Does Bot Gender Matter? Theory and Evidence from a Tense Service Context

Yiting Guo, Ximing Yin, De Liu, Sean Xin Xu

Forthcoming in MIS Quarterly

Abstract

Despite the increasing use of AI-powered voicebots, our understanding of how the choice of bot gender may impact service outcomes in tense service contexts, such as debt collection, remains limited. To address this gap, we draw upon the tensions-based view of customer relationships and gender stereotype theory to hypothesize how and when voicebot gender matters in tense service contexts. We test our hypotheses using a proprietary dataset of debt-collection calls made by AI voicebots. We find that female voicebots increase the odds of a positive repayment intention by 28.3%. This gender effect is more pronounced when service encounters begin with higher tension, such as during weekdays or with initially uncooperative customers. We further show that the gender effect can be explained by the advantages of female voicebots in reducing behavioral and emotional tensions during service interactions.

Keywords: *artificial intelligence; human–computer interaction (HCI); voicebot; tensions-based view; gender stereotype; tense service contexts*

Introduction

Tense customer service contexts, such as debt collection and service failure recovery, are "critical moments of truth" that significantly impact customer relationships and business performance (Heidenreich et al., 2015; Oliveira & Lumineau, 2019). Such tense service settings are characterized by heightened customer emotions, strained relationships, and potential conflicts

(Heidenreich et al., 2015; Zheng et al., 2022), which often induce high turnover among human service representatives. Consequently, companies are increasingly exploring AI-powered voicebots to handle tense customer service interactions (Phillips & Moggridge, 2019). With the global chatbot market valued at USD 5.1 billion in 2023 (Pune, 2024), business leaders, AI bot developers, and service managers increasingly face bot design issues that could directly impact service performance and influence future AI investment decisions.

In this paper, we focus on an important design dimension: the *gender* of AI voicebots. We are first interested in *whether* female voicebots have genuine performance advantages in tense service contexts. The human–computer interaction (HCI) literature has previously studied the effect of bot gender, primarily in friendly or neutral settings, such as virtual partners in healthcare (Borau et al., 2021), gaming assistance (Forgas-Coll et al., 2022), and sales recommendations (Benbasat et al., 2020) (see Appendix A for a review). However, the prior literature documents inconclusive evidence for the bot gender effect on customer perceptions and service performance measured by purchase intention (Beldad et al., 2016; Benbasat et al., 2020). Another notable void in the literature is that bot gender effects remain largely unexplored in tense service settings. Given prior studies, it remains unclear whether the female gender of voicebots has no impact on service performance overall, or whether, in certain contexts—such as tense service situations—female voicebots exhibit performance advantages.

It is crucial to address this question because, in practice, companies frequently default to using female voices and names in bot applications like Cortana and Siri (Cambre & Kulkarni, 2019). There have been suggestions that the preponderance of female bots is rooted in gender bias, with AI products being “designed by men who hold a stereotypical view of females” (Oliver, 2020). Given the research gaps identified above, it remains uncertain whether this design choice leads to tangible performance improvements or merely perpetuates a gender myth in AI. This is particularly critical in high-stakes, tense service contexts, where ineffective management of such scenarios can have significant consequences, including customer churn and payment defaults (Hill, 1994; Kotler & Keller, 2008).

Furthermore, existing studies are conducted in laboratory settings and lack evidence on the real-world impact of bot gender. Lab experiments may not be ideal for studying tense service settings because of inherent difficulties in recruiting participants with experiences in tense service interactions and inducing authentic customer reactions in simulated situations. Existing laboratory studies primarily focus on how bot gender affects user attitudes and perceptions (e.g., likability, attractiveness) toward the *bot* (Borau et al., 2021; Crowell et al., 2009; Lee et al., 2000), rather than on downstream business outcomes such as service performance. There has been a shortage of research on the impact of robot gender on real-world service outcomes. This research aims to fill the gaps by focusing on real-world tense service outcomes, thereby providing more credible and direct support for business decisions in tense service contexts.

Our second goal is to elucidate the underlying mechanisms, i.e., *how* voicebot gender affects tense service outcomes. Although a few studies have examined gender effects through factors such as perceived humanness and competence (Ahn et al., 2022; Borau et al., 2021; Pitardi et al., 2023), they offer limited insights into the mechanisms by which voicebot gender may affect tense service outcomes. It is imperative to delve into the mechanisms specific to tense contexts, as uncovering such mechanisms would lend further credibility to the existence of gender effects in these settings rather than attributing the role of gender to a mere statistical artifact.

Last, it is crucial to comprehend *when* gender effects are more salient. From a theoretical perspective, answers to this question shed light on the boundary conditions that govern the impact of bot gender. From a practical standpoint, understanding such factors enables the selection of personalized bot genders, thereby optimizing voicebot service performance.

In sum, our research aims to address the following three questions:

Whether — Do female voicebots have real-world performance advantages over their male counterparts in tense service contexts?

How — What are the mechanisms through which voicebot gender could affect service performance?

When — When does the effect of voicebot gender become more pronounced?

We address the above questions in the context of AI voicebots for debt collection, a representative tense service setting characterized by stressful, unpleasant, and even threatening interactions with customers (Phillips & Moggridge, 2019). The debt-collection industry is economically significant and growing fast, with a projected total value of USD 43.69 billion by 2027 (Industryresearch, 2022). We collaborate with a prominent Chinese AI voicebot provider to obtain a unique dataset of debt-collection calls made by AI voicebots. We focus on the effect of voicebots' gender on customers' repayment intention—i.e., whether the customer promised or agreed to repay their overdue debt during the debt-collection calls. Repayment intention is the most important performance metric in the debt-collection industry (CX Today, 2021). Prior research shows that repayment intention is a strong predictor of actual repayment (Mazar et al., 2018) and is commonly used as a proxy for actual payment (Li et al., 2024).

Theory and Hypotheses

We draw upon the *tensions-based view (TBV)* of customer relationships (Fang et al., 2011; Zheng et al., 2022) and *gender stereotype theory* (Prentice & Carranza, 2002) to develop hypotheses regarding the effect of bot gender, the underlying mechanisms, and moderating factors in tense service settings, as exemplified by debt-collection calls (see Figure 1 for a summary of the hypothesized relationships).

Gender Effects Under Tense Service Contexts

Originally developed for business-to-business (B2B) relationships, the TBV explains the types of tensions in a relationship and highlights their effects on relationship outcomes (Fang et al., 2011; Gnyawali et al., 2016). According to the TBV, *tension* refers to discomfort, dissonance, stress, strain, or conflict that arises from conflicting interests between actors, contradictory goals, or ambiguity in interactions (Tóth et al., 2018; Zheng et al., 2022). The TBV identifies three types of tension. *Behavioral tension* refers to the tension exhibited through actors' divergent actions, routines, and communication practices, such as a contrast between collaborative and avoiding approaches (Fang et al., 2011; Tóth et al., 2018). *Emotional tension*, also known as psychological tension, refers to the tension exhibited through negative attitudes, perceptions, and emotions formed about other parties (Tóth et al., 2018; Zheng et al., 2022). *Structural tension* refers to the challenges and conflicts arising from how relationships are organized and governed in a network (Fang et al., 2011). The TBV literature has shown that, if not addressed, such tensions in a relationship can result in ailing relationship quality, partnership breakdowns, and poor business performance (Gnyawali et al., 2016; Oliveira & Lumineau, 2019).

Although the TBV is primarily developed and tested in B2B settings, recent research has begun to apply it to business-to-customer (B2C) settings, such as customer–brand relationships (Alvarez et al., 2021) and business-and-customer co-creation (Tsiotsou & Diehl, 2022). Similar to B2B relationships, different types of tension may also arise in B2C interactions, and managing these tensions is vital for effective communication and achieving desirable service outcomes

(Alvarez et al., 2021; Tsiotsou & Diehl, 2022). In the following, we argue that an AI voicebot's gender may influence tensions during service interactions and, consequently, the service outcomes.

To explain the effects of bot gender on tensions in service encounters, we draw on the *gender stereotype theory*, which posits that individuals unconsciously attribute specific traits and behaviors to others based on gender. Specifically, this theory suggests that traits associated with females typically include communality, friendliness, kindness, warmth, and empathy, whereas those associated with males include agency, aggression, assertiveness, dominance, and forcefulness (Bernotat et al., 2021; Fossa & Sucameli, 2022; Smith et al., 2016). Prior research shows that humans project gender stereotypes onto bots based on gender cues such as voice, name, and facial features (Ahn et al., 2022; Eyssel & Hegel, 2012). These projections significantly influence the dynamics between customers and bots, thereby shaping interactions and outcomes (Bernotat et al., 2021; Eyssel & Hegel, 2012).

Applying the TBV and gender stereotype theory to our research context, we first note that the customer–bot relationship in a collection call is inherently tense due to conflicting interests and power imbalance, with customers typically in a low-power position (Deville, 2015). In such tense interactions, the gender stereotypes that customers associate with the voicebot could influence their reactions. A male voicebot would invoke male stereotypes of dominance, aggression, and high power, which may amplify the perceived power imbalance between bot collectors and customers and heighten existing tensions. By contrast, a female voicebot might invoke perceptions of warmth and kindness, which can soothe stressed customers and foster their cooperation, ultimately resulting in better collection outcomes (Harrington, 2018; Hill, 1994). Therefore, when customers encounter a female voicebot, they may experience less tension and be more inclined to cooperate, which increases their repayment intention. Accordingly:

H1: *Female voicebots lead to higher repayment intention than male voicebots do.*

Mechanisms for Gender Effects in Tense Service Settings

In tense service settings such as debt collection, we propose that bot gender may influence service outcomes by impacting these two types of tension identified by TBV: behavioral tension and emotional tension. The TBV literature also identifies structural tension as being related to how relationships are organized and governed (Fang et al., 2011). We do not investigate structural tension in this study because the relationship structure in debt collection is predetermined and is relatively impervious to factors such as voicebot gender. Moreover, we focus on *customers'* behavioral and emotional tensions given that bot collectors' behaviors and emotions are largely predetermined by design.

Mediating Role of Behavioral Tension. In tense service contexts, *behavioral tension* may manifest as avoidance, interruption, or active resistance when the parties interact with each other (Custers, 2017). In debt collection, such behavioral tension naturally arises from the inherent conflict of interest between customers and collectors, with the former interested in minimizing the amount they have to pay and avoiding harassment from collectors, and the collectors aiming to maximize the payment by engaging in active outreach and persuasion.

We argue that bot collectors' voice gender could impact behavioral tension during debt-collection calls. When customers associate bots with masculine traits of agency, assertiveness, and aggression, they may respond more defensively and uncooperatively, exhibiting behaviors such as call avoidance, frequent interruption, antagonistic responses, and premature call termination (Prentice & Carranza, 2002). By contrast, when customers associate bots with feminine stereotypes such as friendliness and communality, they may correspondingly become more prosocial and have a more relational focus, which increases their engagement and reduces their confrontational behaviors during communication (Siegel et al., 2009). Indeed, prior research has shown that female service bots are more effective at maintaining customer dialog and encouraging information sharing (Borau et al., 2021; Pitardi et al., 2023).

Reducing behavioral tension is crucial in tense service settings such as debt collection because active avoidance or antagonistic responses could disrupt communication or render it less effective. These can further lead to failed negotiations, deteriorated relationships, and ultimately poor service performance (Gnyawali et al., 2016; Oliveira & Lumineau, 2019). Therefore, we propose:

H2a: *Behavioral tension mediates the relationship between voicebot gender and customer repayment intention.*

Mediating Role of Emotional Tension. Tense service interactions also lead to customers experiencing emotional tension, which manifests as anxiety, anger, frustration, and guilt, accompanied by negative attitudes such as distrust or animosity (Alvarez et al., 2021; Tóth et al., 2018; Vejnovic et al., 2024). In debt collection, debtors' emotional tension particularly stems from financial distress, the perceived power imbalance between the bot collector and the customer, and embarrassment over being overdue (Custers, 2017; Deville, 2015).

We argue that bot gender can also influence tense service outcomes by affecting emotional tension. Perceived female stereotypes of warmth and empathy are particularly valuable in alleviating the psychological strain inherent in debt collection (Ahn et al., 2022; Nass & Brave, 2005). Such perceptions can help customers feel emotionally supported and trusted (Bernotat et al., 2021; Cheng et al., 2022), thereby reducing their negative emotional reactions (Gelbrich, 2010). By contrast, the perceived dominance and forcefulness of male collectors may make customers feel pressured or judged, exacerbating their stress and negative emotions.

The reduction of emotional tension can significantly improve tense service outcomes. Prior research has demonstrated that high emotional tension significantly impairs customers' ability to process information rationally and make sound decisions (Forgas-Coll et al., 2022). When emotional tension is lower, customers can shift from an emotionally reactive state to a problem-solving mindset (Karimi et al., 2018). The latter enables customers to better comprehend and evaluate information during interactions, which is crucial for positive service outcomes. Therefore, we propose:

H2b: *Emotional tension mediates the relationship between voicebot gender and customer repayment intention.*

Moderation by Contextual Factors

Our tension-based explanation of bot gender effects suggests that these effects could be affected by various contextual factors that contribute to the initial level of tension in a service encounter, resulting in what we refer to as "tense situations." In general, if a contextual factor leads to heightened initial tension, the advantages of female voicebots in de-escalating tension and inducing more cooperative service outcomes will be amplified; thus, we can expect a more pronounced female advantage. Such contextual factors may emerge from circumstantial differences inherent in a service encounter or from individual differences specific to the customer. In the following, we identify three such contextual factors: two factors related to the service encounter circumstance—*stressful time* and *gender incongruence*—and one factor related to customer differences—their *initial uncooperativeness*.[1] We next discuss how these factors

[1] We did not include bot-specific factors because our research questions call for holding bot design constant except for bot gender. We did, however, examine whether other bot design differences (such as pitch and speech rate) may offer an alternative explanation to our findings (see Robustness Checks).

may moderate the gender effects.

Stressful Time. When debt collectors contact customers during a *stressful time*, the overall tension in the interaction is likely to be high from the outset (Stokoe & Sikveland, 2020). For example, customers may experience greater stress when receiving debt-collection calls on weekdays versus weekends, as they are often managing heavier workloads and experiencing greater mental and physical exhaustion (Binnewies et al., 2010). In these high-pressure scenarios, the perceived feminine traits of warmth and empathy can ease the pressure that customers feel, reducing overall tension in the interaction. Consequently, the positive effect of female voicebots on customer repayment intentions may be amplified.

Gender Incongruence. Prior studies show that people feel more comfortable with and psychologically closer to individuals of the same gender (Benbasat et al., 2020; Eyssel & Hegel, 2012). Based on this, we posit that gender incongruence between the voicebot and the customer may lead to higher initial tension. Specifically, when a female voicebot interacts with a male customer, relative to a female customer, the gender incongruence may result in lower initial comfort and greater perceived psychological distance, resulting in elevated initial tension. Consequently, the advantage of female bots could be amplified in opposite-gender interactions.

Initial Uncooperativeness. Customers who display initial uncooperativeness, manifesting as unfriendliness, hostility, or low trust, indicate a higher level of initial tension in the interaction (Kostopoulos et al., 2014). This uncooperative stance creates a challenging communication environment. In such tense situations, female voicebots' advantages become particularly pronounced. Female voicebots, possessing perceived traits such as kindness, may be better equipped to establish rapport, defuse tension, and encourage cooperation from initially uncooperative customers (Eyssel & Hegel, 2012).

In summary, these tense situations suggest heightened initial tension, which can magnify the benefits of female voicebots in alleviating emotional and behavioral tensions. This, in turn, may result in a more pronounced gender effect on customer repayment intention. Therefore, we hypothesize:

H3. *The effect of voicebot gender on customer repayment intention and on (behavioral and emotional) tensions is stronger when debt-collection calls occur at a stressful time (H3a), with gender incongruence (H3b), and with initially uncooperative customers (H3c).*

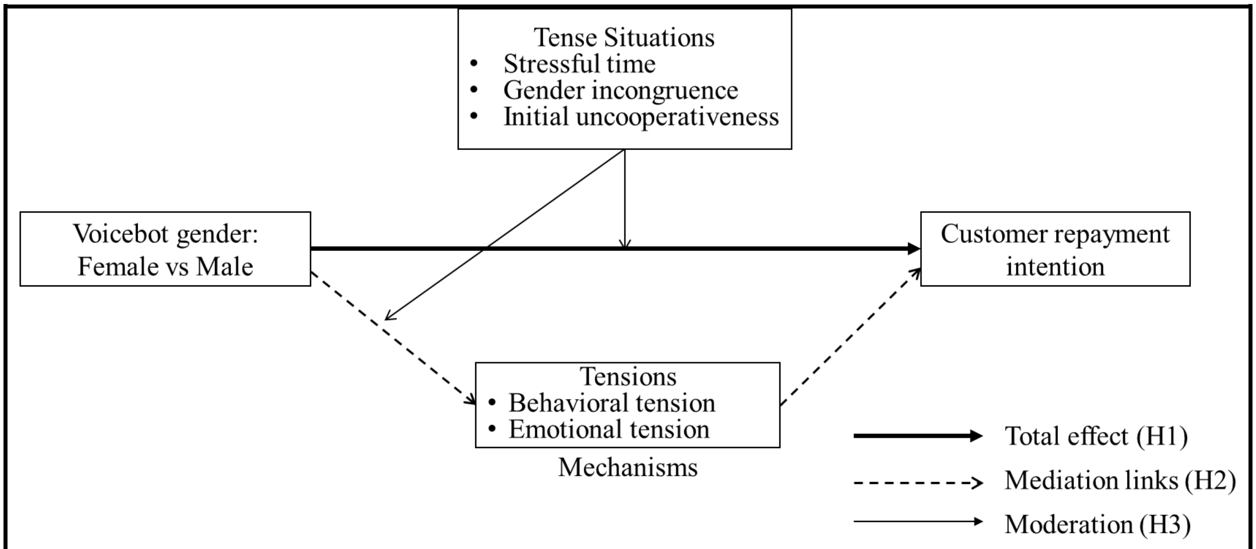

Figure 1. Conceptual Model

Methodology

Data

We leveraged a proprietary dataset comprising outbound debt-collection calls made by AI voicebots, obtained from a prominent AI voicebot provider operating in China. A typical debt-collection call involves the voicebot's self-introduction, verification of the customer's identity to ensure accuracy, reminders regarding the overdue loan (including details such as the amount due), highlight of the potential consequences of non-repayment—such as additional fees and the impact on credit history—and, finally, an inquiry into the customer's intention to repay the debt.

The AI voicebot provider offers client companies two different voicebot genders—female and male. Each gender includes a range of voice options that differ in tone and style. For example, some female voices are sweet and gentle, whereas others are light and enthusiastic. Importantly, the gender of the voice is easily identifiable to any listener. In our dataset, customers were contacted only once, within one to two days after their payment became overdue.

The voicebot is deployed at the level of the line of business (LOB), which corresponds to a specific business unit or a product line within a client company. Based on our interview, the deployment of a voicebot involves two steps. First, a designer from the voicebot provider collaborates with the client company to design the dialog flow and scripts to meet the LOB's needs. Second, a voice, including its gender, is chosen to serve both female and male customers within that LOB. The selection of gender may depend on factors such as the designer's preferences and the characteristics of the LOB. Furthermore, the gender of the bot may not always stay the same. Some LOBs experimented with different bot genders, especially during the early stages of deployment. Even when a different gender was used, the dialog flow and scripts remained the same for a given LOB in our dataset.

Our unit of analysis was a collection call. We chose to focus on calls from LOBs that had used both female and male voicebots, resulting in six LOBs from five credit card companies. One credit card company had two LOBs because one of them was operated by a subsidiary for a different product line. For each LOB–voice combination, we randomly selected 1,200 connected calls from the dataset—if there were fewer than 1,200 connected calls, we included all of them.[2]

[2] We had 20 LOB–voice combinations in total, with each LOB using between two to six available voices. The distribution of calls was highly skewed among the 20 LOB–voice combinations, with one LOB–voice having 12,000 connected calls (71.5% of the total), four having between 400 to 1,200 calls, and 15 having fewer than 300 calls. To avoid over-representation from the LOB–voice with the most calls, and to preserve the data from smaller LOB–voice, we used 1,200 calls, or 10% of all calls from the largest LOB–voice, as our sampling threshold. As a robustness check, we also tested alternative sampling thresholds of 300, 600, and 900 calls and obtained consistent results (available upon request).

This resulted in a total of 5,994 calls. We then removed calls answered by individuals other than the intended borrowers and calls where customers indicated that they had already repaid the debt. Furthermore, we removed 12 calls where the amount due was zero. Ultimately, we obtained a final dataset of 5,136 calls.

Our sample included 1,039 female and 4,097 male customers. A total of 74.8% of female customers and 74.9% of male customers were served by female voicebots. These near-equal proportions suggest that customer gender was independent of bot gender (p = 0.949).

Measurement and Model Specification

Our outcome variable of interest was the customer's intention to repay the overdue debt ($RepayInt_{ij}$), a metric commonly used by voicebot service providers and their clients. Although we did not have data on actual repayment, our interviews with client companies indicated that repayment intention serves as a strong proxy. If a customer promised or agreed to repay the loan, pay the minimum amount due, or pay later, we defined repayment intention as 1; otherwise, as 0. Repayment intention was automatically classified by the voicebot provider using natural language understanding and manually verified by our research assistants. Because our dependent variable was binary, we adopted a logistic regression model to analyze the repayment intention associated with each call:

$$logit\left(RepayInt_{ij}\right) = \alpha + \beta FemaleBot_{ij} + \gamma Controls_{ij} + \varepsilon_{ij} \quad (1)$$

We used i to index the customer and j to index the LOB. Given that each customer was called only once, each call was uniquely indexed by ij. Our main independent variable was the voicebot gender ($FemaleBot_{ij}$), equal to 1 if the bot had a female gender, and 0 otherwise. In equation (1), α and ε_{ij} are the intercept and the random disturbance, respectively. We were mainly interested in the coefficient β, which captures the effect of bot gender. We included several control variables in the model, including customer gender ($FemaleCust_{ij}$), total amount due ($AmtDue_{ij}$), time-of-day dummies—$Morning_{ij}$ (8 am–12 pm) and $Night_{ij}$ (5 pm–9 pm), city-tier dummies, LOB dummies, and week dummies. The city-tier dummies are based on the tier of the customer's city of residence, which ranges from 1 (the most developed) to 5 (the least developed) according to the popular China city-tier system. A city's tier, reflecting factors like population and affluence, approximates a customer's financial status and exposure to AI voicebots. The LOB dummies control for LOB-specific effects and the week dummies control for time effects. The variable descriptions are in Table 1.

Table 1. Descriptions of Key Variables (N=5136)

Variables	Description	Percentage of 1's (except for *AmtDue*)
RepayInt	Whether the customer explicitly promises to repay the overdue debt (Yes=1, No=0)	52.47%
FemaleBot	Whether the voicebot is female (female=1; otherwise, 0)	74.82%
FemaleCust	Whether the customer is female (female=1; otherwise, 0)	20.23%
BehavTens	Whether the customer interrupts the bot frequently (in over 85% of the dialogs)	8.8%
EmoTens	Whether the customer exhibits negative emotions during the interaction	21.16%
StressTime	Whether the call happens during weekdays	76.01%
GenderIncong	Whether the customer and the bot have opposite gender (can be calculated as *abs(FemaleBot – FemaleCust)*)	64.80%
InitialUncoop	Whether the customer behaves uncooperatively (e.g., refuses to provide information, remains silent, or acts unfriendly) during the first round of conversation with the voicebot	22.9%
AmtDue	Natural logarithm of the total amount due	Mean=8.76, SD=1.40, Min=0.11, Max=13.40
Morning	Whether the call happens in the morning (8 am–12 pm)	47.04%
Night	Whether the call happens in the evening (5 pm–9 pm)	14.17%

Mediation Effects

To investigate the proposed mechanisms, we introduced two mediation variables, $BehavTens$ (behavioral tension) and $EmoTens$ (emotional tension). Following the literature (Tóth et al., 2018; Vejnovic et al., 2024; Zheng et al., 2022), we operationalized these variables by measuring tense behaviors and emotions exhibited by customers during the interaction. Specifically, we used frequent interruptions and the presence of negative emotions as signals of behavioral and emotional tensions respectively. Interruptions and negative emotions are behavioral and emotional manifestations of the underlying state of tension or conflict in interpersonal interactions (Barki & Hartwick, 2001). We defined $BehavTens_{ij}$ as whether the customer frequently interrupted the voicebot during the call *ij*. It took a value of 1 when the customer interrupted the voicebot in over 85% of the dialogs, and 0 otherwise.[3] The variable $EmoTens_{ij}$ was a binary indicator of whether customer *i* spoke negative-sentiment words during the interaction with voicebot *j*. We used a dictionary-based text analysis method[4] to detect negative words in the dialogs. If negative-sentiment words were found, $EmoTens_{ij}$ equaled 1; otherwise, it equaled 0. The correlation between $BehavTens$ and $EmoTens$ was 0.123, indicating that these two concepts are distinct. We tested the mediation effects using the procedure outlined by Zhao et al. (2010). Specifically, we estimated three regressions:

$$logit(BehavTens_{ij}) = \alpha + \beta FemaleBot_{ij} + \gamma Controls_{ij} + \varepsilon_{ij} \quad (2)$$

$$logit(EmoTens_{ij}) = \alpha + \beta FemaleBot_{ij} + \gamma Controls_{ij} + \varepsilon_{ij} \quad (3)$$

$$logit(RepayInt_{ij}) = \alpha + \beta FemaleBot_{ij} + \gamma BehavTens_{ij} + \delta EmoTens_{ij} + \theta Controls_{ij} + \varepsilon_{ij} \quad (4)$$

Moderating Effects

We operationalized the moderators as follows. To capture whether a call occurred during a stressful time, we used the variable *StressTime*, equal to 1 if the call occurred during a weekday.

[3] We adopted binary coding for better interpretability. Moreover, the literature suggests that high-frequency interruptions could have a disproportionate effect than moderate or low-frequency interruptions (Speier et al., 1999). Binary coding allows us to focus on high levels of emotional and behavioral tensions and better accommodates the nonlinear relationships.

[4] We used the three most widely applied emotion dictionaries for text-mining in Chinese: the *National Taiwan University Semantic Dictionary*, the *Hownet Semantic Dictionary*, and the *Tsinghua Dictionary of Positive and Negative Sentiment Words*.

We anticipated customers would experience higher stress during weekdays due to heavier workloads and resulting mental exhaustion (Sonnentag et al., 2010). To capture gender incongruence, we defined a variable, *GenderIncong*, equal to 1 if the customer and the voicebot were of different genders. The third moderator, *InitialUncoop*, was a binary variable indicating whether the customer behaved uncooperatively (e.g., refused to provide information, remained silent, or acted unfriendly) during the first round of conversation, which typically consists of the bot's self-introduction and an identity verification question, followed by the customer's response—or lack thereof. We interacted these moderators with the main independent variable $FemaleBot$ in equations (1) through (3), respectively, to test hypotheses H3a through H3c.

Results

Main Findings

In Table 2, Column (1), we present the impact of female voicebot gender on customer repayment intention. The coefficient of *FemaleBot* is positive and significant (β = 0.249, p < 0.01), supporting **H1**, which states that the use of female voicebots leads to higher customer repayment intention. Specifically, compared with male voicebots, female voicebots increase the odds of positive repayment intention by 28.3%, an economically significant increase in repayment intention.

Mediating Effects

In Table 2, Columns (2) to (4), we present the results of the mediation tests. First, as expected, behavioral tension and emotional tension negatively affect customer repayment intention, as indicated by the negative and significant effects of *BehavTens* and *EmoTens* on *RepayInt* (−3.597, p < 0.001; −0.775, p < 0.001, respectively), as shown in Column (4).

Table 2. The Impact of Voicebot Gender

	(1)	(2)	(3)	(4)
Variables	*RepayInt*	*BehavTens*	*EmoTens*	*RepayInt*
FemaleBot	**0.249****	**-0.341***	**-0.277****	0.180
	(0.094)	(0.143)	(0.101)	(0.096)
FemaleCust	0.098	-0.089	0.111	0.101
	(0.076)	(0.130)	(0.087)	(0.080)
AmtDue	0.007	-0.091	0.092*	0.010
	(0.035)	(0.057)	(0.044)	(0.037)
Morning	0.031	-0.041	0.020	0.026
	(0.068)	(0.116)	(0.081)	(0.072)
Night	-0.136	0.277	0.318**	-0.057
	(0.100)	(0.160)	(0.106)	(0.105)
BehavTens				**-3.597*****
				(0.256)
EmoTens				**-0.775*****
				(0.077)
Constant	0.476	-2.109**	-1.917***	0.836*
	(0.388)	(0.734)	(0.457)	(0.413)
LOB FE	Yes	Yes	Yes	Yes
City-tier FE	Yes	Yes	Yes	Yes
Week FE	Yes	Yes	Yes	Yes
Pseudo R2	0.079	0.039	0.024	0.176
N	5136	5136	5136	5136

Notes: * p<0.05, ** p<0.01, *** p<0.001; Robust standard errors in parentheses.

Female voicebots negatively affect the two tension-based mediators, as indicated by the negative and significant coefficients of *FemaleBot* on *BehavTens* (−0.341, $p < 0.05$, Column (2)) and *EmoTens* (−0.277, $p < 0.01$, Column (3)), respectively. In addition, we conducted a causal mediation analysis (Hicks & Tingley, 2011) with 1,000 bootstrap replications, which is well suited to binary or categorical variables. The results of the causal mediation analysis show that both behavioral tension ($p = 0.016$) and emotional tension ($p = 0.016$) significantly mediate the effect of voicebot gender on repayment intention. These findings support **H2a** and **H2b**, indicating that female voicebots increase repayment intention by alleviating behavioral and emotional tensions.

Moderating Effects

We report the moderating effects in Table 3. First, as shown in Column (3), when calls took place during weekdays (*StressTime*), female voicebots were more effective at improving repayment intention (0.448, $p < 0.05$). Second, the coefficient of *FemaleBot* × *GenderIncong* is marginally significant (0.364, $p = 0.051$) and, interestingly, the coefficient of *FemaleCust* becomes significant (0.396, $p < 0.05$). Together, these results reveal that male customers have lower repayment intention (as indicated by the positive coefficient of *FemaleCust*), but their repayment intention increases when interacting with a female voicebot. Third, female voicebots are more successful in persuading initially uncooperative customers to repay (0.503, $p < 0.05$).

We report the effects of moderators on the mediation variables—behavioral and emotional tensions—in Table 3, Columns (1)–(2). As shown in the table, calling customers at a stressful time significantly amplifies the advantage of female voicebots in reducing behavioral tension (−0.974, $p < 0.05$), but not emotional tension, partially supporting **H3a**. The lack of significance regarding emotional tension might be because customers refrained from using negative words during service calls that were often recorded, adding more noise to our emotional tension variable. As mentioned above, gender incongruence positively moderate the main effect of voicebot gender on *customer repayment intention*. However, gender incongruence does not influence the impact of voicebot gender on *behavioral* or *emotional tensions*. Thus, **H3b** is only partially supported. A possible explanation for this finding is that gender incongruence moderates the gender effect on repayment intention through mechanisms other than alleviating tensions. For example, prior research shows that men tend to be more generous in the presence of female observers (Siegel et al., 2009; Van Vugt & Iredale, 2013), which may extend to interactions where gender incongruence elicits similar social dynamics. Last, we find support for **H3c**, where the advantage of female voicebots in reducing behavioral and emotional tensions is significantly amplified by customers' initial uncooperativeness (−0.882, $p < 0.001$; −0.421, $p < 0.05$).

Table 3. Moderation Effects			
	(1)	**(2)**	**(3)**
Variables	*BehavTens*	*EmoTens*	*RepayInt*
FemaleBot	0.905	0.085	-0.327
	(0.506)	(0.239)	(0.234)
FemaleCust	-0.146	0.111	0.396*
	(0.248)	(0.163)	(0.163)
AmtDue	-0.073	0.092*	-0.032
	(0.059)	(0.044)	(0.039)
Morning	-0.013	0.036	0.024
	(0.117)	(0.081)	(0.073)
Night	0.233	0.303**	-0.124
	(0.161)	(0.107)	(0.110)
StressTime	0.786	0.083	-0.521*
	(0.413)	(0.198)	(0.203)
InitialUncoop	1.582***	0.218	-2.541***
	(0.207)	(0.162)	(0.220)
*FemaleBot*StressTime*	**-0.974***	-0.305	**0.448***
	(0.427)	(0.214)	(0.217)
*FemaleBot*GenderIncong*	-0.032	0.006	**0.364+**
	(0.293)	(0.191)	(0.187)
*FemaleBot*InitialUncoop*	**-0.882*****	**-0.421***	**0.503***
	(0.245)	(0.193)	(0.244)
Constant	-3.373***	-2.127***	1.135*
	(0.798)	(0.475)	(0.444)
LOB FE	Yes	Yes	Yes
City-tier FE	Yes	Yes	Yes
Week FE	Yes	Yes	Yes
Pseudo R2	0.070	0.026	0.181
N	5136	5136	5136

Notes: * p<0.05, ** p<0.01, *** p<0.001, + p=0.051; Robust standard errors in parentheses.[5]

Robustness Checks

Selection Bias

One challenge to our estimations was the potential selection bias arising from the non-random assignment of bot gender—as discussed earlier, the choice of bot gender could depend on factors such as the AI designer's preference and characteristics of the LOB. To address such concerns, we focused on LOBs that used both female and male voicebots and introduced several control variables, LOB-fixed effects, and time-fixed effects. To further alleviate this concern, we

[5] *GenderIncog* is not included as a standalone variable in Table 3 because it causes multicollinearity issues and is redundant given the existing *FemaleBot* and *FemaleCust* variables.

employed the Heckman Selection method and matching techniques, as recommended by Hill et al. (2021). We briefly discuss these approaches and findings below. Details of our implementation and results are omitted due to space constraints but are available upon request.

Heckman Selection Model. To account for potential unobserved factors that could influence the selection of bot gender, we employed the Heckman selection model. This approach allowed us to estimate the selection equation and then include the inverse Mills ratio (IMR) derived from the selection equation in the main estimation, helping to control for non-random selection. We used customer gender, amount due, time-of-day, and stressful time in the selection equation, along with an instrumental variable (IV), $DeltaRepay_{jt}$, calculated as the change in the average repayment intention for the voicebot of LOB *j* from time *t-1* to time *t-2*. Results based on the Heckman selection model were similar to our main analyses. Moreover, the coefficient of IMR was not significant, suggesting that selection bias may not be a concern.

Propensity Score Matching. Propensity score matching (PSM) is one of the most popular approaches to addressing the selection bias problem. PSM ensures that treatment and control units have similar probabilities of receiving the treatment given their observable characteristics so that they are comparable. We included several observable covariates that could influence the selection of voicebot gender as matching variables, including customer gender, amount due, time-of-day, and the voicebot's prior performance metrics (number of calls, average call duration, and proportion of customers willing to repay). These metrics served as proxies for the voicebot's effectiveness and may impact future gender choices in interactions. Using the matched sample,[6] we reran all the analyses, and the results were mostly consistent with our

[6] To enhance matching quality, we employed PSM using the kernel matching method within each LOB. This approach ensures that the matching is performed among comparable units within each LOB, reducing potential confounding factors (Caliendo & Kopeinig, 2008). Samples that fell outside the common support region were

previous findings.

Causality of Mediation Variables

Another concern was whether the mediating variables—behavioral and emotional tensions—truly caused repayment intention. We measured mediation variables before repayment intention to limit reverse causality; but omitted variables, such as the customer's unobserved attitude, may impact both tensions and repayment intention, creating spurious mediation effects. We employed three approaches to mitigate this concern.

First, we controlled for confounding customer attitudes. An uncooperative customer may display behavioral and emotional tensions and refuse to repay the loan in the end. We approximated such customer attitudes using the variable *InitialUncoop*, which measures whether the customer was uncooperative in the first round of conversation. By controlling for *InitialUncoop*, we found that the effects of tensions on repayment intention remained significant, supporting a causal relationship between tensions and repayment intention.

Second, we conducted a sensitivity analysis for causal mediation analysis (Hicks & Tingley, 2011) to assess the robustness of the mediation effects to potential unmeasured confounders. We found that the average causal mediation effect became zero only when the sensitivity parameter ρ—representing the correlation between the residuals from the mediator and outcome models—reached −0.5 for behavioral tension and −0.3 for emotional tension. These values indicated a moderate level of robustness relative to other reported values in the literature, supporting the causality of our mediators.

Third, we employed the Two-Stage Residual Inclusion (2SRI) estimation method (Terza

excluded, ensuring that only comparable units were retained. The quality of the PSM was assessed using indicators like pseudo R-squared, chi-square, and likelihood ratio, mean, and median biases (Kim et al., 2016). The matched sample consisted of 953 male bot calls and 1,800 female bot calls. Additionally, attempts to match across different LOBs yielded consistent findings, further validating the robustness of the matching procedure.

et al., 2008), an alternative to the Two-Stage Least Squares for nonlinear models. We identified two IVs each for behavioral and emotional tensions: the average behavioral tension and emotional tension of the same voice within each voice gender, and the average behavioral tension and emotional tension of the same LOB in other calls within the sample. The results of the Likelihood ratio test and Wald test were both equal to 0, validating that these IVs were not weak. The estimation results of 2SRI verified the causal effects of behavioral and emotional tensions on repayment intention.

Acoustic Features as Alternative Mechanisms

Female voices are different from male voices on several acoustic dimensions such as pitch. Prior studies have shown that acoustic features, including the speaker's *speech rate*, *pitch* (mean), *pitch variation*, *loudness* (mean), and *loudness variation*, can influence the perception of dominance and credibility (Chebat & Hedhli, 2007). It is likely that the voice's acoustic features, not voice gender per se, drive the results. To account for such a possibility, we added the aforementioned five major acoustic features as additional controls. Our findings remained robust, which lent further support to our conclusions.

Discussion

Our study addresses a timely issue regarding how the choice of bot gender may impact service performance, particularly in tense service contexts. Drawing on the TBV of customer relationships and gender stereotype theory, we theorize whether, how, and when voicebot gender may play a role in tense service contexts. Our analyses of a bot-powered debt-collection dataset largely confirm our theoretical predictions and yield several practical insights. First (***whether***), female voicebots increase the odds of a positive repayment intention by 28.3%, confirming our theoretical prediction of female advantages in tense service contexts. Second (***how***), the effect of voicebot gender is mediated by behavioral and emotional tensions, confirming our theoretical arguments about underlying mechanisms. Third (***when***), as predicted by our theory, female voicebots lead to higher customer repayment intention in tense situations, such as when calls occur during stressful times, between opposite genders, or with initially uncooperative customers, which further lends credence to our theory.

Contributions and Theoretical Implications

This study contributes to HCI research in two main ways. First, we extend traditional HCI research on bot gender, which has primarily focused on friendly or neutral settings (e.g., Benbasat et al., 2020; Borau et al., 2021), to tense service contexts. We theorize that female bots lead to improved service outcomes in such settings because customers' association of female stereotypes with the bots can lead to reduced behavioral and emotional tensions, which, in turn, can improve service outcomes. This tension-based framework offers a novel theoretical lens for understanding the role of gender in human–bot interactions that involve tensions, moving beyond conventional explanations based on perceived attributes such as humanness or competence (Ahn et al., 2022; Borau et al., 2021). This context-centered theorizing approach opens new avenues for future research in this area. Although our study is empirically tested using debt collection, our findings have implications for other tense contexts such as service failure recoveries, payment disputes, and customer complaints. We hope our study paves the way for more research on such important, yet understudied human–bot interaction contexts.

Second, our study offers one of the first pieces of real-world evidence on the impact of bot gender on service outcomes, complementing existing laboratory-based evidence on human perceptions of and attitudes toward robots (e.g., Bernotat et al., 2021; Qiu & Benbasat, 2010). The natural setting reveals how bot gender effects vary across different situation-specific factors—such as calling time—indicating that gender effects are more nuanced than previously theorized and that optimal gender choice varies based on immediate circumstances rather than remaining fixed within a given context. Therefore, in addition to drawing broad conclusions about bot gender effects for general contexts (Bernotat et al., 2021) or specific industry domains (Ahn et al., 2022; Forgas-Coll et al., 2022), researchers may want to delineate distinct factors favoring female or male robots that could be applied in many different settings.

Our study also makes significant contributions to the research on the TBV of relationships. First, we contribute to the emerging literature that extends the scope of TBV from B2B to B2C settings (Alvarez et al., 2021; Vejnovic et al., 2024). To our knowledge, we are the first to introduce TBV to the information system (IS) literature. We hope our study opens a door to using TBV to study tense IS phenomena, such as tense user–developer interactions, change management, and security- and privacy-related interactions. Second, the original TBV tends to be more descriptive and takes tensions in a relationship as given (Fang et al., 2011). We demonstrate that design and contextual factors (such as gender and calling time) can influence the level of tension in a relationship and thus impact service outcomes. Our findings, therefore, encourage TBV researchers to adopt a more dynamic and proactive view of tensions, examining the role of technology and other contextual factors.

Managerial Implications

As more organizations incorporate bots into their service operations, there is a tremendous interest in understanding consequential design choices for such bots. Our findings show that female voicebots hold a sizable advantage over their male counterparts, dispelling the myth that the preference for female bots is merely a gender bias in the AI industry. Therefore, we suggest deploying female voicebots in tense service contexts and investing more resources in developing female voices. More broadly, our findings emphasize that social design elements are as crucial as technical features when organizations make AI development and investment choices.

In addition, recognizing that female voicebots work by reducing behavioral and emotional tensions, managers should align other service elements to complement this tension-reduction mechanism. Beyond bot gender, AI vendors should leverage complementary design dimensions—such as communication styles, tones, and conversation strategies—to further enhance bots' ability to mitigate tensions. Although our study focuses on tense service contexts such as debt collection, we expect our findings to be relevant for tense service episodes within otherwise non-tense service contexts, such as a sales pitch call to a highly uncooperative customer or a customer service call that leaves the customer waiting for a long time. We recommend that organizations identify such contextual factors and episodes that could benefit from deploying female voicebots.

Limitations and Future Work

Several limitations of this study require further research. First, we use repayment intention as our dependent variable. While repayment intention is highly correlated with actual repayment and is a key industry metric, research linking bot gender and actual repayment can further strengthen this line of investigation. Second, our estimated gender effects may be biased by unobserved temporal confounders. Although our robust checks suggest that it is not a significant threat, further research based on randomly assigned bot gender would further eliminate such concerns. Third, we find that gender incongruence moderates the relationship between voicebot gender and customer repayment intentions, but it does not have a significant moderating effect on the relationship between voicebot gender and tensions. Future research should further explore the mechanisms through which gender incongruence might influence customer–voicebot interactions. Finally, we have limited information on customers due to data privacy restrictions. A promising direction for future research is to explore customer characteristics that alter the effects of bot gender.

References

Ahn, J., Kim, J., & Sung, Y. (2022). The effect of gender stereotypes on artificial intelligence recommendations. *Journal of Business Research*, *141*, 50–59.

Alvarez, C., Brick, D. J., & Fournier, S. (2021). Doing Relationship Work: A Theory of Change in Consumer–Brand Relationships. *Journal of Consumer Research*, *48*(4), 610–632.

Barki, H., & Hartwick, J. (2001). Interpersonal conflict and its management in information system development. *MIS Quarterly*, *25*(2), 195–228. https://doi.org/10.2307/3250929

Beldad, A., Hegner, S., & Hoppen, J. (2016). The effect of virtual sales agent (VSA) gender – product gender congruence on product advice credibility, trust in VSA and online vendor, and purchase intention. *Computers in Human Behavior*, *60*, 62–72.

Benbasat, I., Dimoka, A., Pavlou, P. A., & Qiu, L. (2020). The role of demographic similarity in people's decision to interact with online anthropomorphic recommendation agents: Evidence from a functional magnetic resonance imaging (fMRI) study. *International Journal of Human-Computer Studies*, *133*, 56–70.

Bernotat, J., Eyssel, F., & Sachse, J. (2021). The (Fe)male Robot: How Robot Body Shape Impacts First Impressions and Trust Towards Robots. *International Journal of Social Robotics*, *13*(3), 477–489.

Binnewies, C., Sonnentag, S., & Mojza, E. J. (2010). Recovery during the weekend and fluctuations in weekly job performance: A week-level study examining intra-individual relationships. *Journal of Occupational and Organizational Psychology*, *83*(2), 419–441.

Borau, S., Otterbring, T., Laporte, S., & Fosso Wamba, S. (2021). The most human bot: Female gendering increases humanness perceptions of bots and acceptance of AI. *Psychology & Marketing*, *38*(7), 1052–1068.

Caliendo, M., & Kopeinig, S. (2008). Some Practical Guidance for the Implementation of Propensity Score Matching. *Journal of Economic Surveys*, *22*(1), 31–72.

Cambre, J., & Kulkarni, C. (2019). One Voice Fits All?: Social Implications and Research Challenges of Designing Voices for Smart Devices. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–19.

Chebat, J.-C., & El Hedhli, K. (2007). Voice and Persuasion in a Banking Telemarketing Context. *Perceptual & Motor Skills*, *104*(2), 419–437.

Cheng, X., Zhang, X., Cohen, J., & Mou, J. (2022). Human vs. AI: Understanding the impact of anthropomorphism on consumer response to chatbots from the perspective of trust and relationship norms. *Information Processing & Management*, *59*(3), 102940.

Crowell, C. R., Scheutz, M., Schermerhorn, P., & Villano, M. (2009). Gendered voice and robot entities: Perceptions and reactions of male and female subjects. *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 3735–3741. https://doi.org/10.1109/IROS.2009.5354204

Custers, A. (2017). Falling Behind: Debtors' Emotional Relationships to Creditors. In A.

Mooney & E. Sifaki (Eds.), *The Language of Money and Debt* (1st ed., pp. 163–185). Palgrave Macmillan, Cham, Switzerland.
CX Today. (2021, April 19). *Understanding the Promise to Pay Metric*. https://www.cxtoday.com/loyalty-management/understanding-the-promise-to-pay-metric-and-how-to-improve-it/
Deville, J. (2015). *Lived economies of default: Consumer credit, debt collection and the capture of affect* (1st ed.). Routledge, Taylor & Francis Group, New York, USA.
Eyssel, F., & Hegel, F. (2012). (S)he's Got the Look: Gender Stereotyping of Robots. *Journal of Applied Social Psychology*, *42*(9), 2213–2230.
Eyssel, F., Kuchenbrandt, D., Bobinger, S., De Ruiter, L., & Hegel, F. (2012). 'If you sound like me, you must be more human': on the interplay of robot and user features on human-robot acceptance and anthropomorphism. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, 125-126. https://ieeexplore.ieee.org/document/6249487
Fang, S.-R., Chang, Y.-S., & Peng, Y.-C. (2011). Dark side of relationships: A tensions-based view. *Industrial Marketing Management*, *40*(5), 774–784.
Forgas-Coll, S., Huertas-Garcia, R., Andriella, A., & Alenyà, G. (2022). The effects of gender and personality of robot assistants on customers' acceptance of their service. *Service Business*, *16*(2), 359–389.
Fossa, F., & Sucameli, I. (2022). Gender Bias and Conversational Agents: An ethical perspective on Social Robotics. *Science and Engineering Ethics*, *28*(3), 23.
Gelbrich, K. (2010). Anger, frustration, and helplessness after service failure: Coping strategies and effective informational support. *Journal of the Academy of Marketing Science*, *38*(5), 567–585.
Gnyawali, D. R., Madhavan, R., He, J., & Bengtsson, M. (2016). The competition–cooperation paradox in inter-firm relationships: A conceptual framework. *Industrial Marketing Management*, *53*(02), 7–18.
Harrington, L. (2018). "Helping you to pay us": Rapport management in debt collection call centre encounters. *Journal of Politeness Research*, *14*(2), 193–223.
Heidenreich, S., Wittkowski, K., Handrich, M., & Falk, T. (2015). The dark side of customer co-creation: Exploring the consequences of failed co-created services. *Journal of the Academy of Marketing Science*, *43*(3), 279–296.
Hibbard, J. D., Brunel, F. F., Dant, R. P., & Iacobucci, D. (2001). Does Relationship Marketing Age Well? *Business Strategy Review*, *12*(4), 29–35.
Hicks, R., & Tingley, D. (2011). Causal Mediation Analysis. *The Stata Journal: Promoting Communications on Statistics and Stata*, *11*(4), 605–619.
Hill, A. D., Johnson, S. G., Greco, L. M., O'Boyle, E. H., & Walter, S. L. (2021). Endogeneity: A Review and Agenda for the Methodology-Practice Divide Affecting Micro and Macro Research. *Journal of Management*, *47*(1), 105–143.
Hill, R. P. (1994). Bill Collectors and Consumers: A Troublesome Exchange Relationship. *Journal of Public Policy & Marketing*, *13*(1), 20–35.
Industryresearch. (2022, September 22). *Global Digital Debt Collection Software Industry Research Report*. https://www.industryresearch.biz/global-digital-debt-collection-software-industry-research-report-competitive-landscape-market-21738622
Karimi, S., Holland, C. P., & Papamichail, K. N. (2018). The impact of consumer archetypes on online purchase decision-making processes and outcomes: A behavioural process

perspective. *Journal of Business Research*, *91*, 71–82.
Kim, S. H., Mukhopadhyay, T., Carnegie Mellon University, Carnegie Mellon University in Qatar, Kraut, R. E., & Carnegie Mellon University. (2016). When Does Repository KMS Use Lift Performance? The Role of Alternative Knowledge Sources and Task Environments. *MIS Quarterly*, *40*(1), 133–156.
Kostopoulos, G., Gounaris, S., & Rizomyliotis, I. (2014). How to reduce the negative impact of customer non-compliance: An empirical study. *Journal of Strategic Marketing*, *22*(6), 513–529.
Kotler, P., & Keller, K. L. (2008). *Marketing Management* (13th ed.). Pearson College Div.
Li, Y., Wang, H., Gao, H., Li, Q., & Sun, G. (2024). Credit rating, repayment willingness and farmer credit default. *International Review of Financial Analysis*, *93*, 103117.
Mazar, N., Mochon, D., & Ariely, D. (2018). If You Are Going to Pay Within the Next 24 Hours, Press 1: Automatic Planning Prompt Reduces Credit Card Delinquency. *Journal of Consumer Psychology*, *28*(3), 466–476.
McDonnell, M., & Baxter, D. (2019). Chatbots and Gender Stereotyping. *Interacting with Computers*, *31*(2), 116–121.
Nass, C., & Brave, S. (2005). *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship* (1st ed.). MIT press, MA: USA.
Nass, C., Moon, Y., & Green, N. (1997). Are Machines Gender Neutral? Gender-Stereotypic Responses to Computers With Voices. *Journal of Applied Social Psychology*, *27*(10), 864–876.
Oliveira, N., & Lumineau, F. (2019). The Dark Side of Interorganizational Relationships: An Integrative Review and Research Agenda. *Journal of Management*, *45*(1), 231–261.
Oliver, B. (2020, May 7). AI and me:friendship chatbots are on the rise, but is there a gendered design flaw? *The Guardian*. https://www.theguardian.com/careers/2020/may/07/ai-and-me-friendship-chatbots-are-on-the-rise-but-is-there-a-gendered-design-flaw
Pfeuffer, N., Adam, M., Toutaoui, J., Hinz, O., & Benlian, A. (2019). Mr. And Mrs. Conversational Agent—Gender Stereotyping in Judge-Advisor Systems and the Role of Egocentric Bias. I *Proceedings of the International Conference on Information Systems*. https://core.ac.uk/display/301383720.
Phillips, L., & Moggridge, P. (2019). Artificial intelligence in debt collection. *Credit Control* 40(2):1-11.
Pillai, K. G., Sharma, P., Cornelissen, J., Zhang, Y., & Nair, S. R. (2023). Theorising the dark side of interorganizational relationships: An extension. *Journal of Business & Industrial Marketing*, *38*(12), 2578–2588.
Pitardi, V., Bartikowski, B., Osburg, V.-S., & Yoganathan, V. (2023). Effects of gender congruity in human-robot service interactions: The moderating role of masculinity. *International Journal of Information Management*, *70*, 102489.
Powers, A., Kramer, A. D. I., Lim, S., Kuo, J., Sau-lai Lee, & Kiesler, S. (2005). Eliciting information from people with a gendered humanoid robot. *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005*,158–163. https://doi.org/10.1109/ROMAN.2005.1513773
Prentice, D. A., & Carranza, E. (2002). What Women and Men Should Be, Shouldn’t Be, Are Allowed to Be, and Don’t Have to Be: The Contents of Prescriptive Gender Stereotypes. *Psychology of Women Quarterly*, *26*(4), 269–281.
Pune. (2024, November 27). *Chatbot Market to Reach USD 36.3 Billion by 2032*.

GlobeNewswire. https://www.globenewswire.com/news-release/2024/11/27/2988234/0/en/Chatbot-Market-to-Reach-USD-36-3-Billion-by-2032-Owing-to-Advancements-in-AI-Technology-and-Growing-Demand-for-Automated-Customer-Support-Research-by-SNS-Insider.html

Qiu, L., & Benbasat, I. (2010). A study of demographic embodiments of product recommendation agents in electronic commerce. *International Journal of Human-Computer Studies*, *68*(10), 669–688.

Siegel, M., Breazeal, C., & Norton, M. I. (2009). Persuasive Robotics: The influence of robot gender on human behavior. *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2563–2568. https://doi.org/10.1109/IROS.2009.5354116

Smith, N. A., Martinez, L. R., & Sabat, I. E. (2016). Weight and Gender in Service Jobs: The Importance of Warmth in Predicting Customer Satisfaction. *Cornell Hospitality Quarterly*, *57*(3), 314–328.

Sonnentag, S., Binnewies, C., & Mojza, E. J. (2010). Staying well and engaged when demands are high: The role of psychological detachment. *Journal of Applied Psychology*, *95*(5), 965–976.

Speier, C., Valacich, J. S., & Vessey, I. (1999). The influence of task interruption on individual decision making: An information overload perspective. *Decision sciences*, *30*(2), 337-360.

Stokoe, E., & Sikveland, R. O. (2020). The backstage work negotiators do when communicating with persons in crisis. *Journal of Sociolinguistics*, *24*(2), 185–208.

Terza, J. V., Basu, A., & Rathouz, P. J. (2008). Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. *Journal of Health Economics*, *27*(3), 531–543.

Tóth, Z., Peters, L. D., Pressey, A., & Johnston, W. J. (2018). Tension in a value co-creation context: A network case study. *Industrial Marketing Management*, *70*(04), 34–45.

Tsiotsou, R. H., & Diehl, S. (2022). Delineating transformative value creation through service communications: An integrative framework. *Journal of Service Management*, *33*(4/5), 531–551.

Van Vugt, M., & Iredale, W. (2013). Men behaving nicely: Public goods as peacock tails. *British Journal of Psychology*, *104*(1), 3–13.

Vejnovic, E., Purchase, S., & Tarabashkina, L. (2024). Marketing research agency, creative agency and client relationships: A study of relationship tensions. *Journal of Business & Industrial Marketing*, *39*(3), 461–473.

Zhao, X., Lynch, J., & Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and Truths About Mediation Analysis. *Journal of Consumer Research*, *37*(2), 197–206.

Zheng, L. J., Zhang, Y., Zhan, W., & Sharma, P. (2022). How B2B relationships influence new product development in entrepreneurial firms? The role of psychological tension. *Journal of Business Research*, *139*(02), 1451–1462.

Appendix A

Table A1. Selected Review of Essential HCI Studies on Robot Gender

Study	Context	Method	Findings related to gender effect

Ahn et al. (2022)	industry: e-commerce task: recommendation	online experiment	1. **Gender effect**: males/females are more persuasive for utilitarian/hedonic products. 2. **Mechanism**: warmth mediates gender effects on recommendation evaluation for hedonic products; competence, for utilitarian products.
Beldad et al. (2016)	industry: e-commerce task: recommendation	lab experiment	Gender's effects on credibility, trust, and purchase intention are insignificant.
Benbasat et al. (2020)	industry: e-commerce task: recommendation	lab experiment	Women do not conform to male bots' advice, while men have no preferences.
Benbasat et al. (2010)	industry: e-commerce task: recommendation	lab experiment	The main effect of gender on social presence is not significant. Female users prefer female bots.
Bernotat et al. (2021)	industry: not specified task: not specified	lab experiment	Users trust female bots more. Gender–occupation fit positively affects user preference.
Borau et al. (2021)	industry: healthcare task: unspecified	online experiment	1. **Gender effect**: female bots are more likable. 2. **Mechanism**: perceived humanness and warmth, but not competence, mediate the gender effect.
Crowell et al. (2009)	industry: unspecified task: self-introduction and asking questions	lab experiment	Male-embodied and female-disembodied bots are more reliable than both the male-disembodied and female-embodied bots.
Eyssel & Hegel (2012)	industry: unspecified task: (intended for) math or verbal tasks	lab experiment	Female bots are perceived as more communal. The effect of gender–occupation fit on task fit evaluation is positive.
Eyssel et al.(2012)	industry: unspecified task: reading a sentence	lab experiment	The effect of bot–user gender match on user attitude (e.g., acceptance, perceived humanness) is negative.
Forgas-Coll et al. (2022)	industry: gaming task: assistance	lab experiment	The effect of gender-task fit is positive on user acceptance.
Lee et al. (2000)	industry: unspecified task: offering judge advice	lab experiment	The effect of gender match is positive on opinion conformation. Female voice was perceived to be more socially attractive and trustworthy.
McDonnell & Baxter (2019)	industry: banking or mechanics task: assistance	lab experiment	Male bots enhance user satisfaction, while gender–task congruence yields positive effects.
Nass et al. (1997)	industry: unspecified task: assistance	lab experiment	Praise from males is more convincing than that from females; males are more likable.
Pfeuffer et al. (2019)	industry: math and finance task: assistance	online experiment	1. **Gender effect**: the perceived competence for female bots is higher. 2. **Mechanism**: the mediation effects of agentic, communal, and competence are insignificant.
Pitardi et al. (2023)	industry: airport service task: assistance	online experiment	1. **Gender effect**: gender match elicits positive affect. 2. **Mechanism**: perceived control mediates gender effect only for consumers high on masculinity.
Powers et al. (2005)	industry: dating task: query	lab experiment	Men engage more when communicating with female bots. Female bots are expected to be more knowledgeable.
Qiu & Benbasat (2010)	industry: e-commerce task: recommendation	lab experiment	The effect of gender match on perception (e.g., social presence, usefulness) is insignificant.

Siegel et al. (2009)	industry: nonprofit task: fundraising	lab experiment	Men are more likely to donate when interacting with female robots, while women have no preference.
Tay et al. (2014)	industry: healthcare, security task: medical advice; safety monitoring	lab experiment	The effect of gender–occupation fit is positive on user attitude, evaluation, perceived behavioral control, subjective norms, perceived trust, and acceptance.
This study	**industry: tense service** **task: debt collection**	**archival data analysis**	**1. Gender effect: female bots lead to higher repayment intention.** **2. Mechanism: behavioral and emotional tension reduction mediates the gender effect.**