When a Voice Assistant Asks for Feedback: An Empirical Study on Customer Experience with A/B Testing and Causal Inference Methods

YUQI DENG
Amazon

SUDEEKSHA MURARI
Amazon

Intelligent Voice Assistant (IVA) systems, such as Alexa, Google Assistant and Siri, allow us to interact with them using just the voice commands. IVA systems can seek voice feedback directly from the customers, right after an interaction by simply asking a question such as "did that answer your question?". We refer to these IVA elicited feedbacks as crowdsourced voice feedback (CVF). In this paper, we look to understand the customer experience (CX) during interactions with an IVA that explicitly seeks feedback. We attempt to quantify the CX of providing feedback, identify the driving factors of CX and offer insights into improving CX with the drivers identified. With an A/B test, we collected data from a leading IVA system and found that feedback elicitations did not impair CX in general. To identify drivers of CX, we performed causal inference with Double Machine Learning. Causal inference teases apart multiple confounding factors and avoids CX risks in experimentation of certain variables. We identified multiple CX drivers including elicitation timing and frequency, which can be useful in establishing guardrails for a CVF system. Our results imply opportunities of CVF systems, and we suggest design specifics that can be leveraged for such feedback collection mechanisms.

CCS CONCEPTS • Human computer Interaction • Human-centered computing

Additional Keywords and Phrases: Voice assistant, Feedback, Post-confirmation question, Customer experience, A/B testing, User modeling, Causal inference, Double Machine Learning

ACM Reference Format:

Yuqi Deng, Sudeeksha Murari. 2021. When a Voice Assistant Asks for Feedback: An Empirical Study on Customer Experience with A/B Testing and Causal Inference Methods. In *Workshop Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21), Oct 18–22, 2021, Virtual event, Montreal, Canada.* ACM, New York, NY, USA, 10 pages.

1 Introduction

In order to collect customer feedback on specific aspects of an Intelligent Voice Assistant (IVA), a system such as Crowdsourced Voice Feedback (CVF) can be valuable. In an IVA system, CVF can target specific interactions based on the intention and category of a request, and it can target a particular device type and several other characteristics that define an IVA interaction. Typically, when a customer makes a request to an IVA, the IVA responds with an action or with a response. The customer is then asked a question about the interaction to gauge if the IVA rendered a satisfactory response/action (Figure 1). These questions can be generic or very specific to the interaction. For example, following a

user request such as ‘what is zero degree Celsius in Fahrenheit’, the IVA provides a response, a CVF takes over and asks the user “did that answer your question?” or “on a scale of 1-5 how satisfactory was the response?”. Typical responses are ‘yes’, ‘no’, ‘maybe’, silence or some numeric response. When annoyed, customers also say things such as ‘shut up’ or ‘go away’ or an angry ‘stop’. These responses could be used to improve the IVA.

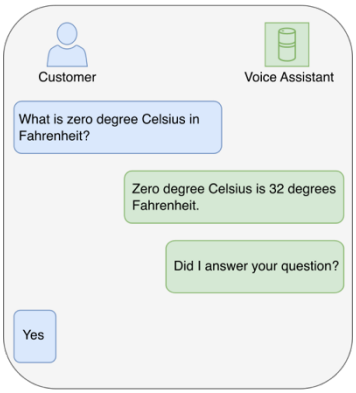

Figure 1: An example dialog of feedback elicitation by an IVA

CVF data is a valuable source of customer feedback, and could be used for evaluation and training of the IVA model. Compared to training and testing data generated by conventional manual annotation, CVF data is less expensive, more instantaneous and customer-centric. However, there is little information available on measuring the customer experience (CX) during customer interactions with such a proactive feedback seeking IVA system. Measuring the CX is important to make sure that feedback elicitations are not annoying customers and driving them away from using the IVA. In addition, CX analysis helps identify customer pain points and facilitates designs of strategies to provide better experience to customers. Such strategies like not asking at a certain time of the day, not asking more than a certain number question per week etc. are called CX guardrails.

In this paper, we aim to answer two questions regarding the CX of CVF systems. First, we investigate in whether or not proactive feedback seeking affects CX in terms of engagement and friction. Traditionally, causal relationship discovery is performed with a randomized experiment approach (A/B testing). Typically in A/B testing, customers in control and treatment group receive identical treatments except for one variation of interest [1]. A/B testing has become a standard way for companies to evaluate user experience for new product features [2], [3]. In this paper, we present results of a controlled A/B testing experiment comparing CX metrics of customers receiving feedback elicitations and those not receiving feedback elicitations.

Second, we aim to explore factors that may positively or negatively drive CX during CVF interactions. In this paper, we selected one key CX metric Feedback Response Rate (rate of customers giving a valid response) and explored some factors that influence it. We adopted causal inference analysis for this study. Causal inference analysis intersecting with machine learning modelling is an alternative method to estimate Conditional Average Treatment Effect (CATE) [4]. Instead of collecting experimental data, causal inference analysis leverages observational data that is not randomized, and machine learning models could be applied to control for the bias introduced by not randomizing. Machine learning based causal inference analysis provides several advantages over a traditional approach of controlled A/B testing which faces several challenges. First, there are multiple confounding factors that may potentially impact CX and bias the treatment outcome. To avoid bias in selection, an A/B experiment needs to be set up with extra caution to include all possible confounders [11], which could be challenging with larger number of confounders. Second, A/B testing on certain variables risks impairing CX. For example, we would like to understand whether repetitive feedback elicitations affect CX. However, eliciting a certain group of customers at a high frequency may negatively impact their experience. Third, it can be expensive to run A/B testing experiments at scale. And finally, A/B testing takes time, during which several confounders may have changed due to changes to components that make up IVAs. To the best of our knowledge, this is a novel application of a machine learning method in the context of IVAs to study the causal factors driving CX.

Therefore, for the second study we applied a causal inference method called Double Machine Learning (DML) to tease apart a large number of factors that may potentially affect CX during CVF. DML is a recently developed causal inference method for estimating heterogenous treatment effect [5]–[7]. A major advantage of DML is that it could be applied to both categorical and continuous treatment variables. On top of that, DML works with high-dimensional and non-parametric confounding factors which could not be resolved by traditional quasi-experimental analysis [8]–[10].

2 Related works

Collecting customer feedback data for self-learning and reinforcement learning by dialog agents is a growing new area of research. Previous studies have shown that using implicit [12], [13] and explicit customer feedback [14] as training data improves model performance of dialog agents. These pioneer works demonstrate attractive potentials and opportunities of CVF systems.

However, there has been limited literature on how interacting with customers through explicit feedback seeking affects CX. A previous psychology study on human-human-interactions has found that asking follow-up questions during social interactions increases interpersonal liking [15]. Whether similar beneficial effects apply to human-computer-interactions is unknown. A less similar previous study has shown that proactive question-asking by an IVA is perceived positively by users and does not negatively affect CX [16]. However, proactively asking questions in their context means task-oriented suggestions rather than post-confirmation feedback seeking, which is a major difference from our study.

Moreover, very few studies have explored how different parameters or factors of a post-confirmation question may influence CX. Using causal analysis methods in user experience studies for business decision making and feature development has been spurring research interest in recent years. Although causal inferencing has been successfully applied to improve CX in several customer centric industries [17], [18], to our knowledge there has been no previous applications in the field of IVAs and dialog agents.

3 Study 1: overall impacts of feedback elicitations

3.1 Methods

3.1.1 A/B Testing

In order to better understand how customers would interact with CVF systems we conducted an A/B testing experiment with data from a subset of consented consumers. Conducting A/B experiment is a common practice in industry to help make product decisions and study customer preferences. This approach enabled us to compare a treatment group of customers who did not receive feedback elicitations and a control group of customers who had a chance to receive feedback elicitations at a relatively low rate (<1%). After a period of testing time, we compared the impact metrics of the control and treatment group with statistical tests to assert customer impact. We computed the following impact metrics for the control and treatment group respectively:

- Number of dialogs. The average number of dialogs a customer had during the testing period. Formula: total number of dialogs of all customers in each group / total number of customers in the group.
- Number of active days. The average number of days for each customer, where a customer had at least 1 interaction. Formula: sum of active days of all customers during the testing period / total number of customers in the group.
- Next day retention. Whether a customer who used the IVA on a given day comes back to use it again the next day. Formula: next day retention of all customers in each group / total number of customers in the group.
- Average satisfactory score. Satisfactory score was calculated on each feedback utterance using the sentiment analysis method proposed by Kim et al. [19] Formula: total satisfaction score of all interactions in each group / number of interactions in each group.

- Unprompted Negative Feedback Rate. This metric measures the number of customers' unprompted negative utterances such as "that was terrible". Formula: total number of unprompted negative feedback utterances/ total number of utterances.
- We collected data from a random subset of fifty thousand customers who were sharing their usage reports. Customers were randomly assigned to be in the control or treatment group. To analyze short-, medium- and long- term impacts we collected data from 3 weeks, 6 weeks and 8 weeks period of time respectively. The duration of data collection was selected to be an integer multiplier of number of weeks to avoid bias of results by day of the week.

3.1.2 Statistics

For each of the impact metrics, we performed a two-sample, independent t-test on the control and treatment group to determine the differences in means of the two groups. We also reported the power [20] and effect size for comparisons that showed statistically significant p-values ($p<0.05$).

3.2 Results

3.2.1 Feedback elicitations do not impair CX in general

To decide whether there is a short-, medium- and long-term impacts of feedback elicitations, we monitored and compared the impact metrics for the control and treatment group over a period of 3, 6 and 8 weeks (Table 1). A positive difference percentage indicates a larger value in treatment group (customers not receiving feedback elicitations), while a negative difference indicates otherwise. The results showed that during these periods of A/B testing time, control group and treatment group did not differ in any of the engagement and friction metrics, except for that the Average Satisfactory Score is slightly lower in treatment group where customers were not receiving feedback elicitations ($p=0.038$). However, the effect size was very low at -0.02%, and such an effect size did not reach 80% of power in our power analysis. None of the impact metrics differed between the control and treatment group with statistical significance ($p < 0.05$; power >80%).

Table 1. A/B test statistical results (p-values) comparing treatment and control group

		Short-term	Med-term	Long-term
Number of dialogs	P-value	0.85	0.75	0.76
	Diff%	-0.06	-0.09	-0.09
Number of active days	P-value	0.43	0.37	0.29
	Diff%	-0.11	-0.13	-0.15
Next day retention	P-value	0.39	0.39	0.26
	Diff%	-0.16	-0.12	-0.2
Unprompted negative feedback rate	P-value	0.73	0.50	0.58
	Diff%	-0.31	-0.08	-0.38
Average satisfactory score	P-value	**0.038**	0.11	0.068
	Diff%	**-0.02**	-0.01	-0.01

4 STUDY 2: FACTORS DRIVING FEEDBACK RESPONSE RATE

4.1.1 Causal Inference

The aim of this study was to understand how parameters of a feedback elicitation may impact CX and customers' willingness to respond. We were interested in the following parameters of a feedback question: elicitation frequency, question type, timing of elicitation (local day, local hour), and we would like to control for confounding factors including original intent, original domain, device type and customer engagement. The definitions of these factors are listed in Table

2. We aimed to estimate the CATE of these parameters and identify any factors that positively or negatively drive response rate.

We used a quasi-experiment [21] approach to estimate these impacts. Specifically, instead of setting up A/B testing experiments, we collected observational data from production data log of a leading IVA system and performed causal inferencing on the observational data. The methodological challenges in this study were: 1) certain factors may confound the likelihood of a question getting a valid response or not. For example, a customer who frequently interact with the IVA may behave differently than a customer that rarely engages with the IVA. A certain device type or a certain type of interaction (e.g., asking for weather vs playing music) may also complicate the results. Therefore, this analysis required controlling for multiple confounding factors to de-bias the final effect estimation; 2) the impacts of such confounding factors may be non-parametric or large in dimension.

The DML method we adopted is comprised of three models. The first model predicted the outcome from control factors and calculate the prediction residual; The second model predicted treatment from control factors and calculate the prediction residual; The final model predicted the residual of model 1 from residual of model 2. Together, DML offers a robust meta-model with great flexibility, where any machine learning algorithms could be used these three models [4]. We implemented the DML model with a recently released Python package EconML [22].

The target variable we were interested in is Feedback Response Rate. Feedback Response Rate is defined as:

Feedback Response Rate = number of valid responses / total number of feedback elicitations.

Valid responses included answers that were yes/no/maybe/numeric. Non-responses included answers that were silent/stop/irrelevant. Feedback Response Rate not only directly measures the efficacy of the feedback elicitations but also reflects customers' willingness to participate in feedback giving. Therefore, Feedback Response Rate is a key metric to measure CX, and it is correlated with multiple CX metrics (Figure 2). Feedback Response Rate is positively correlated with satisfactory scores of feedback utterances (Pearson's $r=0.44$, $p=0.037$). Sentiment score was calculated on each feedback utterance using the method proposed by Kim et al. [19]. Feedback Response Rate is negatively correlated with Feedback Interruption Rate (Pearson's $r=0.94$, $p<0.001$), which is defined as the percentage of interruption (customer expressing any interruption intents such as "stop", "enough") that occurred after the feedback question TTS (text to speech) was rendered by the IVA. We aggregated the data by intent, domain and question type (Table 2) and plotted the results in Figure 2.

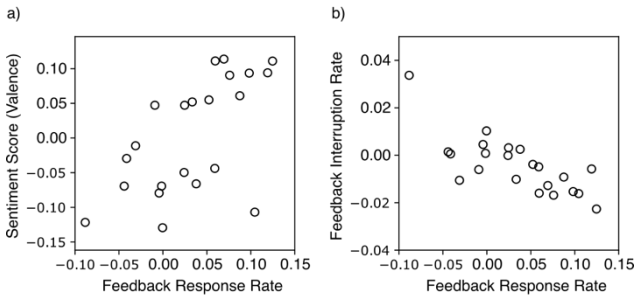

Figure 2. Scatter plots comparing associations among Feedback Response Rate vs. Sentiment and Feedback Response Rate vs. Feedback Interruption Rate. n= 23. Data aggregated by intent, domain and question type. a) Higher Feedback Response Rate is associated with higher Sentiment score. b) Higher Feedback Response Rate associated with lower Feedback Interruption Rate.

4.1.2 Observational Data

To explore the pattern of user interactions with a leading IVA system, we sampled one month of de-identified data from production data log of the IVA system. Our data source was composed of a randomized collection of utterances from customers who have received feedback questions from the IVA. We collected the following data fields for each utterance (Table 2).

Table 2. Data fields collected for each utterance

Data field	Definition
Elicitation Frequency	Number of feedback questions received by a customer per week.
Question Type	Binary if the question seeks a binary answer (e.g., Did I answer your question?); Numeric if the question seeks a numeric answer (e.g., How am I doing on a scale of 1 to 5 stars?).
Local Day	The day of the week in local time when feedback question is asked.
Local Hour	The hour of the day in local time when feedback question is asked.
Customer Response	Customer response to the feedback question. Valid responses include answers that are yes/no/maybe/numeric. Non-responses include answers that are silence/stop/irrelevant. This is classified by a Natural Language Understanding (NLU) model
Original Intent	The customer intent of the experience feedback is elicited on, classified by an NLU model. E.g., Turn off light; what time is it.
Original Domain	The class of CX feedback is elicited on, classified by an NLU model. E.g., Weather, Time, News etc.
Device Type	The type of IVA device on which the interaction occurred.
Customer engagement	The number of dialogs a customer had with the IVA per week.

We adopted a weighted sampling strategy to balance the number of utterances in each condition. Since the majority of customers received none or very small number of feedback elicitations from the IVA, the number of customers who received a small number of elicitations were down-sampled to reduce data size for modeling as well as to ensure each condition has similar number of samples. This allowed the modeling results to have similar statistical power in each elicitation condition. The final dataset constituted of customers who received different number of feedback elicitations per week from the IVA (Table 3).

When a customer opted out of the data collection, their data were removed from the database. We excluded such data in our experiment. In addition, data from customers with extremely high frequency of interactions (top 0.2%) were not included. We screened out customers with extremely frequent interactions to avoid the inclusion of possible developer customers' testing accounts. The final dataset after pre-processing consisted of over 34k utterances.

4.1.3 Double Machine Learning Model

DML is comprised of two stages of models. The first stage consists of two residualization models: a) modelling the outcome Y from the confounders (X, W) and calculate the residual Y_res; b) modeling the treatment T from the confounders (X, W) and calculate the residual T_res. X and W are both confounder features. X are the set of features we would like to model the treatment effect against, usually important features we hypothesized to have significant impact on treatment effects. Whereas W represents all the other variables of less interest that may potentially have impacts on the treatment effect. The residualization model equations are listed in equation 1 and equation 2.

$$Y\_res = Y - E[Y|X,W] \quad (1)$$

$$T\_res = T - E[T|X,W] \quad (2)$$

The final stage model predicts the residual of measured outcome (Y_res) from the residual of treatment variables (T_res) as a function of X (equation 3).

$$Y\_res = \theta(X) \cdot T\_res + \varepsilon \quad (3)$$

A major advantage of DML is that any machine learning models could be applied as residualization models. One of the key reasons we chose this model is that it can be applied to a versatile variety of variables of binary, discrete and continuous treatment variables. Our outcome measurement Y is a binary variable following a weighted Bernoulli distribution. Therefore, we selected a classifier to be our first residualization model. The treatment variable could be either a continuous variable or binary. In case T is continuous, we selected a regressor for the second residualization model; otherwise, we selected a classifier for the second residualization model.

4.2 Results

4.2.1 Causal effects of elicitation frequency

We aimed to investigate in the causal effects of repetitive elicitations on response rate. The probability of customers to provide a valid response reflects customers' willingness to participate in feedback giving, therefore is an important metric to measure CX as well as program effectiveness. We would like to know what is the optimal frequency of eliciting a customer for feedback. However, it was difficult to obtain an accurate estimation of the causal impacts of elicitation frequency due to multiple confounding factors.

From observed data, we found that customer's response rate changed with elicitation frequency in a non-linear manner, and the impact is further confounded by customer's level of engagement (Figure 3a). We separated customers to 3 engagement bins: low engagement (lowest 28% of all customers), medium engagement (lowest 26% of all customers) and high engagement (highest 46% of all customers). Low engagement customers' Feedback Response Rate increases with elicitation frequency until 3 elicitations/week. For medium engagement customers, the elbow point is 4 elicitations/week. Whereas for high engagement customers, there was no elbow point within 5 elicitations/week. It is important to note that the response rates presented here (Figure 3a) were baseline corrected by subtracting a baseline number from each datapoint. The baseline number is the response rate of 1 elicitation per week.

Table 4: Definitions of variables

Y	Outcome measurement, whether the customer provided a valid response. Y ~ Bernoulli (p=Feedback Response Rate calculated from observed data)
T	Treatment. Discrete variable of elicitation frequency (Table 2).
X	Confounder of interest. Continuous variable of customer engagement (dialogs/week, Table 2)
W	Other confounders. Categorical variables of original intent, original domain, question type, device type, local hour and local day (Table 2).

We applied DML to model the impacts of repetitive elicitation on probability of response. We applied polynomial feature transformation (degree =3) to treatment variable T to model the possible non-linear treatment effect (Table 4). The categorical features in W were encoded with one-hot encoding. To model the nonlinear treatment effects we found in observational data (Figure 3a), we used a Generalized Additive Model (GAM) [23] with a polynomial feature transformation (degree = 3) to model the non-linear treatment effects.

$$Y\_res = \theta 1(X) \cdot T\_res + \theta 2(X) \cdot T2\_res + \theta 3(X) \cdot T3\_res + \varepsilon \quad (4)$$

We applied the following models. The heavy regularization in the residualization models was applied to ensure the stability of CATE estimates [4].

Residualization models:

- To model Y_res, we applied a weighted logistic regression classifier with 5-fold cross-validation and L2 regulatory parameter C=0.1.
- To model T_res, we applied a shallow random forest regressor with max depth of 3, 1000 trees, and minimum split sample size of 500.

Final model:

- Nonlinear GAM model with polynomial features (degree = 3). Since we observed that parameters of the nonlinear CATE changes with customer engagement (in Figure 3a, low engagement customers have a lower elbow point), we introduced further non-linearity in the modeling of treatment parameter function θ(X) (equation 5) We applied polynomial feature transformation of degree =3.

$$\theta(X) = \beta 0+\beta 1X+\beta 2X2+\beta 3X3+\varepsilon \qquad (5)$$

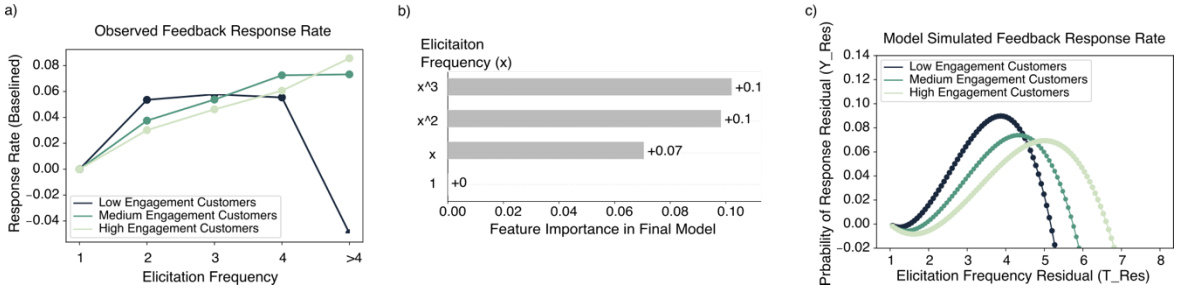

Figure 3. Effects of elicitation frequency (number of elicitations per week) on customer response. a) Observed data. We calculated Feedback Response Rates of customers who received different number of elicitations per week. Customers were separated into three engagement bins: low, medium and high engagement. b) Feature importance in final model. The numbers represent absolute mean feature impacts on CATE. c) Model results of treatment effects. We plotted the treatment effect of repetitive elicitations (T_res) on the probability of receiving a valid answer from customers (Y_res). Since the treatment effect is influenced by customers' engagement level, we plotted the model results for the three customer engagement bins respectively.

The output of our model is visualized in Figure 3c. After accounting for the confounding factors, we observed that repetitive elicitations had a positive impact on driving customer response up to a certain elbow frequency, regardless of customer engagement level. However, the elbow frequency increases with customer engagement level. Low engagement customers' response rate starts to drop as early as 3 elicitations/week, whereas high engagement customers' response rate does not drop until over 5 elicitations/week. In Figure 3c we showed the baseline corrected Y_res by subtracting the residual value at elicitation frequency of 1 from each data point for each engagement bin. Figure 3b showed that the largest impact of elicitation frequency comes from the third-degree polynomial transformed feature. For the third degree of T3_res (see equation (4)), coefficients in θ3(X) had p-values of ($P\beta 0<0.001$; $P\beta 1<0.001$; $P\beta 2=0.0013$; $P\beta 3=0.028$). For the second degree of T2_res (see equation (4)), coefficients in θ2(X) had p-values of ($P\beta 0<0.001$; $P\beta 1<0.001$; $P\beta 2=0.032$; $P\beta 3=0.14$). For the first degree of T_res (see equation (4)), coefficients in θ1(X) (equation (5)) had p-values of ($P\beta 0<0.001$; $P\beta 1<0.001$; $P\beta 2=0.41$; $P\beta 3=0.39$).

4.2.2 Causal effects of question types

We studied the impacts of question type on customer response rate. In our observational datasets, there are two types of questions: binary and numeric. Binary questions seek a yes/no answer, while numeric questions seek a numeric rating (Table 2).

We observed that in our dataset, binary questions had a higher Feedback Response Rate than numeric questions (Figure 4a), and the difference in Feedback Response Rate between the two question types decreased as elicitation frequency increases. Feedback Response Rate was baseline corrected by subtracting the binary condition of 1 elicitation per week and divided by it. Multiple factors could confound this observation. For example, although there was a general upward trend in Feedback Response Rate with higher elicitation frequency, we could not determine a positive causal effect, since customers with higher engagement were more likely to get more elicitations per week and complicate the results. To tease apart the confounding factors and estimate the CATE of eliciting a numeric question vs a binary question, we applied a similar DML approach to model CATE. The variables we included in this model are listed in Table 5. The categorical features in W were encoded with one-hot encoding.

Table 5: Definitions of variables

Y	Outcome measurement, whether the customer provided a valid response. Y ~ Bernoulli (p=Feedback Response Rate calculated from observed data)
T	Treatment. Discrete variable of elicitation question type (whether the question is binary or numeric, Table 2).
X	Confounders of interest. 1) Continuous variable of customer engagement (dialogs/week, Table 2). 2) Continuous variable of customer engagement (dialogs/week, Table 2).
W	Other confounders. Categorical variables of original intent, original domain, device type, local hour and local day (Table 2).

Now that our treatment variable is binary, we selected a classifier model as a residualization model. Both residualization models were heavily regularized to ensure the stability of CATE estimates [4].
Residualization models:

- To model Y_res, we applied a weighted random forest classifier with max depth equals to 3 and 1000 trees.
- To model T_res, we applied a shallow random forest regressor with max depth of 3, 1000 trees, and minimum split sample size of 1000.

Final model:

- Linear model. $\theta(X)=\beta 0+\beta 1X+\varepsilon$ (6)

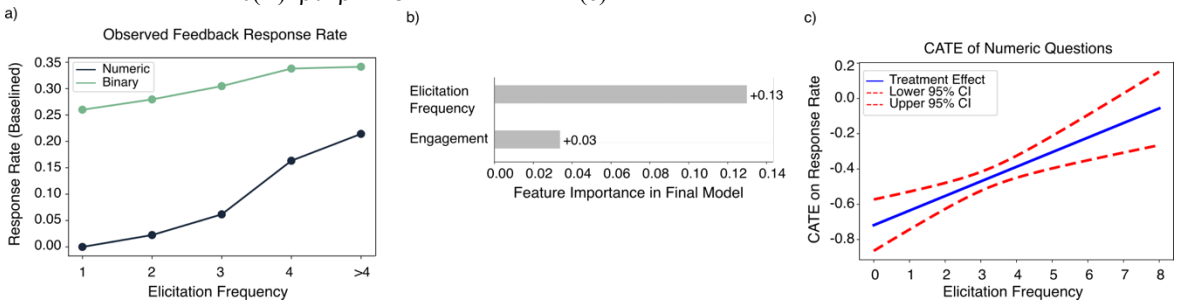

Figure 4. Treatment effect of numeric questions on response probability. a) Observed data. We calculated Feedback Response Rates of customers who received different number of elicitations per week. Data was separated by elicitation question type. b) Feature importance in final model. The numbers represent absolute mean feature impacts on CATE. c) Linear projection of treatment effect on elicitation frequency. The treatment effects (y axis represents treatment effects) of having numeric questions (compared to asking binary questions) are mostly negative, but this impact becomes less obvious with more elicitation repetitions (slope of projection is positive).

Figure 4b is a variable importance plot, that shows that elicitation frequency has the largest impact on model output ($p<0.001$) while engagement level does not impact model outcome ($p=0.075$). We projected the impact of elicitation frequency on the treatment effect (Figure 4c) and found that after controlling for the confounding factors: 1) the overall treatment effect of having a numeric question is negative (the y axis values in Figure 4c are negative), meaning that eliciting a numeric question receives less response than a binary question. 2) this impact decreases with elicitation frequency. On Figure 4c which projects treatment effects on elicitation frequency, the slope of projection is positive. This result suggests that customers found it less intuitive to answer rating questions than simple binary questions. However, repetitive elicitations mitigate this negative effect by helping customers to learn to answer such questions.

4.2.3 Causal effects of elicitation Timing – elicit on weekend vs weekday

We further investigated into the treatment effect of eliciting feedback on weekend vs weekdays. The method we used is similar to 4.2.2, except that the treatment independent variable is now a binary indicator of local weekday (0 for weekday, 1 for weekend). In our observed data (Figure 5a), we found that in general eliciting on the weekends result in

lower Feedback Response Rate than eliciting during weekdays. However, the difference is very small and could potentially be influenced by factors such as original domain and intent. For example, it is possible that certain intents (movie recommendation or play music) occur more frequently on weekends than weekdays. Model output rather than observational data could be a more accurate estimation of CATE. The variables we included in this model are listed in Table 6. The categorical features in W were encoded with one-hot encoding.

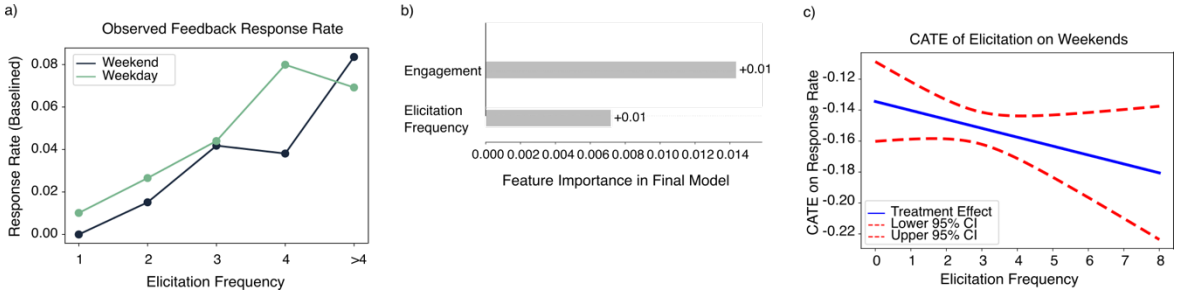

Figure 5. Treatment effect of weekend elicitations on response probability. a) Observed data. We calculated Feedback Response Rates of customers who received different number of elicitations per the week. Data was separated by elicitation timing of the week. b) Feature importance in final model. The numbers represent absolute mean feature impacts on CATE. c) Linear projection of treatment effect on elicitation frequency. The treatment effects of eliciting feedback on weekends (compared to weekdays) are mostly negative (y axis represents treatment effects), and this impact becomes more obvious with more elicitation repetitions (slope of projection is negative).

Figure 5b showed that elicitation frequency has less impact than customer engagement in this case. However, we found that the impact of elicitation frequency is statistically significant despite being small ($p<0.001$), whereas that of engagement is not ($p=0.10$). We plotted the impact of elicitation frequency on treatment effect in Figure 5c. It is shown that the overall impact of eliciting on weekends has a negative effect on customer response rate (the y axis values in Figure 5c are negative). Moreover, this negative effect gets worse when customers are repetitively elicited (projection of treatment effects on elicitation frequency shows a negative slope, Figure 5c). This result suggests that eliciting for feedback on weekend may impair CX and willingness to participate, and the negative impact is driven to be worse when elicited frequently.

4.2.4 Causal effects of elicitation Timing – elicit at night vs daytime

We further investigated the treatment effect of eliciting feedback on daytime vs night. The method we used is similar to 4.2.2, except that the treatment independent variable is now a binary indicator of local weekday (0 for daytime, 1 for night). In our observed data (Figure 6a), we found that in general eliciting at daytime and at night have similar Feedback Response Rate, and the effect of repetitive elicitations is not obvious. Again, we modelled CATE with a DML model. The variables we included in this model are listed in Table 7. The categorical features in W were encoded with one-hot encoding.

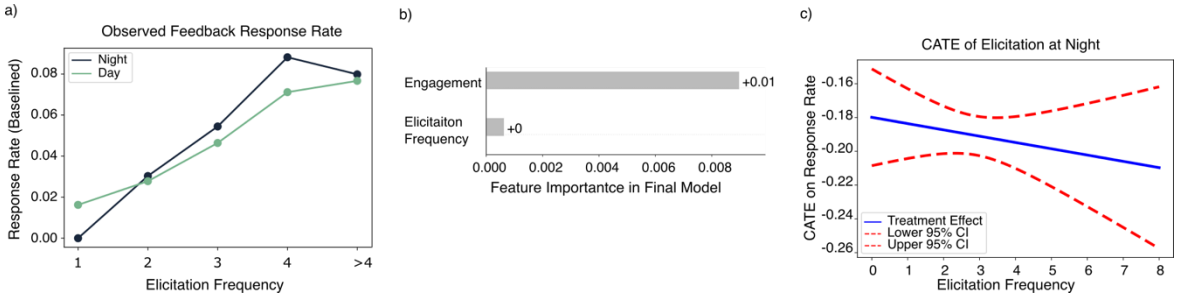

Figure 6. Treatment effect of night time elicitations on response probability. a) Observed data. We calculated Feedback Response Rates of customers who received different number of elicitations per week. Data was

separated by elicitation timing of a day. b) Feature importance in final model. The numbers represent absolute mean feature impacts on CATE. c) Linear projection of treatment effect on elicitation frequency. The treatment effects of eliciting feedback at night (compared to eliciting in daytime) is mostly negative (y axis represents treatment effects), and this impact becomes more obvious with more elicitation repetitions (slope of projection is negative).

Figure 6b showed that elicitation frequency has very little impact in this case. However, we found that the impact of elicitation frequency is statistically significant despite being small ($p=0.031$), whereas that of engagement is not ($p=0.52$). Figure 6c showed that the overall impact of eliciting customers at night vs at day time is negative (the y axis values in Figure 6c are negative), resulting in less valid responses. This negative impact is worsened with repetitive elicitation (projection of treatment effects on elicitation frequency shows a negative slope, Figure 6c). This result suggests that eliciting for feedback at night may impair CX, and the negative impact is worsened with repetitive elicitation.

5 Discussion

The aim of this work is to better understand CX during interactions with an IVA that explicitly seeks feedback. We have presented the results from two studies on this topic. First, we performed an A/B testing experiment which showed that at a relatively low rate (<1%) asking feedback questions does not have significant impact on customer engagement and friction metrics in 3-week, 6-week and 8-week periods of time. Second, we conducted causal inference on observational data to identify factors that may potentially drive Feedback Response Rate, which is a crucial metric for CX and willingness to participate. It was found that Feedback Response Rate is non-linearly impacted by elicitation frequency and customer engagement. We also found that question type and elicitation timing affect the Feedback Response Rate, and the magnitude of impacts changes with elicitation frequency. To our knowledge, our work is the first to provide insights on CX during feedback seeking of an IVA.

Our results suggests that feedback-seeking by an IVA does not cause customer annoyance with proper design parameters. Customer feedback are a valuable resource for the development and growth of products. It helps the understanding of customer pain-points and is essential for designing customer-centric features. The voice user interface (VUI) of an IVA provides a novel channel for customer feedback collection. However, many open questions remain to be answered in how to leverage feedback elicitation to efficiently collect customer feedback without annoying customers with too many or too frequent questions. A previous study has showed users tend to be more satisfied with an interactive dialog agent that asks follow-up questions than a passive one [24]. Similarly in human-human interaction, asking follow-up questions is also perceived to be a favorable behavior [15]. Together with our results, these findings suggest that there are opportunities in VUI as a channel for feedback collection.

We further studied driving factors of customer response rate in a more granulated approach. We found that repetitive elicitation affects customer response rate non-linearly (Figure 3), and an optimal elicitation frequency could be computed to aid elicitation VUI design. Moderately increasing elicitation frequency actually improves customer response rate (Figure 3). We speculate that moderate amount of repetition could help customers get familiar with answering the question. This was consistent with another finding that although numeric questions receive lower responses than binary, repetitive elicitation helps (Figure 4). It is interesting to note that the optimal elicitation frequency depends on customer engagement level, where lower engagement customers have lower tolerance for repetitive elicitations (Figure 3). This results suggests that personalization of the interaction could improve the efficiency and CX [25], [26] during feedback elicitation. Our results also suggests that the timing of elicitation is important, eliciting on weekends and at night hurts CX, which could be worsened even more with repetitive elicitations. In summary, our results present important implications on how and when to ask for feedback to maximize the response rate as well as protect CX. We could further utilize the findings to build alarm metrics and guardrail systems to ensure CX is not degraded by unfavorable feedback elicitations.

Beyond the factors of feedback elicitation that we investigated in this paper, there are many more factors in the interaction context that could potentially affect CX. For example, future studies could integrate tools in natural language processing (NLP) to investigate in metrics such as the length of feedback questions, naturalness of speech [27], formality of speech [28] and so on to investigate this matter on an even more granular level.

Recent years have seen more and more pioneer works applying machine learning methods to the field of causal inference. Compared to traditional A/B testing and quasi-experiment methods, ML based causal inference methods has great versatility to deal with a variety of data types, and enables experimenters to work with observational data with high dimensions of controlling factors. Multiple algorithms and tools have been developed for this purpose such as EconML

[22], causalML [11], DoWhy [29] and so on. In this study we take advantage of both A/B testing and ML based causal inference methods to study the user experience. We chose DML because compared to meta-learners [30] DML could work with both continuous and binary treatment data, making it a more suitable choice for our case. Future meta-analysis study reviewing and comparing the different causal inference machine learning methods and the different tool packages would be helpful for identifying appropriate algorithms and tools for different purposes.

REFERENCES

<BIBL>

[1] C. Sammut, G. Webb, R. Kohavi, and R. Longbotham, "in the Encyclopedia of Machine Learning and Data Mining, edited by Online Controlled Experiments and A/B Tests."

[2] Y. Xu, N. Chen, A. Fernandez, O. Sinno, and A. Bhasin, "From infrastructure to culture: A/B testing challenges in large scale social networks," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015, vol. 2015-August, pp. 2227–2236.

[3] I. Bilogrevic et al., "'Shhh...be quiet!' Reducing the Unwanted Interruptions of Notification Permission Prompts on Chrome."

[4] V. Syrgkanis, V. Lei, M. Oprescu, M. Hei, K. Battocchi, and G. Lewis, "Machine learning estimation of heterogeneous treatment effects with instruments," arXiv, 2019.

[5] V. Chernozhukov, D. Chetverikov, M. Demirer, and E. Duflo, "Double/debiased machine learning for treatment and structural parameters," 2018.

[6] V. Chernozhukov et al., "Double/Debiased Machine Learning for Treatment and Causal Parameters," Jul. 2016.

[7] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey, "Double/Debiased/Neyman Machine Learning of Treatment Effects," 2017.

[8] N. Kumar, L. Qiu, and S. Kumar, "Exit, Voice, and Response on Digital Platforms: An Empirical Investigation of Online Management Response Strategies," Inf. Syst. Res., vol. 29, no. 4, pp. 849–870, 2018.

[9] H. Cavusoglu, T. Q. Phan, H. Cavusoglu, E. M. Airoldi, and N. Jindal, "Assessing the Impact of Granular Privacy Controls on Content Sharing and Disclosure on Facebook," Inf. Syst. Res., vol. 27, no. 4, pp. 848–879, 2016.

[10] M. Yang, Y. Ren, and G. Adomavicius, "Engagement by Design," ACM Trans. Comput. Interact., vol. 27, no. 6, Nov. 2020.

[11] H. Chen, T. Harinen, J.-Y. Lee, M. Yung, and Z. Zhao, "CausalML: Python Package for Causal Machine Learning."

[12] T. Falke, M. Boese, D. Sorokin, C. Tirkaz, and P. Lehnen, "Leveraging User Paraphrasing Behavior In Dialog Systems To Automatically Collect Annotations For Long-Tail Utterances," pp. 21–32, 2021.

[13] D. Muralidharan et al., "Leveraging User Engagement Signals For Entity Labeling in a Virtual Assistant," arXiv, Sep. 2019.

[14] B. Hancock, A. Bordes, P. E. Mazaré, and J. Weston, "Learning from dialogue after deployment: Feed yourself, chatbot!," arXiv, pp. 3667–3684, 2019.

[15] K. Huang et al., "It Doesn't Hurt to Ask: Question-Asking Increases Liking," J. Pers. Soc. Psychol., vol. 113, no. 3, pp. 430–452, 2017.

[16] M. Schmidt, W. Minker, and S. Werner, "How users react to proactive voice assistant behavior while driving," Lr. 2020 - 12th Int. Conf. Lang. Resour. Eval. Conf. Proc., no. May, pp. 485–490, 2020.

[17] T. Harinen and B. Li, "Using Causal Inference to Improve the Uber User Experience," 2019.

[18] F. Tan, Z. Wei, A. Pani, and Z. Yan, "User response driven content understanding with causal inference," in Proceedings - IEEE International Conference on Data Mining, ICDM, 2019, vol. 2019-November, pp. 1324–1329.

[19] Y. Kim, J. Levy, and Y. Liu, "Speech Sentiment and Customer Satisfaction Estimation in Socialbot Conversations," Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, vol. 2020-October, pp. 1833–1837, Aug. 2020.

[20] C. R. Wilson Vanvoorhis and B. L. Morgan, "Understanding Power and Rules of Thumb for Determining Sample Sizes," 2007.

[21] J. D. Angrist and J.-S. Pischke, Mostly Harmless Econometrics : An Empiricist ' s Companion Mostly Harmless Econometrics : An Empiricist ' s Companion. Princeton University Press, 2009.

[22] "EconML/Double Machine Learning Examples.ipynb at master · microsoft/EconML · GitHub." [Online]. Available: https://github.com/microsoft/EconML/blob/master/notebooks/Double Machine Learning Examples.ipynb. [Accessed: 03-Mar-2021].

[23] T. Hastie and R. Tibshirani, Generalized additive models. 1990.

[24] S. Quarteroni and S. Manandhar, "Designing an interactive open-domain question answering system," 2009.

[25] Y. Yuan et al., "Speech interface reformulations and voice assistant personification preferences of children and parents," Int. J. Child-Computer Interact., vol. 21, pp. 77–88, Sep. 2019.

[26] C. E. Rhee and J. Choi, "Effects of personalization and social role in voice shopping: An experimental study on product recommendation by a conversational voice agent," Comput. Human Behav., vol. 109, p. 106359, Aug. 2020.

[27] A. Hindle, E. T. Barr, M. Gabel, Z. Su, and P. Devanbu, "On the naturalness of software," Commun. ACM, vol. 59, no. 5, pp. 122–131, 2016.

[28] S. Rao and J. Tetreault, "Dear sir or madam, may i introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer," in NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 2018, vol. 1, pp. 129–140.

[29] A. Sharma and E. Kiciman, "DoWhy: An End-to-End Library for Causal Inference," arXiv, Nov. 2020.

[30] S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu, "Metalearners for estimating heterogeneous treatment effects using machine learning," Proc. Natl. Acad. Sci. U. S. A., vol. 116, no. 10, pp. 4156–4165, Mar. 2019.

</BIBL>