	(0.096)	(0.092)	(0.098)
Fix x Status Quo Shown	-0.169**	-0.257***	0.000NA	0.191**	0.115*
	(0.072)	(0.063)	()	(0.084)	(0.061)
Fix x Status Quo Shown (only if clicked)	-0.190**	-0.303***	0.000	0.205**	0.091
	(0.076)	(0.062)	(0.000)	(0.082)	(0.061)
Female	-0.158**	-0.109	-0.147**	0.107	0.286***
	(0.070)	(0.072)	(0.071)	(0.066)	(0.071)
Black	-0.250*	-0.218*	-0.048	0.176	0.224*
	(0.138)	(0.132)	(0.131)	(0.111)	(0.130)
Asian	-0.165	-0.027	-0.084	-0.026	-0.162
	(0.134)	(0.134)	(0.160)	(0.138)	(0.149)
Other Ethnicity	-0.332***	-0.262*	-0.212	0.083	-0.086
	(0.124)	(0.142)	(0.164)	(0.140)	(0.152)
Political Conservatism (Standardized)	0.094**	0.082**	0.026	-0.067**	-0.079**
	(0.038)	(0.037)	(0.036)	(0.034)	(0.038)
AI Decisions	0.121*	0.121*	0.078	0.025	0.019
	(0.069)	(0.072)	(0.075)	(0.067)	(0.074)
Prepared by Educ.	0.131	0.139*	0.099	-0.123*	-0.115
	(0.080)	(0.083)	(0.082)	(0.075)	(0.080)
Knows ML	0.005	0.003	-0.056	0.062	-0.097
	(0.070)	(0.074)	(0.074)	(0.068)	(0.076)
Num.Obs.	1992	1992	1992	1992	1992
R2	0.103	0.122	0.121	0.117	0.097

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

5 Discussion

The industries have benefitted from the development of Artificial Intelligence technology in liberating manual efforts, replacing rule-based models, and facilitating the decision-making process in business operations, meanwhile they also live under the fear of lawsuits and negative PR inflicted by algorithmic bias. The purpose of the research is to educate the business decision-maker of the future workhorse in the big data era and consequently to promote the business adoption in AI technology.

This paper replicates the first experiment of Cowgill, Dell'Acqua, and Matz (2020) to further study the managerial impact of AI fairness activism through the channel of its belief manipulation on impression on technology impact and manager's perception on algorithm fairness fixability. The heterogeneity in adoption response by subject traits is examined with a multiple linear regression as a supplementary model to identify subgroup heterogeneity within the panel fixed effects.

5.1 Bias and ethical concerns

Despite the fact that Cowgill, Dell'Acqua, and Matz (2020) paper centers around the topic of AI algorithmic fairness, there are still some potential ethical biases in its experimental design.

5.1.1 Imbalanced sampling proportion of race

The random stratified sampling method replicates the real U.S. demographic distribution on the sample distribution for the recruited subjects, it also introduces data imbalance. Data imbalance happens when the

number of observations per class is not equally distributed; often one class, the majority group, has a large amount of data while one or more other classes (minority groups) have much fewer observations.

As shown in Table 2, the majority in race group are white (81%), leaving black, Asian, and Hispanic to be the minority groups; each taking up around 5% of the population by race. Data imbalance is a less explored problem in statistical regressions than in the domain of machine learning classification. Many statistical models simplify under the assumption of balanced data such as ANOVA, which is closely related to experimental design (Henry and O 2018).

In regression, the problem with imbalanced data is that essentially the sample size for the minority group is much smaller than the number of subjects in the experiment. The data points in the small sample and the observed behavior pattens might not be representative enough for the whole minority race group, which would cause bias in regression coefficients and statistical inference.

5.1.2 Additional demographic factors to consider

Besides gender, race, political conservatism, and education, other factors that also can proxy people's acceptance level on new technology are age and industry working for. Age for the working class embodies their years of working experience and seniority of employment (junior, senior, about to retire), which can indicate people's willingness to be exposed to and learn about new things. Eventually, their learning attitude plays a role in shaping people's belief and altering decision-making process.

For the specific topic of Artificial Intelligence, depending on the sectors being applied to, AI brings different scales of working efficiency increase and varying risk implications. According to Shah (2019), AI is most applied in the industries of healthcare, education, marketing, retail and E-commerce, and financial services. Managers from different industries might bring their implicit context into the business case setting. To account for these omitted factors, age and working industry information should be collected and added to the model.

5.2 Model results

5.2.1 Effects of "Counterfactual" and "Fatalism" Activism

Based on model 1 regression results, the counterfactual arguments on AI fairness significantly encourages AI adoption, especially through promoting the binary recommendation, by 0.14 standard deviation. The fatalistic op-ed dispels people's consideration in adopting the technology by persuading them with the increased likelihood of lawsuits and PR problems.

5.2.2 Heterogeneity managerial effects from demographic differences

As for the multiple OLS regression on subject's response towards AI adoption, heterogenous effects in managers from different demographic groups are observed. All else held constant, on average, female managers view AI impact as less positive; furthermore, they are less likely to recommend AI adoption as they would picture a much larger damage (+0.29) caused by AI bias. Similar but less extreme behaviors (in terms of coefficient magnitude and statistical significance) are also observed in the African American group. For people who are politically conservative, they welcome the adoption of AI, but to a limited extent.

5.2.3 Belief on the fixability of AI bias

The study intends to manipulate people's belief on the fixability of AI bias by telling them about a six-month of focused engineering fix. Their change in responses is measured by the repeated survey questions after the algorithm `fix` scenario. From both regression models we can see that algorithm fix is effective in persuading people of the positive algorithmic impact, increasing recommendation scale, and dispelling their

negative concerns. However, the algorithm fix condition is futile to sway people's final decision on whether to recommend adoption; its coefficient in the binary recommendation question is close to 0.

In Model 1, the joint information of counterfactual op-ed about fixability makes the business decision-makers think that fairness problems have less room to be fixed compared receiving no op-ed. In addition, the fatalism op-ed about algorithm fix brings about ambivalent persuasive impact on subject's decision-making behavior.

5.3 Real world implications

Based on the observed effects in AI adoption related response from subjects in the sequential business case studies, this paper confirms that persuasive arguments can have real impact on business decision-maker's AI adoption decision through altering people's belief on algorithmic fairness problem and its fixability. As for the reason behind their choice, the study finds that AI abandonment is correlated with fear of lawsuits and negative PR.

For the activists and policymakers looking to encourage AI adoption through activism campaigns, one of the most effective ways is to use the counterfactual arguments which emphasizing AI's eminence in reducing human bias. To address the relatively less approving attitude among women, African American, and politically liberal managers, a pilot AI activism program can be held to educate them on the edge of AI algorithm to assure them of its promising landscape and controllable risks involved.

5.4 Internal validity & external validity of model

5.4.1 Internal validity

The internal validity of the model can be justified by the procedures of sampling and RCT (randomized control trial) in experiment interventions, so that alternative explanations but the treatment can be eliminated from the cause of effect on the outcome.

Sampling: Participant recruitment targeted adults with management experience nationwide in the U.S. Subjects were stratified sampled so that the randomization in sampling simulates real world demographic distributions.

Randomization in conditions, treatment & control, and business case order: The randomization in assignment of status quo conditions mimics the real-life scenario where some people are shown the information, some choose to view it, while the rest do not have access. The intervention of receiving what kind of persuasive argument on AI fairness (counterfactual, fatalistic, or none) was randomized to avoid this viewpoint biasing their beliefs on algorithm fixability and thus the experiment outcome (Meldrum 2000). The order of the sequential business cases were randomized to ensure that the variations on outcome does not come from the order of the case content but the experiment interventions.

5.4.2 External validity

External validity of the conclusions on AI activism was supported by the comparison of outcome results across five dimensions relating AI adoption decisions (Devroe, n.d.). Thereby, the observed outcomes are applicable to all the perspectives in manager's response, from thoughts on algorithm positivity, scale of recommendation, to damage size. The sampling frame covers a range of nationwide population, so that the study conclusion and subject trait summary can be extrapolated to the domestic level. The generalizability of the model findings can be enhanced by adding various new business topics in the case studies, such as topics on medical care and education to make the model discussion more inclusive.

5.5 Weakness and opportunities for future work

5.5.1 Weakness

Data imbalance within demographic subgroups: Among 498 experiment subjects, white people account for 81% of the race groups, while the minority races (i.e. Asian, black, Hispanic) constitute only 5-10% of the sample total respectively. The decision-making behavior from the minority race groups have not been fully delineated due to a lack of data points in the regression model.

Discussion on proxy variables for new technology acceptance: For the working class, besides the demographic information collected by the study, age and industry can proxy people's acceptance level on new technology. Years of working experience and seniority of employment can indicate people's willingness to learn about new things, which plays a role in shaping people's belief and decision-making. Managers from different industries might have difference tolerance level for algorithmic mistakes. New technology acceptance level should be taken into consideration when studying business decision-maker's response on AI adoption.

Bad variable choice for Instrumental Variable: In the original paper, Cowgill, Dell'Acqua, and Matz (2020) use op-ed conditions, counterfactual and fatalism to serve as instrumental variables for the variable `status quo seen in the second case`. For a factor to a valid instrumental variable for an endogenous variable, it needs to have a causal effect on this endogenous variable, and it cannot directly influence the outcome variable. Since in the second business case, subjects are randomly assigned with the op-ed treatment before the status quo viewing condition, the viewpoints in the op-ed arguments can affect those who has the option to view the status quo or not, the first IV assumption is satisfied.

However, the latter, exclusion restriction (Labrecque and Swanson 2018), is not satisfied in this case. In fact, op-ed is the experiment treatment of interest that can directly influence subject responses through altering their belief on AI fixability; it is a confounder that impacts both the status quo seen variable and the outcome variable of adoption response.

5.6 Differences and difficulties

5.6.1 Differences

Supplementary multiple linear regression model: Model 2 explores time t invariant demographic s heterogeneity and status quo conditions within the subject-level fixed effects from the replication model. It examines the components of fixed effects in detail by the dimensions of demographic traits and status quo conditions.

Skip the replication of IV specifications: Cowgill, Dell'Acqua, and Matz (2020) use the op-ed conditions as instrumental variables for `status quo seen second`. However, the variable op-ed does not satisfy the exclusion restriction which require op-ed to affect subject responses only through the channel of status quo seen in the second business case. In fact, op-ed is the experiment treatment of interest that can directly influence subject responses through altering belief on AI fixability. Therefore, I chose to not replicate the IV regression model on the use of counterfactual information.

5.6.2 Difficulties

Exact replication of the STATA clustered error: For the panel linear regression with individual-level fixed effect and clustered standard error, there seems to be no library or package in R that would output the exact same standard errors as STATA `areg` clustered error function. The `felm` linear model in the `lfe` library by R in can replicate fairly close clustered standard errors (slightly smaller by 0.005); it does not affect the statistical significance inference at 5% level.

Create experiment flow chart from scratch: Given the complex nature of the experiment procedures, ambiguous instructions on details and unclear wordings in some of the experiment descriptions, it took me multiple rounds of reading to figure out the experiment settings and several drafts before finalizing on the flow chart shown in appendix.

5.6.3 Opportunities

In term of experimental design, alternative sampling methods such as random sampling and oversampling can be used to obtain a more balanced dataset for different races to fix the data imbalance problem. Extra demographic information such as age and industry of work can be collected during survey to construct proxy variables for people's acceptance level on new technology, which serve as additional regressors in the model.

As for improvements in model design, the difference-in-difference technique can be used to model the panel data controlling for individual fixed effect and time fixed effect. Difference-in-difference compares the average change over time in the outcome variable in the treatment group with the control group. The DiD modeling is suitable for the study of "policy" effect (Zeldow and Hatfield 2019), in our case, the treatment of persuasive argument on AI activism.

A Appendix

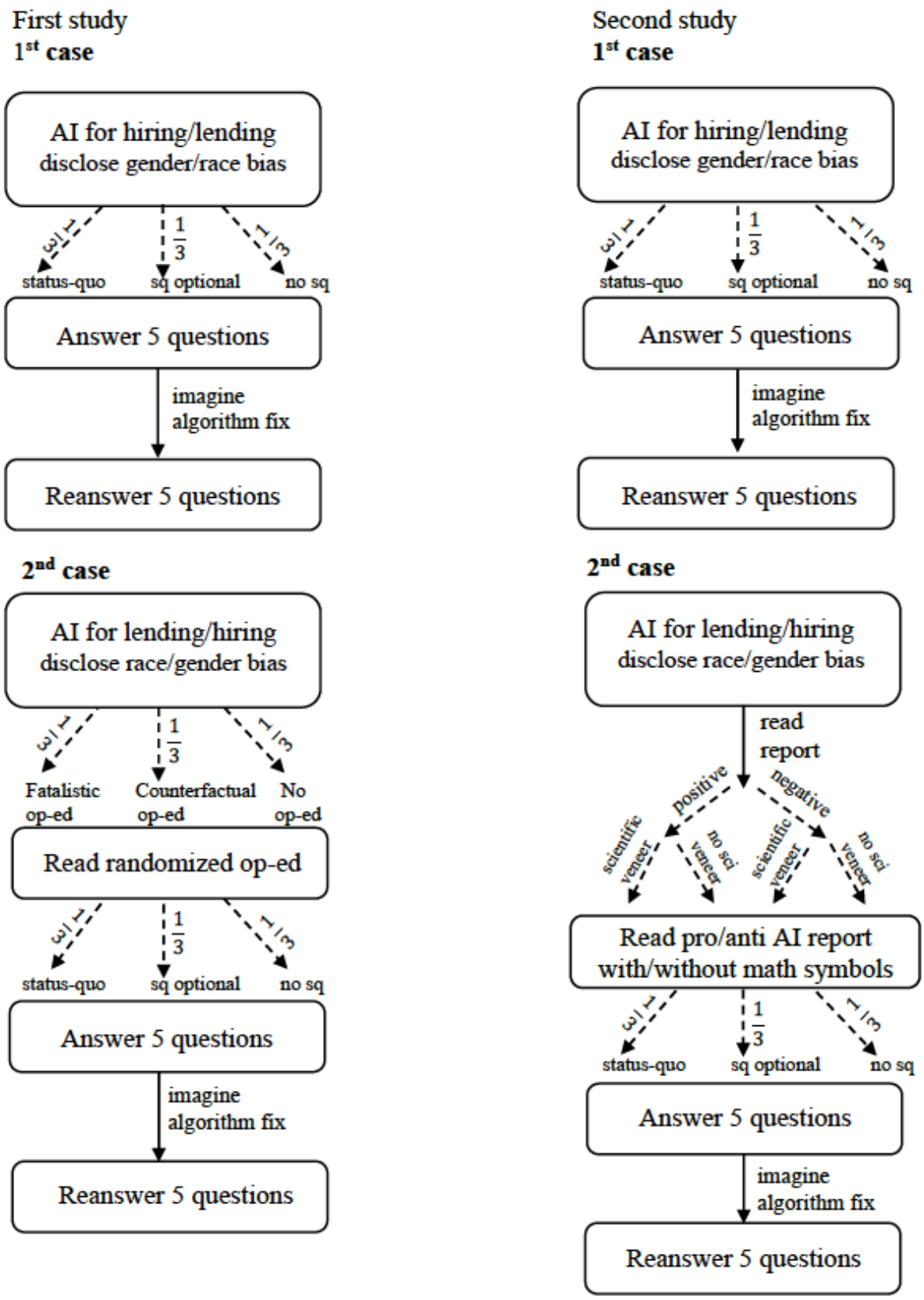

Figure 5: Overall experiment flow

Five survey questions in Study 1

- How positively do you consider the impact of AI technology? (1-7)
- On a scale 1-7, how much do you recommend adopting AI technology?
- Do you recommend adopting AI? (0/1)
- On a scale 1-7, how likely are lawsuits or PR problems by AI algorithmic bias?
- How large is the damage size of AI lawsuits or PR problems if they ever occur (1-7)?

B Regression Specifications

B.1 Model 1 Panel linear regression: `Counterfactual` and `Fatalism`

$$Y_{i,t,fix} = \beta[1(t=1) + 1(fix=1) + Hiring + FE_i + CF_{i,t} + FAT_{i.t} +$$
$$CF_{i,t} \times 1(fix=1) + FAT_{i,t} \times 1(fix=1)] + \epsilon$$

where:

- $Y_{i,t,fix}$: the response to question Y individual i on case t regarding the engineering effort to fix. Separate regressions are conducted for each of the five questios.
- $i \leq 498$: index for 498 subjects in Study 1
- $t \in \{1, 2\}$: order of the business case (first or second)
- $fix \in 0, 1$: indicator variable of whether before (fix=0) or after ($fix = 1$) the engineering effort to fix algorithm
- $hiring$: indicator variable of whether being the hiring business case
- FE_i: individual-level fixed effect
- CF: dummy variable of whether reading the counterfactual op-ed condition in the second case
- FAT: dummy variable of whether reading the fatalistic op-ed condition in the second case
- No op-ed is the excluded condition

B.2 Model 2 Multiple linear regression: `Status Quo Conditions` and `Demographics`

$$Y_{i,t,fix} = \beta[1(t=1) + 1(fix=1) + Hiring + SqShown_{i,t} + SqOpt_{i,t} + SqShown_{i,t} \times 1(fix=1) +$$
$$SqOpt_{i,t} \times 1(fix=1) + SqSeen_{i,t} \times 1(t=1) + Female + Black + Asian + Hispanic + OtherEthn +$$
$$PoliticConsrv + AIDecision + EducPrepared + KnowsML] + \epsilon$$

where:

- $SqShown$: dummy variable of whether the subject is shown with the status quo (yes:1, no: 0)
- $SqOpt$: dummy variable of whether the subject is assigned to the optional status-quo-viewing conditon group
- $SqSeen$: dummy variable of whether the subject have viewed the status quo (either being shown the status quo or opt to click on it)
- $PoliticConsrv$: subject's political conservativeness level (standardized)
- $AIDecision$: dummy for whether the subject make AI-related decisions on their jobs
- $EduPrepared$: dummy variable for whether the subject feels educationally-prepared

References

Arel-Bundock, Vincent. 2021. *Modelsummary: Summary Tables and Plots for Statistical Models and Data: Beautiful, Customizable, and Publication-Ready.* https://CRAN.R-project.org/package=modelsummary.

Arnold, Jeffrey B. 2021. *Ggthemes: Extra Themes, Scales and Geoms for 'Ggplot2'.* https://CRAN.R-project.org/package=ggthemes.

Cameron, A. Colin, and Douglas L. Miller. 2015. "A Practitioner's Guide to Cluster-Robust Inference." *J. Human Resources*, 317–72. https://doi.org/10.3368/jhr.50.2.317.

Cowgill, Bo, Fabrizio Dell'Acqua, and Sandra Matz. 2020. "The Managerial Effects of Algorithmic Fairness Activism." *AEA Papers and Proceedings* 110 (May): 85–90. https://doi.org/10.1257/pandp.20201035.

Devroe, Robin. n.d. "How to Enhance the External Validity of Survey Experiments? A Discussion on the Basis of a Research Design on Political Gender Stereotypes in Flanders." In *ECPR General Conference*. https://core.ac.uk/download/pdf/74712652.pdf.

Gaure. 2013. *Lfe: Linear Group Fixed Effects. The R Journal*. Vienna, Austria.

Henry, and Adam O. 2018. "When Is Unbalanced Data Really a Problem in Machine Learning?" Cross Validated. https://stats.stackexchange.com/q/283942.

Jha, Gaurav. 2017. "6 Sampling Techniques: How to Choose a Representative Subset of the Population." https://humansofdata.atlan.com/2017/07/6-sampling-techniques-choose-representative-subset/.

Labrecque, Jeremy, and Sonja A. Swanson. 2018. "Understanding the Assumptions Underlying Instrumental Variable Analyses: A Brief Review of Falsification Strategies and Related Tools." *Springer*. https://doi.org/10.1007/s40471-018-0152-1.

Masterov, Dimitriy V. 2018. "Clustered Standard Errors Are Completely Different in R Than in Stata." Cross Validated. https://stats.stackexchange.com/q/359047.

Meldrum, Marcia L. 2000. "A Brief History of the Randomized Controlled Trial: From Oranges and Lemons to the Gold Standard." *Hematology/Oncology Clinics of North America* 14: 745–60. https://doi.org/https://doi.org/10.1016/S0889-8588(05)70309-9.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Shah, Hardik. 2019. "5 Industries That Are Using Artificial Intelligence the Most." *Datafloq*. floq.to/ahGXH.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Zeldow, Bret, and Laura Hatfield. 2019. healthpolicydatascience. https://diff.healthpolicydatascience.org/.

Zhu, Hao. 2020. *KableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.