The Value of Humanization in Customer Service

Yang Gao
Simon Business School
University of Rochester
yang.gao@simon.rochester.edu

Huaxia Rui
Simon Business School
University of Rochester
huaxia.rui@simon.rochester.edu

Shujing Sun
Naveen Jindal School of Management
University of Texas at Dallas
shujing.sun@utdallas.edu

Abstract

As algorithm-based agents become increasingly capable of handling customer service queries, customers are often uncertain whether they are served by humans or algorithms, and managers are left to question the value of human agents once the technology matures. The current paper studies this question by quantifying the impact of customers' enhanced perception of being served by human agents on customer service interactions. Our identification strategy hinges on the abrupt implementation by Southwest Airlines of a signature policy, which requires the inclusion of an agent's first name in responses on Twitter, thereby making the agent more humanized in the eyes of customers. Multiple empirical analyses consistently show that customers are more willing to engage, and upon engagement, more likely to reach a resolution, with more humanized agents. Furthermore, we find that customers do not behave more aggressively to more humanized agents, hence humanization incurs no additional cost to agents.

1. Introduction

Consider a disgruntled customer who complains to a customer service agent without any face-to-face interaction, through email or live chat for example. In the first scenario, the agent is a real human being, while in the second scenario, the agent is an artificial intelligence (AI) algorithm. For the customer, does knowing which scenario he is in affect his behavior? In other words, even if there is little difference regarding how agents respond in these two scenarios, will the *binary* information of whether the agent is an actual human or an algorithm change how the customer behaves? For instance, will the customer be less engaged or less satisfied simply because he is not dealing with a human being?

With rapid advances in AI, the aforementioned scenario is no longer futuristic or merely philosophical. Increasingly, managers are faced with the looming question of how much automation should be incorporated into their customer service operations. The question is rooted in two seemingly incompatible prescriptions for the future of customer service provision. On the one hand, since call centers have long been perceived as cost centers,[1] firms have been leveraging information technologies for years to deliver customer service as cost-effectively as possible. The recent development of AI chatbot technology presents the latest opportunity and probably the ultimate solution to such a quest for cost reduction.[2] On the other hand, anecdotal evidence suggests that customers prefer to engage with human agents in the context of customer service. While current limitations of AI technology constitute an important cause of customers' preference for engaging with human agents, these limitations will likely be overcome as AI technology matures. However, the hypothetical scenarios discussed at the beginning of the paper may foreshadow a fundamental limitation of AI applications in customer service: *do emotionally-charged customers have an inherent preference for engaging with human agents over algorithmic agents?*

Intrigued by this question and motivated by the hypothetical scenarios, we study how customers' changing perception regarding the probability of the two scenarios affect their behavior, in terms of the *outcome* and the *process* of their interactions with agents. More specifically, we exploit a quasi-experiment induced by a policy change from Southwest Airlines on Twitter. Starting from March 16, 2018, all customer service agents of Southwest Airlines include their first names in their service responses on Twitter. We refer to this

[1] According to IBM, 265 billion customer services are requested every year, and it costs companies $1.3 trillion to address these requests., For details, see https://www.ibm.com/blogs/watson/2017/10/how-chatbots-reduce-customer-service-costs-by-30-percent/

[2] AI is predicted to power 95% of all customer interactions by 2025, including live telephone and online conversations. For details, see https://www.financedigest.com/ai-will-power-95-of-customer-interactions-by-2025.html

URI: https://hdl.handle.net/10125/71293
978-0-9981331-4-0

sudden change as the **signature experiment** because the inclusion of a first name (henceforth referred to as the *signature*) essentially signifies the authorship of a response from a live human being, as opposed to a chatbot or a corporate script. The change also seems exogenous from the customers' perspective because there is no prior notice or hint from Southwest Airlines. Customers can neither anticipate nor influence the timing of the change. By observing a signature, customers are more likely to perceive their social media customer service interactions as between humans. Using the example of the two hypothetical scenarios, we expect the *subjective probability* for the first scenario (i.e., interaction with a real human being) to be higher after the launch of the signature experiment.

To measure the *outcome* of a social media customer service interaction, we construct two variables. First, we consider a customer's decision to continue engaging with an agent upon receiving the agent's initial response as a measure of the customer's willingness to engage. In many cases, a customer's follow-up is the prerequisite for an agent to continue the service. Conditional on a customer's follow-up, we further examine whether the conversation leads to a resolution, which is based on manual annotation and a supervised machine learning classifier we developed. Second, to measure the *process* of a social media customer service interaction, we focus on verbal aggression, which is a well-recognized customer misbehavior detrimental to the cognitive and task performance of customer service agents [1, 2]. We infer customers' verbal aggression from the usage of profanity words in their communications with agents.

We carry out the empirical analyses in two steps. First, we follow the one-group before-and-after design [3, 4] to detect whether there is any change in the outcome and process of customer service interactions following the signature experiment. We find that, with the enhanced humanization induced by the signature, customers are more willing to engage and the likelihood of reaching a resolution is also higher. Moreover, we find no evidence that customers behave either more or less aggressively after the change. These findings remain consistent in various robustness checks and falsification tests. Second, to minimize the chance that our findings are driven by unaccounted time-varying factors, we conduct difference-in-differences analyses by constructing a synthetic control group that is similar to the Southwest Airlines [5, 6, 7] based on the donor pool consisting of three other major U.S. airlines (i.e., American Airlines, Delta Airlines, and United Airlines). Additionally, we also propose a conversation-level two-way matching procedure, which is in the spirit of the synthetic control method but implemented at a more granular level. Both sets of results consistently support our findings.

2. Literature Review

2.1. Social Media Customer Service

Social media customer service has drawn increasing attention from IS and marketing researchers over recent years. One stream of this literature focuses on the customer side [8, 9, 10, 11]. For example, using a dynamic choice model and accounting for customer relationships with the firm, Ma et al. found that redress seeking is a major driver of customer complaints. While service intervention can improve customers' relationships with the firm, it also increases individuals' propensity to complain in the future [8]. Another stream of this literature focuses on the firm side [12, 13]. For example, by analyzing over three million tweets to seven major U.S. airlines, Gunarathne et al. found that airlines are more likely to respond and respond faster to customers with higher social media influence [12]. By studying the value of humanization in the context of social media customer service, the current paper introduces an important and novel angle to this literature and connects it with the emergent literature on the implications of AI.

2.2. Conversational AI

Most works in this literature stream are from the computer science field, with a particular focus on improving the algorithmic performance of chatbots [14, 15, 16]. Meanwhile, some computer science researchers explored how people interact with a chatbot [17, 18]. For instance, Corti and Gillespie found that people are more likely to repair misunderstandings when speaking to an algorithmic agent represented in a human body interface compared with one in a text screen interface [17]. Business researchers have explored the potential of AI-based customer service from various perspectives [19, 20, 21]. For example, Xiao and Kumar proposed a conceptual framework that includes the antecedents (e.g., customer acceptance of robots) and consequences (e.g., service quality) of firms' adopting and integrating robotics in their customer service operations [21]. To the best of our knowledge, the current paper is the first to quantitatively study the value of humanization in customer service which is the key difference between human agents and algorithmic agents.

3. Hypotheses Development

3.1. Humanization and Service Outcome

The perception that a service agent is a human agent, rather than an algorithmic agent, is associated with the perception of the agent's expertise. For instance, customers consider human sales agents as more knowledgeable than technology-based chatbots [22] and they also resist medical AI because of the belief that AI provides inferior care compared to human providers [23]. Such a belief is likely rooted in people's past experience of the poor performance of many automated systems employed by companies, especially in the early days. As a result, customers may be less willing to engage with an agent if the perceived level of humanization is lower.

Upon customers' follow-up engagement, their biased perceptions can further affect the resolution of customer service interactions. Giffin has shown that the perception of a speaker's expertise facilitates interpersonal trust in the communication process [24] and interpersonal trust plays a critical role in persuasion [25, 26]. Because a critical aspect of customer service is persuading customers to forgive a firm's service failure or defective products, we expect that a more humanized service agent can be more persuasive and therefore is more likely to resolve an issue.

Even if the perceived level of expertise of an algorithmic agent is the same as or even higher than that of a human agent, a customer may nevertheless still prefer to engage with a human agent because of empathy. Empathy, the capacity to understand or feel what others are experiencing from their perspective, is a critical factor that affects the customer service outcome and a uniquely human characteristic that we do not expect from a machine. No matter how well an algorithmic agent can imitate the responses of an empathetic human agent, from a customer's perspective, these responses are merely outputs from a carefully-designed algorithm, sophisticated but emotionless. Therefore, customers are less willing to engage with an algorithmic agent, and even upon engagement, they are less likely to accept any reasoning or apology from an algorithmic agent as opposed to a human agent. We propose the following two hypotheses for empirical tests.

Hypothesis 1: *Humanization increases a customer's willingness to engage.*

Hypothesis 2: *Humanization increases the chance of reaching a resolution.*

3.2. Humanization and Service Process

We focus on customer verbal aggression as a key measure for the process of customer service encounters. To understand how humanization might affect customer verbal aggression, we need to first understand the underlying motives of customer complaining behavior, which can be categorized into *goal-oriented* and *emotion-focused* [27]. Driven by different motivations, customers' attitudes toward agents can be affected differently by humanization.

Goal-oriented customers, after weighing costs and benefits, intend to seek redress or economic compensation through complaining. They may pressure customer service agents through more aggressive behavior to better achieve their goals. Since dehumanized agents, such as chatbots, are merely algorithms lacking empathy, they are unlikely to respond to pressure, at least from the customer's perspective. In contrast, human agents, exactly because of their empathy, are susceptible to emotional pressure. Hence, we conjecture that goal-oriented customers would act more aggressively to customer service agents whom they perceive as more likely to be humans.

The complaining behavior from emotion-focused customers is evoked by frustration and the desire to express emotional dissatisfaction [27]. With such a motivation, customers complain not only because they expect changes to be made but also because the act of complaining itself makes them *feel* better. The aggressiveness of these complaints partly depends on how humanized the recipient of the actions is perceived [28]. In particular, Bandura et al. find that people prefer not to behave cruelly toward those they perceive as more humanized because of human empathy and social norm [29]. Hence, the aggressiveness of emotion-focused complaints can be alleviated by a higher degree of perceived humanization.

Considering that the motivations of customer complaints may be mixed, customers can be either *less* aggressive towards more humanized agents if the emotion-focused motive dominates the goal-oriented motive, or *more* aggressive if the goal-oriented motive dominates the emotion-focused motive. In the latter case, humanization shall have the side effect of hurting customer service employees' performance. As previous literature suggests, customer service agents tend to respond to customers' aggressive behaviors by exhibiting customer service failures in return [1, 2, 30]. Moreover, such a side effect can also increase employee turnover. To examine the overall effect of higher perceived humanization, we propose the following competing hypotheses for empirical examination.

Hypothesis 3a: *Humanization increases a customer's aggressiveness in the interaction.*
Hypothesis 3b: *Humanization decreases a customer's aggressiveness in the interaction.*

4. A Quasi-Experiment

Although we believe most companies provide their customer service on social media through human agents, we are unaware of any that explicitly and clearly states so. Customers are sometimes left wondering whether they are served by human agents or algorithmic agents, especially when agent responses seem monotone or completely out of context. Therefore, we consider *humanization* as a *subjective probability* reflecting customers' belief that they are served by human agents.

Customer service agents from Southwest Airlines on Twitter started including their first names in responses to customer tweets on March 16, 2018. Prior to this, each response is accompanied by a two-letter code following the carat symbol which some interpret as an abbreviation. Clearly, a signature can significantly increase a customer's perception that he or she is dealing with a human agent rather than an algorithmic agent. Figure 1 shows that the percentage of customer service agents who use a signature jumped from 0 to 100% on March 16, 2018. The abruptness of the change and the lack of advanced notice or discussion about the change from Southwest Airlines' official website, Twitter account, and news media, suggests that this is likely an exogenous shock to customers who can neither anticipate nor influence the timing of the change. Therefore, the change offers us a nice quasi-experiment setting to investigate the effect of enhanced humanization perceived by customers.

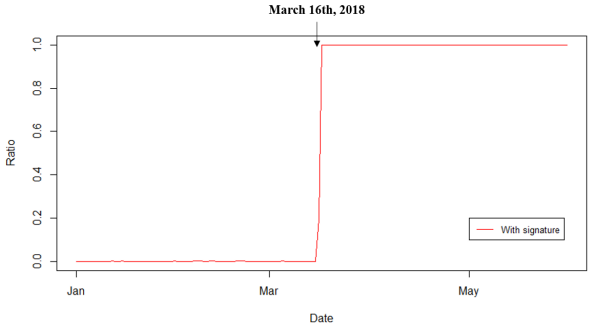

Figure 1. Daily Ratio of Agent Replies Ended With A Signature

4.1. Data

Our data set contains all conversations between an individual customer and Southwest Airlines from February 16, 2018, to April 16, 2018 on Twitter. As our primary focus is social media customer service, we only include conversations that are initiated by customers in our analysis. To distinguish customer service-related conversations from all other types of conversations,[3] we hired an annotator to label 25,530 tweets and then train a support vector machine (SVM) classifier using the labeled data.[4] Next, we use the classifier to identify customer service-related conversations in our data. For ease of illustration, we refer to these customer service-related conversations as *conversations*.

4.2. Variables

We consider three dependent variables in this study. In evaluating the hypotheses regarding the customer service outcome, we first construct the variable $engagement_i$ to measure a customer's willingness to engage with an agent. Conditional on a customer's further engagement, we define $resolution_i$ as a binary variable indicating whether a resolution is reached at the end of the conversation.[5] Since the resolution is difficult to track, an annotator is hired to read through all the conversations and determine whether a resolution is reached. To test the hypothesis regarding the customer service process, we use the Python package *profanity-check* to construct $aggressiveness_i$, which captures customers' attitudes toward agents in customer service encounters.[6]

To alleviate the endogeneity concern regarding the potential shift of customer service engagement over time, we control for a large number of conversational characteristics. On the customer side, we first use the Latent Semantic Analysis (LSA) and the k-means clustering algorithm to group similar tweets into seven clusters.[7] We control for the number of followers, the number of followings, and the number of updates, for every customer. We construct $initialAggressiveness_i$ as the proxy for customers' aggressiveness at the beginning of the conversation,

[3] We define customer service-related conversations as those that start with an inquiry or a complaint. There are several types of conversations that do not fall into the customer service category. For example, customers may initiate a conversation with the airline to participate in a marketing event.

[4] The performance of the SVM classifier is available upon request.

[5] As it is difficult to determine the resolution without customers' further engagement, we focus only on conversations that customers are willing to engage after agents' first reply.

[6] For details of the python package, please see `https://pypi.org/project/profanity-check/`. We also construct an alternative measure for customer aggression based on Google Perspective API (see `https://www.perspectiveapi.com/#/home`), which aims to identify the toxic or abusive language. The results are qualitatively the same and available upon request.

[7] We choose seven because it is the optimal number of clusters suggested by the silhouette score [31]. We also tried alternative numbers of clusters and obtained qualitatively the same results.

which might affect the customer service process and outcomes. We derive customers' Big Five personality traits (i.e., agreeableness, conscientiousness, extraversion, neuroticism, and openness) based on the Linguistic Inquiry and Word Count (LIWC) dictionary and customers' historical tweets on their public Twitter pages. On the agent side, we control for *responseTime*$_i$, *numReplies*$_i$, and *avgWords*$_i$ for each conversation i, which capture the efficiency of agents' interventions. Further, we use a lexicon-based method to create *dm*$_i$, which is a binary variable equal to one if an agent mentions keywords like "direct message" in the reply to request the customer to communicate privately. To capture agents' writing styles, we apply the R package *politeness* to create a list of dummy variables and quantify the quality of agents' responses.[8]

We conduct a balance check of the control variables.[9] There are small and insignificant paired differences for most covariates, suggesting that the comparison of customer service interactions before and after the signature experiment is a good starting point to evaluate the effect of humanization.

5. One-Group Before-and-After Analysis

To test the impact of the enhanced humanization, we start with the standard one-group before-and-after design following previous literature [3, 4] and estimate the following model at conversational level, indexed by i:

$$Y_i = \beta_0 + \beta_1 signature_i + X_i + Z_t + HourofDay_t + DayofWeek_t + TimeTrend_t + \epsilon_{i,t}$$

The main variable of interest is *signature*$_i$ whose coefficient β_1 captures the effect of enhanced humanization. We control conversation-specific characteristics, X_i, which includes the circumstance, customers' characteristics, and agents' service quality. We control time-varying airline characteristics, Z_t, which includes the Google Trend and the number of offline incidents at time t. We include the linear time trend, day-of-week fixed effects, and hour-of-day fixed effects to adjust for any unobserved seasonality.

5.1. Baseline Results

Table 1 reports the regression results with four different estimation windows around the event date. The rationale for using different estimation windows is to alleviate the endogeneity concern of unobserved events during the sample period that can potentially affect the outcome measures. Shortening the estimation window alleviates such a concern, albeit at the expense of statistical power. In columns 1, 4, 7, and 10, the coefficients of *signature*$_i$ are positive and statistically significant ($p < 0.01$), suggesting that including agents' signatures in the reply increases customers' propensity to engage with agents, thereby supporting **Hypothesis 1**. In columns 2, 5, 8, and 11, the coefficients of *signature*$_i$ are positive and significant, although the significance level naturally decreases as the sample size shrinks. Therefore, enhanced humanization increases the likelihood of reaching a resolution, thereby supporting **Hypothesis 2**. In columns 3, 6, 9, and 12, most coefficients of *signature*$_i$ remain insignificant, supporting neither **Hypothesis 3a** nor **3b**. The observed null effect is probably due to the mix of customers' goal-oriented and emotion-focused motives, which can change customers' aggressiveness in opposite directions. Nevertheless, the null effect suggests that the benefits of enhanced humanization at least do not come at the expense of increased cost on customer service agents in the form of elevated customer verbal aggression.

5.2. Robustness Check: Entropy Balancing and Coarsened Exact Matching

Although the aforementioned balance check indicates the comparability of treated and control groups, we further balance the sample to check the robustness of our findings. We use two popular methods. The first one is Entropy Balancing (EB), which relies on a maximum entropy reweighting scheme to produce a more balanced sample [32]. The other method is the coarsened exact matching (CEM), a popular matching method proposed by [33]. The CEM algorithm coarsens the observed variables (i.e., the circumstance, customer characteristics, and agent reply quality) and then applies the exact matching on the coarsened data to determine the matches. Estimation results using both methods are reported in Table 2, which are consistent with the baseline results.

5.3. Falsification Tests: Pseudo Treatments at Different Times Before the Signature Experiment

If the identified effects are largely due to some unobserved performance-improvement initiatives before the event date, then, by assuming a pseudo treatment before the actual policy change, our econometric model would falsely detect similar effects as the baseline analysis. To implement this idea, we assume two pseudo

[8]For details, please see `https://cran.r-project.org/web/packages/politeness/politeness.pdf`

[9]The specific results are available upon request.

treatments, one is on March 1, 2018, which is two weeks before the event date; and the other one is on March 8, 2018, which is one week before the event date. We estimate the regression model with those pseudo treatments and report the results in Table 3. From the insignificant coefficient estimates of $signature_i$ in both tests, we conclude that there is no evidence that our main findings are driven by unobserved events before the signature experiment.

5.4. Falsification Tests: Pseudo Treatments at Different Airlines and at Southwest Airlines in 2017

We consider two types of confounding factors right at or very close around the signature experiment. First, if our findings are driven by unobserved industry-level shocks, then we should be able to falsely detect the humanization effect for other airlines that did not initiate the policy change on March 16, 2018. To implement this idea, we estimate the pseudo treatment effect for other airlines including American Airlines, Delta Airlines, and United Airlines. Columns 1 through 9 of Table 4 report the estimation results. We do not find any significant effect on any of the outcome variables for any of the three airlines. Second, if our findings are driven by unobserved seasonality specific to Southwest Airlines (e.g., annual event by Southwest Airlines around the time of the signature experiment), then we should be able to falsely detect the humanization effect for Southwest Airlines on March 16, 2017. To implement this idea, we conduct a falsification test assuming a pseudo treatment on March 16, 2017, for Southwest Airlines. Columns 10 through 12 of Table 4 report the estimation results. Again, we do not find any significant effects on any of the outcome variables. In summary, these falsification tests suggest that our main findings are unlikely driven by unobserved time-varying confounding factors that are either shared by other airlines or driven by seasonality specific to Southwest Airlines around the time of the signature experiment.

6. Difference-in-Differences Analysis with Synthetic Control

Although our falsification tests alleviate the endogeneity concern due to time-varying confounding factors that are not systematically modeled, it is more common in the literature to employ a difference-in-differences (DID) strategy. However, the validity of the DID approach crucially relies on finding a control airline for which the dependent variable parallels those of Southwest Airlines in the absence of the signature experiment. Unfortunately, such a "parallel trend" assumption is violated for the other three major airlines (see Table 5 for demonstration). An increasingly popular approach to overcome the lack of a natural control is to construct a synthetic control from the so-called donor pool of candidate controls [5, 6, 34, 7].

6.1. Synthetic Control

We utilize the matrix completion method proposed by [7], which uses the observed covariates and outcomes of control units to predict the counterfactual outcomes for the treated unit and period combinations.

To implement the matrix completion method, we consider customer service-related conversations from American Airlines, Delta Airlines, and United Airlines as potential control units. We choose these three airlines as control groups because they have similar passenger volume as Southwest Airlines[10] and did not initiate similar policy change during our sample period. Since the matrix completion method only works on panel data, we aggregate our conversation-level data at a daily level for each airline. For instance, $engagement_{f,t}$ measures the ratio of conversations with customers' further engagement for airline f at day t. As seen in Table 5, there is no significant difference in dependent variables between the treated and the synthetic control before the treatment, suggesting a high quality of the synthetic control. Table 6 reports the estimation results from the matrix completion method, which are consistent with the one-group before-and-after analysis.

6.2. Two-Way Matching

Since the synthetic control is constructed at the aggregate level, conversation-level characteristics such as customer characteristics and agent response quality may not be fully accounted for. Therefore, we propose a two-way matching at the conversation level. Denote the vector of observed covariates and the outcome for a conversation relating to a treated airline before the treatment by X_{t0} and Y_{t0} respectively. For another conversation relating to the treated airline (i.e., Southwest Airlines) but after the treatment, denote the vector of observed covariates and the outcome by X_{t1} and Y_{t1} respectively. In a similar fashion, we use the notations X_{c0}, Y_{c0}, X_{c1}, and Y_{c1} for conversations selected from the donor pool (i.e., all conversations relating to American Airlines, Delta Airlines, or United Airlines).

We illustrate the proposed two-way matching method in Figure 2. First, for each (X_{t0}, Y_{t0}), we use

[10] Please see https://en.wikipedia.org/wiki/List_of_largest_airlines_in_North_America

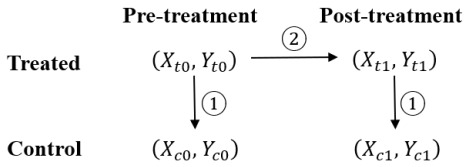

Figure 2. Two-way Matching Procedure

the Mahalanobis distance and one-to-one matching to identify a matched conversation relating to the control airline in the pre-treatment period, (X_{c0}, Y_{c0}), based on X_{c0}. To ensure the matching quality, we keep only the matched pairs whose distance between X_{t0} and X_{c0} is within a specified caliper.[11] The difference in outcomes for the treated group (i.e., Southwest Airlines) and the control group before the signature experiment, $\Delta Y_0 = Y_{t0} - Y_{c0}$, is then calculated. Similarly, by matching conversations in the post-treatment period, we obtain $\Delta Y_1 = Y_{t1} - Y_{c1}$. Second, for each $(X_{t0}, \Delta Y_0)$, we match it with a $(X_{t1}, \Delta Y_1)$ based on the Mahalanobis distance between X_{t0} and X_{t1}. Finally, for each matched tuple $(X_{t0}, X_{t1}, \Delta Y_0, \Delta Y_1)$, we calculate the treatment effect as $\Delta Y_1 - \Delta Y_0$. A t-test is then conducted to test whether the treatment effect is significantly different from zero.

Table 7 reports the results, with two different calipers for the one-to-one matching. Take the first set of results corresponding to the caliper of 3 for example. In Column 1, the positive and significant coefficient of $signature_{i,t}$ suggests that, upon receiving a response, a customer is more willing to engage with a customer service agent perceived as more humanized. In terms of the magnitude of the effect, being served by an agent with a signature increases a customer's likelihood to engage by 5.4 percentage points, which represents a 14% increase in the engagement level at the mean (i.e., from 40% to 45.4%). In Column 2, the positive and significant coefficient of $signature_{i,t}$ suggests that a customer is more likely to reach a resolution with a customer service agent perceived as more humanized. In terms of the magnitude, the estimated coefficient indicates that including a signature increases the probability of reaching a resolution by 10 percentage points, which represents a 19% improvement compared to the previous resolution rate (i.e., from 52% to 62%). On the other hand, the insignificant coefficient of $signature_{i,t}$ in Column 3 again suggests the lack of evidence that enhanced humanization increases (or decreases) customer verbal aggression. In summary, results from this proposed two-way matching analysis reinforce our previous results based on the one-group before-and-after analyses and from the synthetic control analyses.

[11] For instance, if the caliper is equal to three, the matched pairs whose distance is more than three times the standard deviation of all distances among matched pairs will be dropped.

7. Conclusion

The current paper studies the value of humanization in customer service by quantifying how a simple policy change that enhances customers' perception of them being served by human agents affects the outcome and process of customer service interactions. Through a quasi-experiment on Twitter, we find that customers are more willing to engage, and upon engagement, more likely to reach a resolution, with more humanized agents. Furthermore, we find no evidence of increased customer verbal aggression towards more humanized agents, despite the theoretical prediction that this could happen. These findings are robust in a series of robustness checks, falsification tests, and two sets of synthetic control analyses.

This paper contributes to the IS and marketing literature by being the first quantitative study on the value of humanization in customer service, thereby advancing the existing literature that qualitatively discusses the relationship between automated engagement and human agents [20, 21]. Our paper also provides two important insights to practitioners. First, despite the popular trend of automating customer service through AI technology, our findings offer a cautionary tale that there is a natural limit to the effectiveness of algorithmic agents in customer service, no matter how advanced the technology is. Firms should always keep in mind that AI-augmented customer service or human-assisted customer service is likely the long-run equilibrium for customer service in the age of AI. Second, the specific empirical setting of our paper suggests that a simple policy change of requiring agent signatures in responses to customer inquires can go a long way towards more engaging and more satisfying interactions. Moreover, such a policy change does not appear to have the unintended consequence of increasing customer verbal aggression. Given the easiness and negligible cost of such a policy change, we encourage all firms to do so when they deliver customer service on social media.

References

[1] A. A. Grandey, D. N. Dickter, and H.-P. Sin, "The customer is not always right: customer aggression and emotion regulation of service employees," *Journal of Organizational Behavior*, vol. 25, pp. 397–418, May 2004.

[2] A. Rafaeli, A. Erez, S. Ravid, R. Derfler-Rozin, D. E. Treister, and R. Scheyer, "When customers exhibit verbal aggression, employees pay cognitive costs.," *The Journal*

of Applied Psychology, vol. 97, no. 5, pp. 931–950, 2012.

[3] X. M. Zhang and F. Zhu, "Group size and incentives to contribute: A natural experiment at Chinese Wikipedia," *American Economic Review*, vol. 101, no. 4, pp. 1601–15, 2011.

[4] J. Claussen, T. Kretschmer, and P. Mayrhofer, "The effects of rewarding user engagement: The case of Facebook apps," *Information Systems Research*, vol. 24, no. 1, pp. 186–200, 2013.

[5] A. Abadie and J. Gardeazabal, "The economic costs of conflict: A case study of the Basque Country," *American Economic Review*, vol. 93, no. 1, pp. 113–132, 2003.

[6] A. Abadie, A. Diamond, and J. Hainmueller, "Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program," *Journal of the American Statistical Association*, vol. 105, no. 490, pp. 493–505, 2010.

[7] S. Athey, M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi, "Matrix completion methods for causal panel data models," tech. rep., National Bureau of Economic Research, 2018.

[8] L. Ma, B. Sun, and S. Kekre, "The squeaky wheel gets the grease—An empirical analysis of customer voice and firm intervention on Twitter," *Marketing Science*, vol. 34, pp. 627–645, Sept. 2015.

[9] J. S. Gans, A. Goldfarb, and M. Lederman, "Exit, tweets and loyalty," Working Paper 23046, National Bureau of Economic Research, Jan. 2017.

[10] S. He, S.-Y. Lee, and H. Rui, "Open voice or private message? The hidden tug-of-war on social media customer service," in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[11] P. Gunarathne, H. Rui, and A. Seidmann, "Whose and what social media complaints have happier resolutions? Evidence from Twitter," *Journal of Management Information Systems*, vol. 34, pp. 314–340, Apr. 2017.

[12] P. Gunarathne, H. Rui, and A. Seidmann, "When social media delivers customer service: Differential customer treatment in the airline industry," *MIS Quarterly*, vol. 42, pp. 489–A10, June 2018.

[13] Y. Hu, A. Tafti, and D. Gal, "Read this, please? The role of politeness in customer service engagement on social media," in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[14] L. Cui, S. Huang, F. Wei, C. Tan, C. Duan, and M. Zhou, "SuperAgent: A customer service chatbot for e-commerce websites," in *Proceedings of ACL 2017, System Demonstrations*, pp. 97–102, July 2017.

[15] A. Xu, Z. Liu, Y. Guo, V. Sinha, and R. Akkiraju, "A new chatbot for customer service on social media," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 3506–3510, 2017.

[16] T. Hu, A. Xu, Z. Liu, Q. You, Y. Guo, V. Sinha, J. Luo, and R. Akkiraju, "Touch your heart: A tone-aware chatbot for customer care on social media," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, pp. 1–12, 2018.

[17] K. Corti and A. Gillespie, "Co-constructing intersubjectivity with artificial conversational agents: People are more likely to initiate repairs of misunderstandings with agents represented as human," *Computers in Human Behavior*, vol. 58, pp. 431–442, May 2016.

[18] Y. Mou and K. Xu, "The media inequality: Comparing the initial human-human and human-AI social interactions," *Computers in Human Behavior*, vol. 72, pp. 432–440, July 2017.

[19] S. Ba, J. Stallaert, and Z. Zhang, "Balancing IT with the human touch: Optimal investment in IT-based customer service," *Information Systems Research*, vol. 21, pp. 423–442, Sept. 2010.

[20] M.-H. Huang and R. T. Rust, "Artificial intelligence in service," *Journal of Service Research*, vol. 21, pp. 155–172, May 2018.

[21] L. Xiao and V. Kumar, "Robotics for customer service: A useful complement or an ultimate substitute?," *Journal of Service Research*, p. 1094670519878881, Sept. 2019.

[22] X. Luo, S. Tong, Z. Fang, and Z. Qu, "Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases," *Marketing Science*, vol. 38, no. 6, pp. 937–947, 2019.

[23] C. Longoni, A. Bonezzi, and C. K. Morewedge, "Resistance to medical artificial intelligence," *Journal of Consumer Research*, vol. 46, pp. 629–650, Dec. 2019.

[24] K. Giffin, "The contribution of studies of source credibility to a theory of interpersonal trust in the communication process.," *Psychological Bulletin*, vol. 68, no. 2, pp. 104–120, 1967.

[25] C. I. Hovland and W. Weiss, "The influence of source credibility on communication effectiveness," *Public Opinion Quarterly*, vol. 15, no. 4, pp. 635–650, 1951.

[26] B. Sternthal, L. W. Phillips, and R. Dholakia, "The persuasive effect of scarce credibility: a situational analysis," *Public Opinion Quarterly*, vol. 42, no. 3, pp. 285–314, 1978.

[27] R. M. Kowalski, "Complaints and complaining: Functions, antecedents, and consequences.," *Psychological Bulletin*, vol. 119, no. 2, p. 179, 1996.

[28] A. Bandura, "Social learning theory of aggression," *Journal of Communication*, vol. 28, pp. 12–29, Sept. 1978.

[29] A. Bandura, B. Underwood, and M. E. Fromson, "Disinhibition of aggression through diffusion of responsibility and dehumanization of victims," *Journal of Research in Personality*, vol. 9, no. 4, pp. 253–269, 1975.

[30] D. D. Walker, D. D. van Jaarsveld, and D. P. Skarlicki, "Sticks and stones can break my bones but words can also hurt me: The relationship between customer verbal aggression and employee incivility.," *Journal of Applied Psychology*, vol. 102, no. 2, pp. 163–179, 2017.

[31] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[32] J. Hainmueller, "Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies," *Political Analysis*, vol. 20, no. 1, pp. 25–46, 2012.

[33] S. M. Iacus, G. King, and G. Porro, "Causal inference without balance checking: Coarsened exact matching," *Political Analysis*, vol. 20, no. 1, pp. 1–24, 2012.

[34] Y. Xu, "Generalized synthetic control method: Causal inference with interactive fixed effects models," *Political Analysis*, vol. 25, no. 1, pp. 57–76, 2017.

Table 1. Baseline Results

	$\pm$ 1 *month*			$\pm$ 3 *weeks*		
	engagement$_i$ (1)	*resolution*$_i$ (2)	*aggressiveness*$_i$ (3)	*engagement*$_i$ (4)	*resolution*$_i$ (5)	*aggressiveness*$_i$ (6)
signature$_i$	0.0842***	0.0695**	-0.0093*	0.1177***	0.0822*	-0.0083
	(0.0216)	(0.0352)	(0.0050)	(0.0254)	(0.0425)	(0.0058)
Controls	Y	Y	Y	Y	Y	Y
Time Trend	Y	Y	Y	Y	Y	Y
Seasonality FE	Y	Y	Y	Y	Y	Y
Observations	8214	3258	3258	5771	2249	2249
R^2	0.06	0.14	0.07	0.07	0.15	0.09
	$\pm$ 2 *weeks*			$\pm$ 1 *week*		
	engagement$_i$ (7)	*resolution*$_i$ (8)	*aggressiveness*$_i$ (9)	*engagement*$_i$ (10)	*resolution*$_i$ (11)	*aggressiveness*$_i$ (12)
signature$_i$	0.2256***	0.2082***	0.0074	0.2262***	0.2810**	-0.0092
	(0.0378)	(0.0686)	(0.0082)	(0.0632)	(0.1071)	(0.0123)
Controls	Y	Y	Y	Y	Y	Y
Time Trend	Y	Y	Y	Y	Y	Y
Seasonality FE	Y	Y	Y	Y	Y	Y
Observations	3885	1518	1518	2010	744	744
R^2	0.08	0.16	0.10	0.10	0.19	0.12

Notes. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Robust standard errors in parentheses.

Table 2. Robustness Check: Entropy Balancing and Coarsened Exact Matching

	EB			CEM		
	engagement$_i$ (1)	*resolution*$_i$ (2)	*aggressiveness*$_i$ (3)	*engagement*$_i$ (4)	*resolution*$_i$ (5)	*aggressiveness*$_i$ (6)
signature$_i$	0.0620***	0.0670*	-0.0086	0.0791***	0.0853**	-0.0077
	(0.0229)	(0.0380)	(0.0057)	(0.0249)	(0.0420)	(0.0051)
Controls	Y	Y	Y	Y	Y	Y
Time Trend	Y	Y	Y	Y	Y	Y
Seasonality FE	Y	Y	Y	Y	Y	Y
Observations	8214	3258	3258	7733	3110	3110
R^2	0.01	0.02	0.01	0.02	0.07	0.06

Notes. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Robust standard errors in parentheses.

Table 3. Falsification Test with Pseudo Treatment at Different Times Before the Signature Experiment

	2018-3-1			2018-3-8		
	engagement$_i$ (1)	*resolution*$_i$ (2)	*aggressiveness*$_i$ (3)	*engagement*$_i$ (4)	*resolution*$_i$ (5)	*aggressiveness*$_i$ (6)
signature$_i$	0.0501	0.0358	-0.0026	0.2056	-0.1895	0.0972
	(0.0407)	(0.0661)	(0.0126)	(0.1975)	(0.3636)	(0.0645)
Controls	Y	Y	Y	Y	Y	Y
Time Trend	Y	Y	Y	Y	Y	Y
Seasonality FE	Y	Y	Y	Y	Y	Y
Observations	3513	1315	1315	1649	574	574
R^2	0.09	0.14	0.10	0.12	0.21	0.20

Notes. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Robust standard errors in parentheses.

Table 4. Falsification Test with Pseudo Treatment at Different Airlines and at Southwest Airlines in the Previous Year

	American Airlines - 2018-3-16			Delta Airlines - 2018-3-16		
	$engagement_i$	$resolution_i$	$aggressiveness_i$	$engagement_i$	$resolution_i$	$aggressiveness_i$
	(1)	(2)	(3)	(4)	(5)	(6)
$signature_i$	0.0255	0.0340	0.0032	0.0065	-0.0101	-0.0034
	(0.0191)	(0.0304)	(0.0072)	(0.0076)	(0.0313)	(0.0049)
Controls	Y	Y	Y	Y	Y	Y
Time Trend	Y	Y	Y	Y	Y	Y
Seasonality FE	Y	Y	Y	Y	Y	Y
Observations	10609	4138	4138	13267	3462	3462
R^2	0.05	0.14	0.08	0.07	0.18	0.05
	United Airlines - 2018-3-16			Southwest Airlines - 2017-3-16		
	$engagement_i$	$resolution_i$	$aggressiveness_i$	$engagement_i$	$resolution_i$	$aggressiveness_i$
	(7)	(8)	(9)	(10)	(11)	(12)
$signature_i$	0.0274	-0.0687	0.0067	-0.0085	-0.0076	0.0014
	(0.0229)	(0.0425)	(0.0084)	(0.0213)	(0.0352)	(0.0059)
Controls	Y	Y	Y	Y	Y	Y
Time Trend	Y	Y	Y	Y	Y	Y
Seasonality FE	Y	Y	Y	Y	Y	Y
Observations	5656	2017	2017	8013	2936	2936
R^2	0.13	0.12	0.08	0.08	0.16	0.07

Notes. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Robust standard errors in parentheses.

Table 5. Paired t-tests of Differences in the Pre-treatment Period (in %)

	$engagement_{f,t}$	$resolution_{f,t}$	$offensiveness_{f,t}$
Treated - American Airlines	-0.511	1.269	-1.367***
Treated - Delta Airlines	10.912***	4.413***	0.123
Treated - United Airlines	4.288***	5.741***	-0.095
Treated - Synthetic Control	-0.040	-0.005	0.040

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

Table 6. Synthetic Control

	$engagement_{f,t}$	$resolution_{f,t}$	$aggressiveness_{f,t}$
	(1)	(2)	(3)
$signature_{f,t}$	0.022***	0.010***	0.0012
	(0.005)	(0.004)	(0.0019)
Controls	Y	Y	Y
FirmFE	Y	Y	Y
TimeFE	Y	Y	Y
Observations	240	240	240

Notes. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Bootstrapped standard errors in parentheses.

Table 7. Conversational-level Matching

	$caliper = 3$			$caliper = 4$		
	$engagement_{i,t}$	$resolution_{i,t}$	$aggressiveness_{i,t}$	$engagement_{i,t}$	$resolution_{i,t}$	$aggressiveness_{i,t}$
	(1)	(2)	(3)	(4)	(5)	(6)
$signature_{i,t}$	0.0541***	0.1002**	-0.0114	0.0436***	0.0759**	-0.0081
	(0.0186)	(0.0467)	(0.0086)	(0.0164)	(0.0358)	(0.0065)
Observations	2,608	409	409	3,349	711	711

Notes. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Robust standard errors in parentheses.