AI Chatbots in Customer Service: Adoption Hurdles and Simple Remedies

Evgeny Kagan
Carey Business School, Johns Hopkins University, ekagan@jhu.edu

Maqbool Dada
Carey Business School, Johns Hopkins University, dada@jhu.edu

Brett Hathaway
Carey Business School, Johns Hopkins University, bhathaw2@jhu.edu

Despite recent advances in language processing algorithms, chatbot technology continues to face adoption hurdles. We survey chatbot users about their experiences and use their testimonies to construct a decision model of customer choice between the chatbot service channel and the live agent service channel. The fundamentals of this choice are the time spent in line and in service, the chatbot's success rate, and the qualitative differences in the service experience provided by the chatbot and by the live agent. We then conduct experiments in which participants choose, and then experience, the chatbot or the live agent channel as we vary operational (i.e., times spent and chatbot success rates) and qualitative features of the chatbot. We find that users respond positively to improvements in chatbot operational performance; however, the chatbot channel remains underutilized relative to what expected time minimization would predict. Additional experiments show that this underutilization is caused by two separate mechanisms: algorithm aversion (aversion to an algorithmic service provider), and gatekeeper aversion (aversion to any service format that may involve multiple stages). Examining potential remedies, we find that algorithm aversion can be mitigated by making salient the expected time savings offered by the chatbot. However, gatekeeper aversion is more persistent and harder to overcome. We conclude by building and estimating a structural model of channel demand and by proposing a behavior-aware service design that reduces the firm's staffing costs by up to 22%.

Key words: Experiments, human-AI interfaces, behavior in queuing systems, service operations

1. Introduction

Chatbots are algorithmic agents that automate customer support in online settings. Recent advances in language processing algorithms have significantly increased chatbot capabilities, improved their speed, enabled them to handle more-complex, often-unstructured customer queries, and reduced training and maintenance costs (Johannsen et al. 2018). These improvements have significantly reduced the staffing needs for live operators, lowering payroll and other costs related to providing live customer support. The cost savings can be substantial – a recent report estimates an average cost reduction of up to $0.70 per customer interaction, and an annual savings of 8 Billion US Dollars in the retail banking industry alone (Maynard and Crabtree 2020).

The technological maturity and the cost savings offered by chatbots have shifted the burden of successful chatbot deployment from AI developers to managers implementing this technology in their organizations. In this paper we seek to understand and mitigate these implementation hurdles. While there is a growing literature on human-chatbot interactions (Goot and Pilgrim 2019, Goot et al. 2020, Sheehan et al. 2020, Schanke et al. 2021), it is focused mainly on questions related to chatbot design; for example, whether anthropomorphism (human-likeness) helps or hurts adoption, and how engagement and adoption vary for different customer demographics. These studies help developers build chatbots with more desirable appearance and behavior; however, they provide little or no insight into the process design implications of chatbots, their integration into the broader service delivery strategy, and their effects on the cost and performance of a service system.

Operationally, chatbot systems resemble gatekeeper systems (Shumsky and Pinker 2003, Freeman et al. 2017, Hathaway et al. 2022), where the chatbot plays the role of a gatekeeper that handles only a subset of the incoming requests, with the remaining requests being diverted to a live, human agent. This is because certain requests may be difficult to communicate or categorize, or because the chatbot may not be authorized to handle certain requests, for example, ones that involve large cash transfers. Thus, the chatbot serves as the entrance point to, but not necessarily the final step of the service encounter, similar to a nurse in a hospital or a front desk receptionist in a hotel. Different from the healthcare or hospitality settings, which *require* the patient or customer to go through the gatekeeper to begin service, chatbot operators may allow customers to *choose* between a live agent and a chatbot.[1] In this study we examine the determinants, and the implications of this choice, focusing on the following questions:

- What are the key trade-offs driving customer willingness to engage with chatbots?
- What are some simple remedies to increase chatbot uptake?

To begin addressing these questions we first conducted an exploratory survey of online users, asking them to describe a recent customer service encounter. In one version of this survey ($N = 100$) we asked respondents to describe a recent encounter with a live customer service agent. In the other version ($N = 98$) we asked a different set of respondents to describe an encounter with a chatbot.[2]

Notably, 85.00% (resp., 78.57%) of respondents reported having had an interaction with a live agent (resp., chatbot) in the past 12 months. Their testimonies reveal two common themes. First, chatbots require minimal waiting to start the interaction. Users appreciated the instant availability of the chatbot and commented on the expediency with which their request was handled:

[1] The choice may either be explicit (i.e., a direct link to the chatbot and a phone number for live support) or more implicit. For example, there may be an automatic chatbot pop-up on a website, and a live support phone number located in the FAQs.

[2] The survey was conducted on the Prolific platform and consisted of two versions or “treatments”, assigned at random. See Appendix EC.1 for survey details and data summary.

- *I was on a website for a college that I attend that had a chatbot enabled so I decided to use it. I had a question about financial aid and where on campus I could find the financial aid office. I used the bot because it was quicker than trying to navigate through their messy site. (Participant ID: 51)*
- *I had a question about air-travel and I searched up an airline's website, I think American. The quickest option was to use a chatbot, so that's what I decided to do. (Participant ID: 197)*
- *I contacted my internet provider in regards to an unexpected bill increase, so I went to their website and was connected with the chatbot for my billing issue. I used the chatbot because it is faster than calling customer support. Using the chatbot was also faster than looking through the FAQ for my answer. (Participant ID: 105)*

Second, low chatbot success rates were frequently mentioned as a negative feature of the chatbot channel. Consider the following statements related to the inability of chatbots to correctly diagnose or solve the problem:

- *The chatbot sent me in circles and didn't help me with my issue at all. (Participant ID: 124)*
- *There were 4-5 options that I had to choose from, my issue was not among them. (Participant ID: 126)*
- *The chatbot gave me a list of options that had nothing to do with what I was asking. (Participant ID: 180)*

To make more formal comparisons we also asked the respondents to provide several details of their encounter. Their responses show that the chatbot channel is significantly faster to access, with 75.31% of respondents reporting in-line waits of "less than one minute", compared to 23.53% for live agents (rank sum test: $p \ll 0.01$). However, we also asked the respondents about the outcome of their interaction and found that chatbots were able to successfully resolve only 33.77% of requests, relative to 78.82% for the live agents (rank sum test, $p \ll 0.01$). The majority of the chatbot users with unresolved requests were either transferred to a live agent (23.38%) or had to call a live agent (29.87%), thus spending additional time in the system until their request was resolved. Together, these comparisons suggest a speed-performance trade-off in channel choice: chatbots have a lower success rate but are faster to access, while live agents may require a longer wait but are usually the final step of the encounter.

In addition to this (decision-theoretic) trade-off there are qualitative differences between channels. First, chatbots and live agents differ in their communication mode: as pointed out in one of the testimonies quoted above (Participant ID 126), chatbot interactions often involve multiple, semi-structured message exchanges to diagnose and resolve the problem. Such fragmented interactions may be less desirable because they lack the steady experience of being served by a live agent.

Second, users may have an innate preference for working with other humans over working with algorithms – a phenomenon often referred to as algorithm aversion (Dietvorst et al. 2015). Finally, users may dislike any multi-stage service system with a possibly failing gatekeeper and a second, expert service provider (gatekeeper aversion). Whether the gatekeeper is a human or a chatbot may be secondary.

We conduct three experiments to evaluate these explanations. In all three experiments we examine user response to the operational performance features of the chatbot by asking participants to make a series of choices between the live agent and the chatbot channels, as we vary (1) the failure rate of the chatbot, (2) the service time of the chatbot and (3) the waiting time in line for the live agent after chatbot failure. To incentivize truthful reporting of preferences we require participants to experience a random subset of their choices prior to receiving experimental payments.

In the first experiment we study the role of communication mode in two between-subject treatments. In the first treatment the chatbot channel is only different from the live agent channel in its operational parameters (waiting and service times, and failure rate), while in the second treatment the chatbot channel uses a distinctly robotic mode of communication. The results suggest that chatbot uptake is reduced as the chatbot gets slower, has a higher failure rate, and when the customer needs to wait longer for a live agent after chatbot failure. However, chatbot uptake is below what expected time minimization would predict. In addition, there are significant differences in chatbot uptake rates across treatments: chatbot uptake decreases when the chatbot channel requires the customer to send back-and-forth messages and the live agent channel requires the customer to be continually engaged, relative to the treatment where the communication mode is held constant (back-and-forth messages) across channels.

To understand the mechanisms driving the above results we conduct a second experiment, in which we replicate the first experiment but remove the chatbot context, framing the problem as a neutral choice between two unnamed alternatives. This experiment provides two additional insights. First, the neutral-labeled live agent remains the preferred alternative even when the chatbot provides a significant time saving (in expectation) relative to the live agent, and when communication mode is held constant. This suggests a process-related adoption hurdle that is not tied to the algorithmic nature of the chatbot – a behavior that we term "gatekeeper aversion". Second, the treatment effect (the effect of communication mode differences) is no longer significant, suggesting that the treatment effect in the first experiment is caused not primarily by users' real channel experiences, but by their ex ante beliefs. In other words, communication differences matter mainly because they *amplify* the algorithmic nature of the chatbot.

To examine potential solutions for increasing chatbot uptake we conduct a third experiment in which we examine the effects of a nudge that displays total expected waiting times for each channel.

We find that this nudge significantly increases chatbot uptake. We also show that the effectiveness of the nudge is tied to a reduction in algorithm aversion, as customer attention is directed away from the qualitative differences between channels and towards the operational parameters of the channel choice.

We conclude by examining the chatbot service design problem from the perspective of a manager minimizing the total cost of a service system. We use our experimental data to perform a counterfactual analysis that examines the effects of the expected waiting time nudge and of queue priority rules (prioritizing chatbot customers by shortening their wait for a live agent after chatbot failure) on staffing and system cost. Staffing level analysis in a M/D/1 queuing regime suggests that these interventions achieve up to 22% in cost reduction in moderately congested systems.

In addition to advancing the conversation on human-AI interfaces in service operations, our study and findings contribute to the growing experimental literature on queue joining (Kremer and Debo 2016, Flicker and Hannigan 2022), and queue switching and reneging behaviors (Aksin et al. 2020, Buell 2021), as well as to the studies examining the internal decision trade-offs between the value and the cost of waiting in lines (Naor 1969, Ülkü et al. 2020, Hathaway et al. 2021, Luo et al. 2022). Allon and Kremer (2018) offer a comprehensive review of this literature. Our contribution to this literature is as follows. First, while most existing research in this stream focuses on waiting in line, we study both in-line and in-service waits and the trade-offs between the two. Second, while the existing work looks at waiting in general, we focus on the specific context of the channel choice between an algorithmic chatbot and a live server; to achieve this, we study both the operational and the experiential differences between service channels.

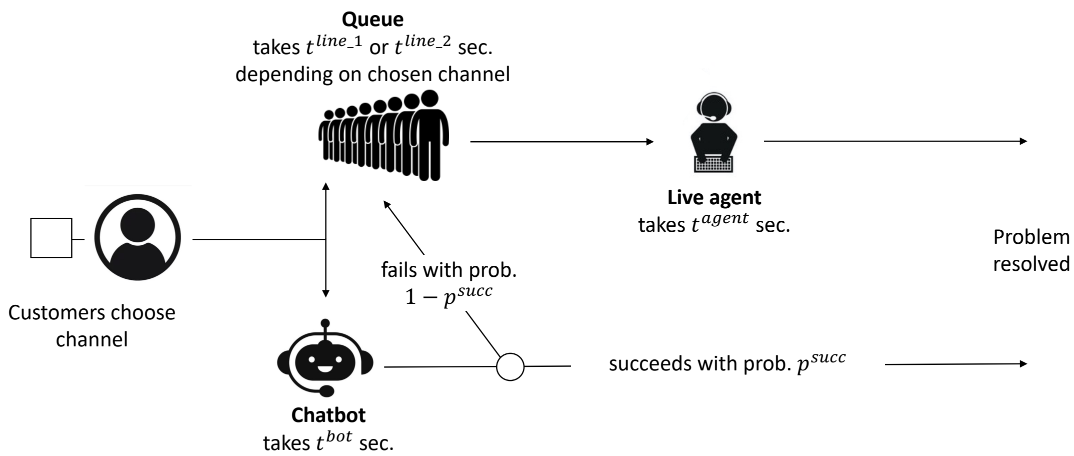
2. Customer Channel Choice Model

In this section we introduce the customer channel choice model. We focus on the main ideas necessary to develop our experimental design and hypotheses. Model details and estimation are deferred to §6.

To represent customer preferences we follow the classic utility-maximization approach in the queueing literature, where the utility of obtaining service is the value of the service minus the cost of waiting (Allon and Kremer 2018, and references therein). If we assume that the cost increases linearly in time, then the total utility of obtaining service can be expressed as $r - w^{line}t^{line} - w^{service}t^{service}$, where r is the value of the service, $w^{line} > 0$ and $w^{service} > 0$ are the stage-specific unit waiting costs and $t^{line} \geq 0$ and $t^{service} \geq 0$ are the times spent in each waiting stage.

Our customer utility model builds on the above framework with two modifications. First, different channels lead to different progressions of waiting stages. Consider first the live agent channel. Due to finite staffing capacity, the customer needs to wait in line until the next live agent becomes

Figure 1 Customer Channel Choice



available. The wait in line takes t^{line_1} seconds, and the wait in service with the agent takes t^{agent} seconds. The live agent always succeeds in resolving the request and the customer exits the system.[3] Consider next the chatbot channel. There is no line, so that the customer immediately proceeds to interacting with the chatbot and spends t^{bot} seconds to complete this interaction. The chatbot has limited cognition and problem-solving skills, resulting in some portion of chatbot interactions being redirected to a live agent, with p^{succ} denoting the probability of chatbot success. If the chatbot succeeds in resolving the request, the customer exits the system. If the chatbot fails, the customer has to wait in line for t^{line_2} and then in service with a live agent agent for t^{agent} seconds. After that, the customer exits the system. To focus on the main decision trade-offs we assume that $t^{bot}, t^{agent}, t^{line\_1}$ and t^{line_2} are constant parameters known to the customer. The sequence of stages is shown in Figure 1.[4]

The second modification is related to the qualitative differences between channels. To minimize the expected time spent in the system, a customer can choose the smaller of $t^{line\_1} + t^{agent}$ and $t^{bot} + (1 - p^{succ})(t^{line\_2} + t^{agent})$. However, customers may also have preferences regarding *how* to spend time – a phenomenon long known to service managers (Maister et al. 1984) and documented in the literature (Kumar et al. 1997, Buell et al. 2017). For example, waiting in service may be perceived to be more pleasant than waiting in line. Further, customers may react disproportionately to certain stages of the wait such as the most recent or the most dissatisfying ones (Redelmeier and Kahneman 1996, Das Gupta et al. 2016), so that different arrangements of the wait may lead

[3] To focus on the main trade-off, we do not consider the additional uncertainty that the customer may have about the ability of the live agent to resolve their request, or about the amount of time the agent will need to do so. Similarly, for simplicity we will assume that all requests are eventually resolved and that the eventual outcome (the customer obtaining a reward r) is the same for both service channels.

[4] Note that t^{line_1} is typically equal to t^{line_2} when the two channels are not integrated. However, t^{line_2} can be chosen to be different from t^{line_1} if the service provider chooses to prioritize one of the two customer classes.

to different perceived costs, even when the total duration of the wait is the same. For example, an initial wait in line for the live agent may be perceived and evaluated differently than a wait in line for the live agent after a chatbot failure. To account for such behaviors we will examine richer representations of waiting costs that go beyond simple linear weights of the line and service times in the customer's utility function.

If we assume that the customer *must* choose either the chatbot or the live agent channel (i.e., has no other outside options), and that the service value r does not depend on channel, then the choice is determined by the comparison of the two channel utilities, defined as follows:

$$U^{agent} = r - g^{agent}(t^{line\_1}, t^{agent}) \tag{2.1}$$

$$U^{bot} = r - g^{bot}(p^{succ}, t^{bot}, t^{line\_2}, t^{agent}), \tag{2.2}$$

where $g^{agent}(\cdot)$ and $g^{bot}(\cdot)$ are the channel disutilities of waiting.

While we defer until §6 the estimation of $g^{agent}(\cdot)$ and $g^{bot}(\cdot)$, the general form of equations (2.1)-(2.2) already provides intuition and structure for our experiments, summarized in Table 1. For example, one would expect $g^{bot}(\cdot)$ to decrease as chatbot performance improves (i.e., as p^{succ} goes up, or as t^{bot} or t^{line_2} go down). A lower disutility incurred in the chatbot channel should

Table 1 Experiment Design

Experiment objectives	Treatment	Treatment description	No. of subjects (recruited/ passed screening/ w. consistent choices)
Experiment 1 (§3): What are the key determinants of the customer channel choice between a chatbot and a live agent?	*Gen*	Contextualized + generic communication mode	109/101/85
	Spec	Contextualized + channel-specific communication mode	106/97/81
Experiment 2 (§4): What are the mechanisms behind chatbot aversion?	*NeutGen*	Neutral frame + generic communication mode	107/98/81
	NeutSpec	Neutral frame + channel-specific communication mode	116/98/78
Experiment 3 (§5): What can managers do to increase chatbot uptake?	*Spec + Nudge*	Contextualized + channel-specific communication mode + expected waiting time information	105/95/81
	NeutGen + Nudge	Neutral frame + generic communication mode + expected waiting time information	103/94/85

Notes: Screening questions included quiz questions (administered prior to the main task) as well as questions about any technical difficulties subjects experienced during the task. Subjects were allowed at most 3 quiz errors to pass the screening questions. Consistency is defined as having no more than one switching point between alternatives (i.e., no violations of monotonicity) during the preference elicitation.

increase the observed chatbot uptake. These and related behaviors are examined in Experiment 1 (§3). Further, we would expect $g^{bot}(\cdot)$ to depend on whether we make salient the algorithmic nature of the chatbot. Experiment 2 (§4) unpacks this and other mechanisms. Similarly, $g^{bot}(\cdot)$ may be lower if we make explicit the expected time savings afforded by the chatbot. In Experiment 3 (§5) we examine the reaction to such a nudge. Finally, in §6 we use our data to estimate U^{agent} and U^{bot} under several functional form specifications. We then use these estimates to develop recommendations for the firm's service design decisions and quantify their impact on staffing costs.

3. Experiment 1: Channel Choice

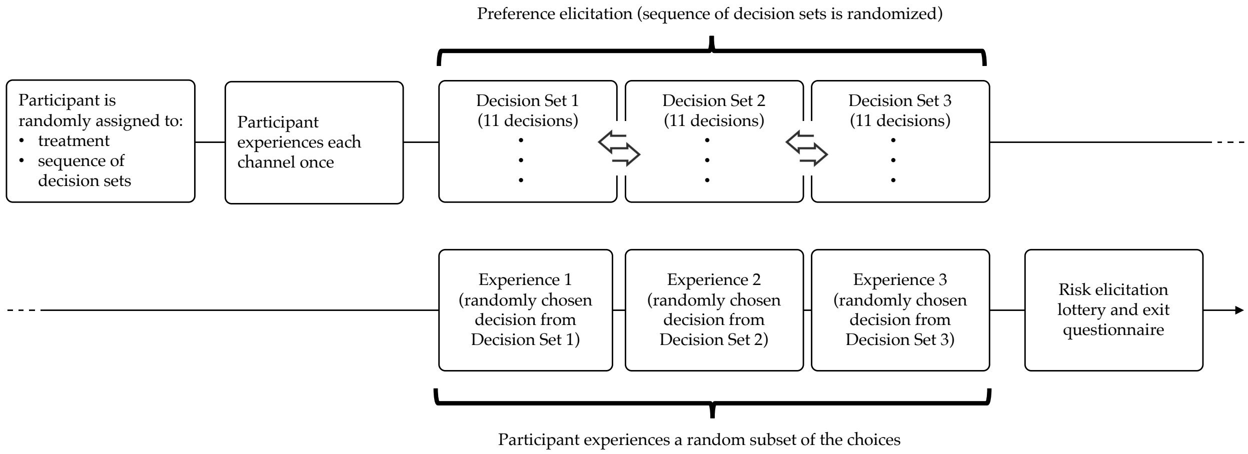
In Experiment 1 we focus on two questions. First, we examine how customers respond to differences in operational performance of chatbots (probability of success, time in service, time in line for live agent after chatbot failure). Second, we examine whether the mode of communication matters. That is, do users respond negatively to the distinctly robotic flow of interaction with a chatbot? To perform the experiment, we recruited 215 participants on the Prolific platform. A total of 198 participants passed the screening (comprehension questions and attention checks). All experiments were programmed in oTree (Chen et al. 2016).

3.1. Experiment Design and Hypotheses

3.1.1. Experimental Flow Figure 2 shows the flow of the experiment. After being randomly assigned to a treatment, participants first experience both the live agent channel and the chatbot channel. We describe the channel experience in each treatment in more detail in §3.1.4. Thereafter, participants make a total of 33 decisions, subdivided into three decision sets of 11 decisions. The sequence of decision sets was randomized to control for any order effects. Each of the 33 decisions is a binary choice between the live agent channel and the chatbot channel. The live agent channel requires an in-line wait of t^{line_1} seconds to begin service, and a wait of t^{agent} seconds to complete service. The live agent resolves the request with probability one. In contrast, the chatbot requires no wait in line and a wait of t^{bot} seconds in service. The chatbot fails with $1-p^{succ}$, which leads to the participant spending an additional $t^{line\_2}+t^{agent}$ seconds in the system. The chatbot parameters, p^{succ}, t^{line_2} and t^{bot} vary for each of the 33 decisions in ways that will be discussed below.

3.1.2. Elicitation Within each decision set we used the Multiple Price Lottery mechanism (Holt and Laury 2002) to elicit preferences. The basic idea of this mechanism is to present participants with a list of binary decisions, where one of the alternatives becomes more desirable as one goes down the list. In Decision Set 1 we varied the success rate of the chatbot (p^{succ}) in increments of 0.05 from 0.25 to 0.75 while all other parameters were held constant. In Decision Set 2 we varied the service time of the chatbot (t^{bot}) in increments of 2 from 30 to 10. In Decision Set 3 we varied

Figure 2 Experimental Flow



the time spent in line if the chatbot fails (t^{line_2}) in increments of 4 from 40 to 0. Across all three decision sets we kept constant the difference in expected times between the two alternatives for each decision within a decision set (i.e., Decision 1 in Decision Set 1 has the same expected time difference as Decision 1 in Decision Set 2 and Decision 1 in Decision Set 3, and similarly for the remaining 10 decisions). See Table 2 for the parameters of all 33 decisions.

3.1.3. Incentives Once a participant completed all their decisions, a subset of the participant's decisions was selected to be experienced in real time. Specifically, one decision from each decision set was selected at random for the real experience; thus, each participant experienced three of their 33 choices prior to receiving their (fixed) dollar payment and exiting the experiment. Participants received $1.50 as a show up fee, and an additional $3 at the end of the experiment, once they had completed all their decisions and waiting experiences. The average time spent in the experiment was 18 minutes and the average payment was $5.70.[5]

3.1.4. Treatments In addition to studying the response to chatbot performance, we wanted to examine whether non-operational parameters, i.e., factors other than the times spent in line and in service and the probability of chatbot success, affect channel preferences. To this end we conducted two between-subjects treatments in which we held constant across treatments the decision sets that participants progress through, but varied the communication mode. We call these two treatments the *Gen* treatment and the *Spec* treatment. We provide details below.

Both Treatments The waiting in line screen and the chatbot screens are the same in both treatments. First consider the waiting in line screen shown in Figure 3b. The participant observes the progress bar fill until the waiting time (t^{line_1} or t^{line_2}, depending on the channel) is over. The

[5] In addition to the main task, at the end of the experiment we elicited the participants' risk aversion (with respect to money) using an incentivized version of the Eckel-Grossman single lottery test (Eckel and Grossman 2002, 2008), which could earn participants up to an additional $2. Additional details and instructions are provided in EC.1.2.

Table 2 Decision Sets in Experiments 1, 2 and 3

Decision Set 1 (Varying Chatbot Success Rate)

	Live Agent Channel		Chatbot Channel				Expected time difference
	t^{line_1}	t^{agent}	p^{succ}	t^{bot}	t^{line_2}	t^{agent}	
Decision 1	20	20	0.25	20	20	20	10
Decision 2	20	20	0.3	20	20	20	8
Decision 3	20	20	0.35	20	20	20	6
Decision 4	20	20	0.4	20	20	20	4
Decision 5	20	20	0.45	20	20	20	2
Decision 6	20	20	0.5	20	20	20	0
Decision 7	20	20	0.55	20	20	20	-2
Decision 8	20	20	0.6	20	20	20	-4
Decision 9	20	20	0.65	20	20	20	-6
Decision 10	20	20	0.7	20	20	20	-8
Decision 11	20	20	0.75	20	20	20	-10

Decision Set 2 (Varying Chatbot Service Time)

	Live Agent Channel		Chatbot Channel				Expected time difference
	t^{line_1}	t^{agent}	p^{succ}	t^{bot}	t^{line_2}	t^{agent}	
Decision 1	20	20	0.5	30	20	20	10
Decision 2	20	20	0.5	28	20	20	8
Decision 3	20	20	0.5	26	20	20	6
Decision 4	20	20	0.5	24	20	20	4
Decision 5	20	20	0.5	22	20	20	2
Decision 6	20	20	0.5	20	20	20	0
Decision 7	20	20	0.5	18	20	20	-2
Decision 8	20	20	0.5	16	20	20	-4
Decision 9	20	20	0.5	14	20	20	-6
Decision 10	20	20	0.5	12	20	20	-8
Decision 11	20	20	0.5	10	20	20	-10

Decision Set 3 (Varying Line Duration After Chatbot Failure)

	Live Agent Channel		Chatbot Channel				Expected time difference
	t^{line_1}	t^{agent}	p^{succ}	t^{bot}	t^{line_2}	t^{agent}	
Decision 1	20	20	0.5	20	40	20	10
Decision 2	20	20	0.5	20	36	20	8
Decision 3	20	20	0.5	20	32	20	6
Decision 4	20	20	0.5	20	28	20	4
Decision 5	20	20	0.5	20	24	20	2
Decision 6	20	20	0.5	20	20	20	0
Decision 7	20	20	0.5	20	16	20	-2
Decision 8	20	20	0.5	20	12	20	-4
Decision 9	20	20	0.5	20	8	20	-6
Decision 10	20	20	0.5	20	4	20	-8
Decision 11	20	20	0.5	20	0	20	-10

Notes: The sequence of decision sets was chosen at random for each participant. Depending on the treatment, decision alternatives were either labeled as "Alternative A" and "Alternative B" (in *NeutGen* and *NeutSpec* treatments) or as "Chatbot" and "Live Agent" (in *Gen* and *Spec* treatments). Expected time difference was not shown to participants, but is added here for clarity.

screen for waiting in line does not differ by channel or treatment. (Recall, however, that the live agent channel always requires waiting in line, but the chatbot channel only requires waiting in line if the chatbot fails.)

The service experience differs by channel. The chatbot channel screens are shown in Figure 3c and are kept constant across both treatments. In the chatbot channel each participant begins service by seeing a message from the system: *To begin "service" press the **a** key on the keyboard.* After the participant presses ***a*** and waits for a prespecified number of seconds, the chatbot asks: *Hi, how can I help you?* The participant then sees the following system prompt: *To answer "service" press the **s** key on the keyboard.* The chatbot then takes a prespecified number of seconds until the next prompt is displayed. The interaction continues in this manner until service is completed. The prompts and the number of steps are the same for all participants. The chatbot speech bubble is gif-animated to emulate real-world chatbots.

Gen Treatment In the *Gen* treatment we keep the communication mode *gen*eric, i.e., do not vary it across channels. Specifically, the experience of interacting with the live server (including the text of the system prompts and the keys to be pressed) is exactly the same as that of interacting with the chatbot, with the sole difference being that the images and the animations show an interaction between a live agent and a customer. See the top row of Figure 3d for an illustration.

Spec Treatment In the *Spec* treatment communication mode is *spec*ific to the channel. In particular, in the chatbot channel the communication mode continues to be the same sequence of keystrokes as before. However, when interacting with the live agent, the participant is required to press and hold down a button for t^{agent} seconds.[6] The live agent screens are shown in the bottom row of Figure 3d. Note that the gif animations are the same across treatments – the only difference is whether the service process with the live agent requires the customer to read and submit messages or to press and hold down a button.

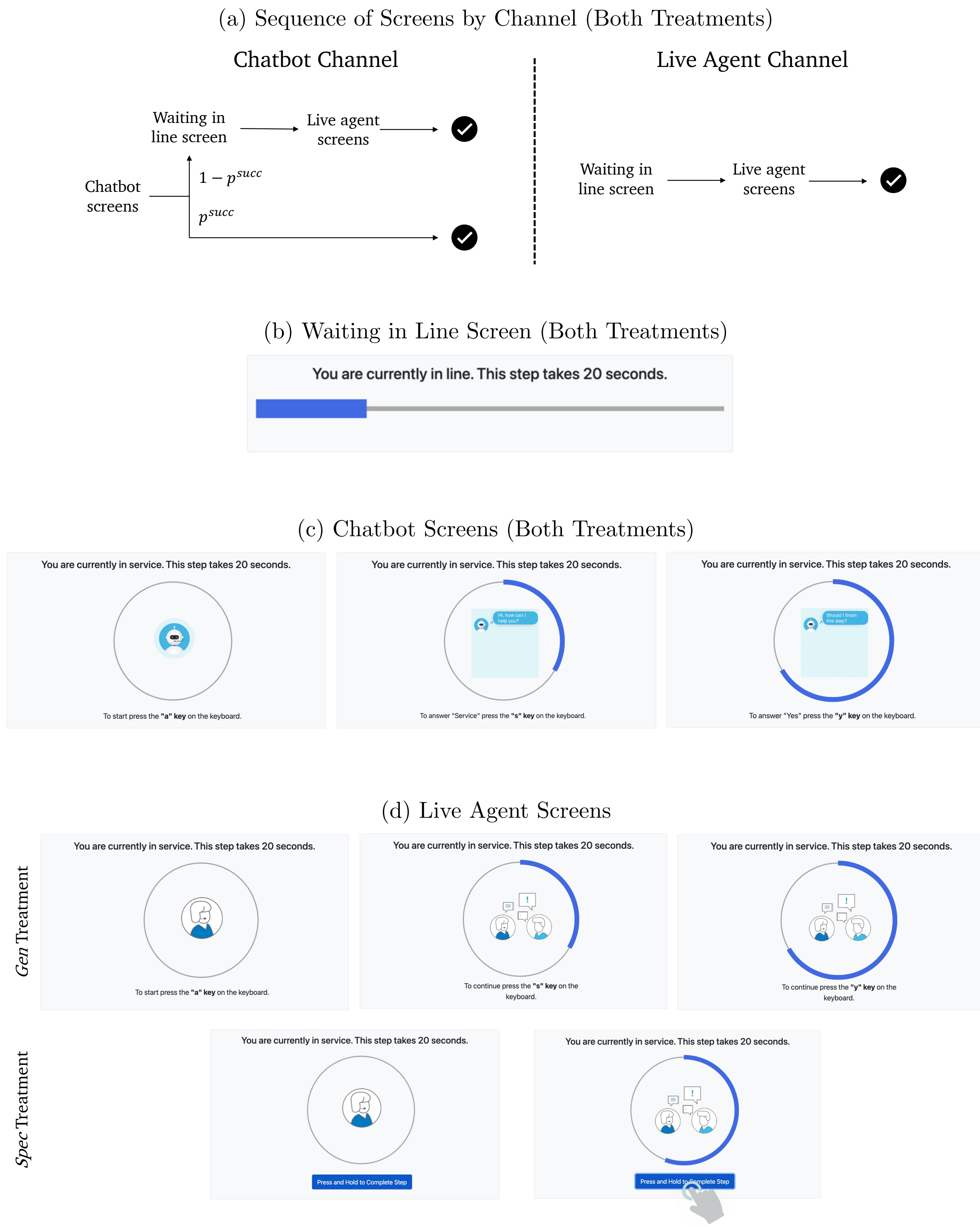
3.1.5. Hypotheses and Testing Approach Our first hypothesis is related to the response to chatbot performance, i.e., to increases in its success probability, to decreases in the chatbot service time, and to decreases in the time spent in line after chatbot failure.

H1 (Chatbot Performance): *Chatbot uptake increases in p^{succ} and decreases in t^{bot} and in t^{line_2}.*

In addition to responding to chatbot performance (reliability and speed), we expect the communication mode to matter. First, a message exchange between the server and the customer results in a service experience that is more fragmented relative to holding down a button. This may lower the willingness to engage with the chatbot. Additionally, the literature on robot and algorithm aversion suggests that customers may be more critical to automated (algorithmic) agents simply because the work is performed by a non-human (Dietvorst et al. 2015, and references therein). The distinctly robotic communication mode of the chatbot in the *Spec* treatment may amplify the

[6] If the participant releases the "press and hold down" button before service is completed, and then presses it again they resume from where they left off (as opposed to having to start from the beginning).

Figure 3 Screenshots of the Experiment



algorithmic nature of the chatbot. While the detailed investigation of mechanisms is postponed until Experiment 2 (§4), both of these arguments suggest that participants should be less inclined to choose chatbots when the communication mode differs between channels.

H2 (Communication Mode): *Chatbot uptake is lower when the communication mode is channel-specific (chatbot: keystrokes, live agent: holding) than when the communication mode is generic (chatbot: keystrokes, live agent: keystrokes).*

To test H1 and H2 we use random effects logistic regressions with the channel choice ($Chatbot_{ij}$) indicating whether the chatbot channel was chosen by participant i in decision j, where $p_{ij}^{succ}, t_{ij}^{bot}$ and $t_{ij}^{line\_2}$ denote the operational parameters of the chatbot channel faced by participant i in decision j, and $Spec_i$ denotes whether participant i was assigned to the *Spec* treatment group, and $Controls_{ij}$ denotes individual and time controls:

$$log\frac{P(Chatbot_{ij}=1)}{P(Chatbot_{ij}=0)} = \alpha_0 + \alpha_1 p_{ij}^{succ} + \alpha_2 t_{ij}^{bot} + \alpha_3 t_{ij}^{line\_2} + \alpha_4 Spec_i + Controls_{ij} + \epsilon_{ij} \tag{3.1}$$

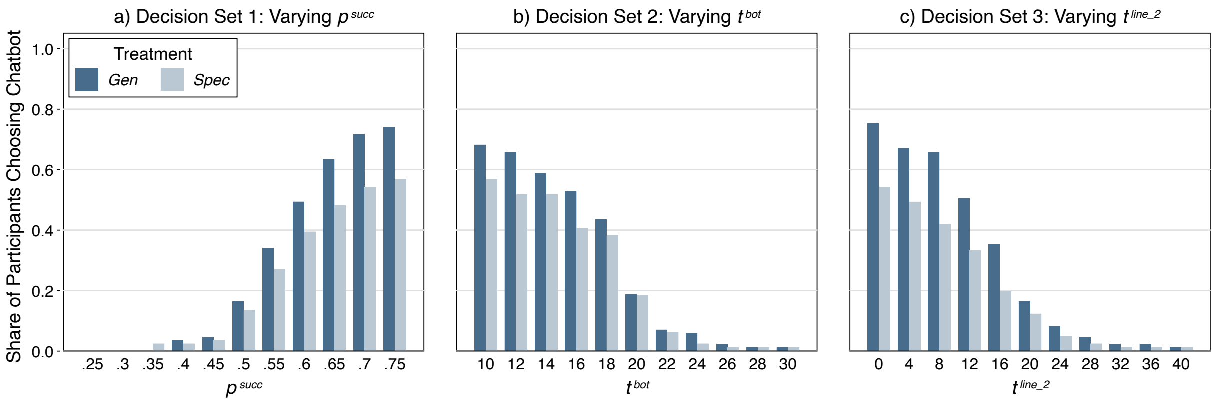
To test H1 we will examine coefficients α_1, α_2 and α_3. To test H2 we will examine α_4.

3.2. Results

Before presenting the results we comment briefly on the overall level of consistency in the collected data. We refer to a participant as "consistent" if, for each decision set in Table 2, they switch at most once from left (the live agent channel) to right (the chatbot channel). Multiple switching points within a decision set violate basic choice axioms (Charness et al. 2013). In our data, 166 out of 198 participants (83.84%) make choices that are consistent throughout the experiment (i.e., their chatbot uptake is always weakly increasing in p^{succ} and weakly decreasing in t^{bot}, and t^{line_2}). This compares favorably to the consistency numbers reported in the prior literature using our elicitation method (Holt and Laury 2002, Charness et al. 2013). In our analysis (§3.2.2) we will present the results for both the full set of subjects and for the consistent subjects only.

3.2.1. Descriptive Statistics Figure 4 shows the share of participants choosing the chatbot in each of the 33 decisions. Note first that very few people (between 0% and 10%) choose the chatbot when the parameters are such that the expected time in the chatbot channel is higher than the expected time in the live agent channel ($p^{succ} \leq 0.45$ in panel a, $t^{bot} \geq 22$ in panel b, $t^{line\_2} \geq 24$ in panel c). Second, even in decisions for which both channels have the same expected waiting time ($p^{succ} = 0.5$ in panel a, $t^{bot} = 20$ in panel b, $t^{line\_2} = 20$ in panel c) there is a preference for the live agent channel. Finally, a large group of participants (between 25% and 45%, depending on treatment and decision set) *never* chooses the chatbot channel. Importantly, across all three decision sets, chatbot uptake appears to be higher in the *Gen* treatment relative to the *Spec* treatment, providing some initial evidence that communication mode may affect channel preferences.

Figure 4 Chatbot Uptake in Experiment 1



Note. Figures show the fraction of participants choosing the chatbot in each of the 33 decisions (only participants with consistent choices are displayed).

3.2.2. Hypothesis Tests We next formally test Hypotheses 1 and 2 using logistic regression analysis, in which we account for the panel nature of the data (repeated measures), as well as control for demographic variables. Table 3 shows the estimates. The dependent variable in all four specifications is the channel choice ($Chatbot_{ij} = 1$ if the chatbot was chosen and $Chatbot_{ij} = 0$ otherwise). In columns (1) and (2) we use the data from all 198 participants who passed the screening, which includes both the consistent participants (i.e., participants who crossed over at most once from the live agent channel to the chatbot channel in each of the three decision sets) and the inconsistent ones. In columns (3) and (4) we use the data from all 166 participants who made consistent choices in all three decision sets. Further, in columns (2) and (4) we control for individual differences (see table notes for details).

The results confirm the patterns suggested by the descriptive statistics in §3.2.1. First, the coefficients for all three determinants of chatbot performance (p^{succ}, t^{bot} and t^{line_2}) are highly significant ($p \leq 0.01$). The size of the coefficients indicates a strong response, especially if we restrict our attention to the consistent subjects in columns (3) and (4). For example, the coefficients in column (4) suggest that a one percentage point increase in p^{succ} leads to a 0.49 percentage point increase in chatbot uptake, a one second increase in t^{bot} leads to a 3.19 percentage point decrease in chatbot uptake, and a one second increase in t^{line_2} leads to a 1.68 percentage point decrease in chatbot uptake. Note that the t^{bot} coefficient is approximately half the size of the t^{line_2} coefficient. This is intuitive given that on average half of the chatbot interactions involve waiting in line; hence, we would expect the average marginal response to one second of waiting in line to be half as strong as to one second of waiting with the chatbot. Second, channel-specific communication significantly reduces chatbot uptake (*Spec* treatment dummy coefficient is significant at $p < 0.05$

Table 3 Channel Preferences in Experiment 1

Dependent Variable:	(1) Chatbot Channel	(2) Chatbot Channel	(3) Chatbot Channel	(4) Chatbot Channel
p^{succ}	12.22***	12.22***	26.43***	26.45***
	(0.558)	(0.558)	(1.221)	(1.221)
t^{bot}	-0.356***	-0.356***	-0.564***	-0.565***
	(0.016)	(0.016)	(0.026)	(0.026)
t^{line_2}	-0.172***	-0.172***	-0.297***	-0.297***
	(0.008)	(0.008)	(0.014)	(0.014)
Spec Treatment	-0.886***	-0.655**	-1.480**	-1.174**
	(0.334)	(0.315)	(0.594)	(0.546)
Intercept	3.091***	3.206***	1.244	1.296
	(0.504)	(0.777)	(0.864)	(1.332)
Sample	All subjects	All subjects	Consistent subjects	Consistent subjects
Demographic Controls?	No	Yes	No	Yes
Observations	6534	6534	5478	5478
Subjects	198	198	166	166

Notes: Random effect logit regression coefficients are reported. Dependent variable is the chatbot channel choice (yes = 1). Baseline is *Gen* treatment. Specifications (1) and (2) include all subjects that passed the screening questions. Specifications (3) and (4) include only the subjects with consistent choices throughout the task (no more than 1 switching point from live agent channel to chatbot channel in each decision set). All specifications control for the decision set number (which serves as the time period variable in the panel data set). Specifications (2) and (4) control for the following demographic variables: age, gender, number of quiz errors and the Eckel-Grossman risk aversion measure (administered after the main task). *** $p < 0.01$ ** $p < 0.05$ * $p < 0.1$.

in all four specifications). The size of the treatment effect is also meaningful and ranges between -0.655 and -1.480 depending on the specification, corresponding to an average decrease of chatbot uptake between -6.01 and -8.33 percentage points. We summarize these results below.

Result 1 (Chatbot Performance): *H1 is supported. Chatbot uptake increases in* p^{succ} *and decreases in* t^{bot} *and in* t^{line_2}.

Result 2 (Communication Mode): *H2 is supported. Chatbot uptake is lower when the communication mode is channel-specific (chatbot: keystrokes, live agent: holding) than when the communication mode is generic (chatbot: keystrokes, live agent: keystrokes).*

3.2.3. Discussion Experiment 1 suggests that chatbot uptake goes up as p^{succ} increases and as t^{bot} and t^{line_2} decrease (H1). These behaviors align with intuition and with classic utility-maximization logic: more reliable and faster chatbots generate a lower disutility relative to chatbots that fail frequently and are slow. However, a significant portion of participants continue to choose the live agent channel even when the expected time in the chatbot channel is the same or lower than in the live agent channel, suggesting that the operational attributes of the channels alone do not explain the observed channel preferences. Further, communication mode matters: conditional on their operational performance, chatbots are chosen less frequently when they use a different communication mode than live agents (H2).

What can plausibly explain the observed deviation from the expected time minimization benchmark? For one, participants may choose the live agent despite having to spend more time in the system, because the live agent channel does not involve a handover from one server to a second one. The experience of being transferred, and the resulting discontinuity in the service experience, may be perceived as unpleasant. Conversely, the seamlessness of the interaction with the live agent may compensate for a longer duration. Another potential explanation is that decision-makers may have adverse beliefs against chatbot technology, caused by their past experiences with and preconceptions about chatbots, i.e., algorithm aversion (Dietvorst et al. 2015). Similarly, communication mode may matter not because users have a preference for one mode over the other, but because the differences in communication mode reinforce prior beliefs and biases about the chatbot technology. To be able to parse these alternative explanations we next present a second experiment, in which we remove the contextual details and examine choices in a more neutral setting, where any ex ante biases towards the chatbot technology (algorithm aversion) will be muted.

4. Experiment 2: Mechanisms

To better understand the drivers of customer channel choice, we replicate Experiment 1 using a neutral framing. As before, we administer two treatments; we label them *NeutGen* and *NeutSpec*. Similar to Experiment 1, these treatments vary the mode of communication. A total of 223 participants were recruited on Prolifitc. Participation was restricted to workers who did not participate in Experiment 1.

4.1. Treatments

NeutGen Treatment. In the *NeutGen* treatment we use neutral labels for the two channels ("Format 1" for live agent and "Format 2" for chatbot) and also keep the nature of the communication mode constant ("generic") for both channels. For the service stages in both channels, the service process is represented by the word "server" and the graphical animations are removed (See Figure EC.3 for a screenshot). This treatment isolates the effects of the operational parameters on channel choice.

NeutSpec Treatment. In the *NeutSpec* treatment we keep the neutral labeling and graphical representation of the channels, but make the communication mode specific to the channel. Similar to the *Spec* treatment, communication mode depends on the chosen channel. To complete service, participants need to hold down a button in the first (formerly referred to as live agent) channel, and use keystrokes to respond to messages in the first stage of the second (formerly referred to as the chatbot) channel. This is to emulate the continuous experience of receiving service from a live agent vs. the more fragmented experience of the message exchange with the chatbot.

4.2. Hypotheses

We first ask whether part of the aversion to engage with chatbots in Experiment 1 is driven by a negative response to being transferred from one server (a gatekeeper) to a second one, even when the (algorithmic) identity of the first server is not disclosed. The idea that *continuity*, i.e., receiving service from a single provider (rather than being transferred from one provider to a second one) may be a concern has been studied in the healthcare context (Kajaria-Montag et al. 2021, and references there) and is a known issue in call center management (Hathaway et al. 2022). Experiment 1 suggests that continuity may also matter for the channel choice between a chatbot and a live agent. To separately identify whether continuity – or the lack thereof in a gatekeeper system – affects channel choices, we remove the context and frame the decision problem as a neutral choice between two unnamed alternatives.

H3 (Gatekeeper Aversion): *Under neutral framing and generic communication mode (NeutGen treatment), chatbot uptake is below the expected time minimization benchmark.*

We next seek to understand the treatment effect in Experiment 1, i.e., the decrease in chatbot uptake when the communication mode is different between channels. Is this effect driven by the different experiences (keystroke vs. holding) in the two channels? In this case we should observe a robust treatment effect in the neutral setting of Experiment 2, even when the context and thus any potential bias against chatbots is removed. Or, is the treatment effect in Experiment 1 caused by an ex ante bias against chatbots, and that bias is reinforced by the more robotic communication mode? In that case we should not observe the treatment effect under the neutral framing of Experiment 2. Both alternatives are plausible, so we formulate two competing hypotheses for this effect.

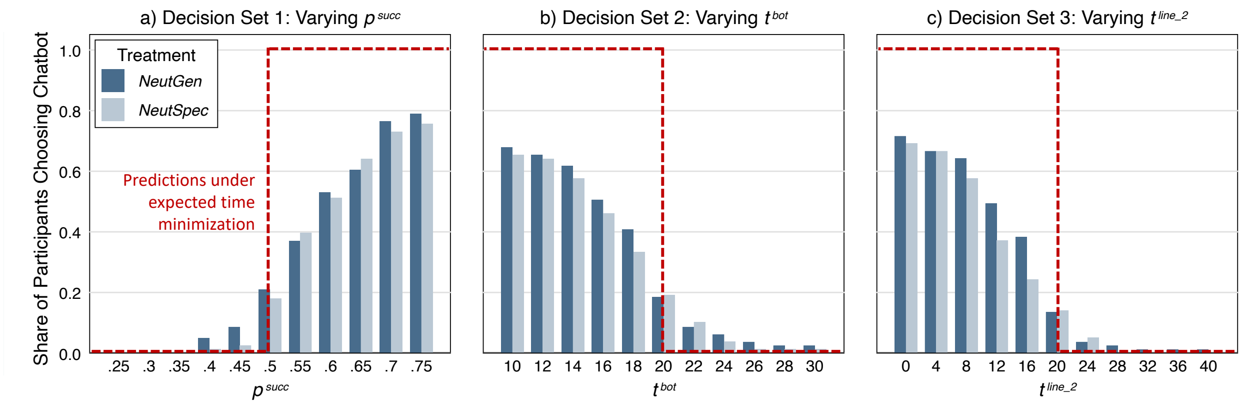
H4a (Communication Mode Mechanism): *Under neutral framing chatbot uptake is lower when the communication mode is channel-specific (chatbot: keystrokes, live agent: holding) than when communication mode is generic (chatbot: keystrokes, live agent: keystrokes).*

H4b (Algorithm Aversion Mechanism): *Under neutral framing chatbot uptake is not significantly different when the communication mode is channel-specific (chatbot: keystrokes, live agent: holding) than when the communication mode is generic (chatbot: keystrokes, live agent: keystrokes).*

4.3. Results

4.3.1. Descriptive Statistics Figure 5 shows the Experiment 2 channel choices by treatment, decision set, and decision. In this figure we also highlight the hypothetical decisions that would result from all participants simply minimizing expected time (dotted red lines). Note that these predictions follow directly from the last column of Table 2 in which we present the expected time

Figure 5 Chatbot Uptake in Experiment 2



Note. Figures show the fraction of participants choosing the chatbot in each of the 33 decisions (only participants with consistent choices are displayed).

difference for each decision. Several observations are in order. First, the increasing pattern in panel a) and the decreasing patterns in panels b) and c) mirror the pattern observed in Experiment 1: chatbot adoption continues to be higher for chatbots that have higher performance. Further, as before, the choices are well below the expected time minimization benchmark. Second, the treatment effect appears to be substantially weaker compared to Experiment 1. This suggests that the communication mode (exchanging text messages vs. holding) appears to matter not because of the inherent differences in the service experience provided, but mainly because it amplifies the preexisting biases against the chatbot channel. We next formally test Hypotheses 3 and 4.

Table 4 Participant Types in NeutGen Treatment

	Participant Type			
	(1) Never chooses chatbot channel (very gatekeeper-averse)	(2) Moderately gatekeeper-averse	(3) Expected time minimizer	(4) Other
Share of participants (%)	18.52	45.68	17.28	18.52

Notes: Table uses *NeutGen* treatment data. Type (1) participants choose the live agent format in all 33 decisions. Type (2) participants choose the live agent channel at least once for a decision where the chatbot channel yields a shorter expected time. Type (3) make strictly expected time-minimizing choices. Type (4) collects the remaining participants. Only consistent participants are included.

4.3.2. Hypothesis Tests To understand how gatekeeper aversion drives decisions (H3) we divide the participants into discrete types. The distribution of types is shown in Table 4. Here, we only use the *NeutGen* treatment data, in which any psychological biases or associations are muted and communication mode is held constant across channels; thus, any choice of the live agent in scenarios where the live agent has a shorter expected time (in line + in service) is driven by the aversion against being transferred. The data in Table 4 reveal that 18.52% of participants never choose the chatbot channel and that 45.68% of participants choose the chatbot channel less frequently than predicted under expected time minimization. Thus, the majority of participants (64.20%) fall into categories (1) and (2), both of which avoid the chatbot channel even in situations when it has a shorter expected time, relative to the live agent channel. Proportion tests confirm that the proportion of gatekeeper-averse participants (64.20%) is significantly higher than the proportion of expected time minimizers (17.28%) and higher than the proportion of all the remaining groups (35.80%), both at $p \leq 0.001$. These comparisons offer robust support for Hypothesis 3.[7]

Result 3 (Gatekeeper Aversion): *Hypothesis 3 is supported. Under neutral framing and generic communication mode (NeutGen treatment), chatbot uptake is below the expected time minimization benchmark.*

Next, we examine the treatment effect of varying the communication mode. Table 5 reports the estimate of random effects logit regressions with the chatbot channel choice as the dependent variable. As before, we observe robust increases of chatbot uptake as chatbot performance is improved, i.e, the chatbot becomes more reliable (p^{succ} increases) or faster (t^{bot} or t^{line_2} decrease). However, different from Experiment 1, the treatment effect of varying the communication mode is smaller (between -0.059 and -0.491) and is no longer statistically significant ($p-$values between 0.371 and 0.769). These estimates suggest that the decrease in chatbot uptake observed in Experiment 1 in the *Spec* treatment is driven not by the differences in experience (communication mode) but are driven by ex ante associations and beliefs, and these beliefs are simply reinforced by the communication differences.

Result 4 (Algorithm Aversion): *Hypothesis 4a is rejected. Hypothesis 4b is supported. Chatbot uptake is not affected by the communication mode in the neutral decision environment.*

4.4. Discussion

The results of Experiment 2 reveal two hurdles to chatbot adoption. First, there is significant gatekeeper aversion, i.e., the unwillingness to use the service channel that results in a potential handover to a second server. In particular, more than 60% of our participants avoid the chatbot channel even in situations where the chatbot would save them time (in expectation). A potential explanation of this result is risk-aversion. Indeed, the literature on risk preferences with respect to time (Leclerc et al. 1995, Abdellaoui and Kemel 2014, Flicker and Hannigan 2022) suggests that most people prefer certain over uncertain waits, even for relatively short durations. In our setting,

[7] For additional analysis of types, the reader is referred to Table A1, in which we report the distribution of types as well as the demographics for all three experiments and all treatments.

Table 5 Channel Preferences in Experiment 2

	(1)	(2)	(3)	(4)
Dependent Variable:	Chatbot Channel	Chatbot Channel	Chatbot Channel	Chatbot Channel
p^{succ}	13.48***	13.48***	31.49***	31.49***
	(0.581)	(0.581)	(1.458)	(1.455)
t^{bot}	-0.330***	-0.330***	-0.575***	-0.575***
	(0.014)	(0.014)	(0.028)	(0.027)
t^{line_2}	-0.184***	-0.184***	-0.356***	-0.356***
	(0.008)	(0.008)	(0.017)	(0.017)
NeutSpec Treatment	-0.100	-0.059	-0.431	-0.491
	(0.327)	(0.320)	(0.646)	(0.630)
Intercept	1.701***	1.311*	-0.831	-2.343
	(0.499)	(0.789)	(0.961)	(1.591)
Sample	All subjects	All subjects	Consistent subjects	Consistent subjects
Demographic Controls?	No	Yes	No	Yes
Observations	6468	6468	5247	5247
Subjects	196	196	159	159

Notes: Random effect logit regression coefficients are reported. Dependent variable is the chatbot channel choice (yes = 1). Baseline is *NeutGen* treatment. Specifications (1) and (2) include all subjects that passed the screening questions. Specifications (3) and (4) include only the subjects with consistent choices throughout the task (no more than 1 switching point in each decision set). All specifications control for the decision set number (which serves as the time period variable in the panel data set). Specifications (2) and (4) control for the following demographic variables: age, gender, number of quiz errors and the Eckel-Grossman risk aversion measure (administered after the main task). *** $p < 0.01$ ** $p < 0.05$ * $p < 0.1$.

however, the decision-maker does not know the outcome of the random draw until the first stage of service (the interaction with the chatbot) is completed. Thus, risk aversion is further compounded by the loss of time already budgeted for the transaction, as well as potential disappointment about chatbot failure. Indeed, almost a third of participants in our setting choose the live agent channel even when choosing the chatbot is the dominant strategy in *both* states of the world (see last row in Table 2 and the corresponding choice in the left-most column in panel c of Figure 5). In §6 we will further characterize gatekeeper aversion by examining different utility specifications that are consistent with the observed choices.

Second, once we remove the context and use neutral channel labels, the treatment effect disappears. It is worth noting that we used a minimal treatment manipulation (keystrokes vs. holding), which may have contributed to the relatively weak effect of communication. We did not examine the effects of speaking vs. typing, which may have further effects on channel uptake in practice. It is thus remarkable that even our relatively benign manipulation of communication mode appeared to reinforce prior beliefs and biases against chatbots (algorithm aversion), leading to a reduced chatbot uptake. More generally, this suggests that algorithm aversion is not a universal, but a more transient, situational phenomenon that depends on the specifics of AI design and appearance.[8]

The result that chatbots are avoided in part due to the users' preexisting biases and beliefs suggests that de-emphasizing the algorithmic nature of the chatbot may be an effective tool for increasing adoption. A related solution is to direct customer attention away from the agent providing service and towards the performance benefits (i.e., towards the expected time savings) offered by the chatbot. In the next section we will examine the effectiveness of this solution.

5. Experiment 3: Nudges

To examine potential ways to increase chatbot adoption we replicated the *Spec* treatment from Experiment 1 and the *NeutGen* treatment from Experiment 2 with the addition of a nudge that presents participants with the expected waiting time for the chatbot channel. A total of 208 participants were recruited on Prolific. Participation was restricted to workers who did not participate in Experiment 1 or Experiment 2.

5.1. Treatments

The nudge was shown to participants in all 33 decisions, and consisted of adding the total expected waiting time (line + service) in the chatbot channel. Figure 6 shows a screenshot of both the original decision screen (Experiment 1 and 2) in panel a) for Decision Set 1 and the decision screen for the same decision set with a nudge (Experiment 3) in panel b). The average waiting time information was added in all 33 decisions across the three decision sets in Experiment 3. To better understand how the nudge operates, we conducted two between-subject treatments.

Spec + Nudge Treatment This treatment is identical to the *Spec* treatment from Experiment 1, with the addition of the expected waiting time information.

NeutGen + Nudge Treatment This treatment is identical to the *NeutGen* treatment from Experiment 2, with the addition of the expected waiting time information.

5.2. Hypotheses

The theoretical support for using nudges in this setting is based on prior results in the experimental economics literature on choice elicitation under uncertainty. This literature shows that decision-makers do not always aggregate the outcomes and the attendant probabilities in ways consistent with the expected utility paradigm, but rather focus on some salient component or combination of components when making decisions (Arieli et al. 2011, Aimone et al. 2016a,b). Following this logic, we present participants with the expected waiting time, which we expect to shift their attention

[8] Notably, the effect of context (context = 1 in Experiment 1, context = 0 in Experiment 2) on chatbot uptake was not significant on average, or conditional on communication mode (See Table A3 for a full regression specification).

Figure 6 Screenshots of the Experiment

(a) Decision Screen Shot in Experiments 1 and 2

Scenario 1	20 + 20 = 40 sec. w. prob. 100%	○○	20 sec. w. prob. **25%**,	20 + 20 + 20 = 60 sec. w. prob. **75%**.
Scenario 2	20 + 20 = 40 sec. w. prob. 100%	○○	20 sec. w. prob. **30%**,	20 + 20 + 20 = 60 sec. w. prob. **70%**.
Scenario 3	20 + 20 = 40 sec. w. prob. 100%	○○	20 sec. w. prob. **35%**,	20 + 20 + 20 = 60 sec. w. prob. **65%**.
Scenario 4	20 + 20 = 40 sec. w. prob. 100%	○○	20 sec. w. prob. **40%**,	20 + 20 + 20 = 60 sec. w. prob. **60%**.
Scenario 5	20 + 20 = 40 sec. w. prob. 100%	○○	20 sec. w. prob. **45%**,	20 + 20 + 20 = 60 sec. w. prob. **55%**.
Scenario 6	20 + 20 = 40 sec. w. prob. 100%	○○	20 sec. w. prob. **50%**,	20 + 20 + 20 = 60 sec. w. prob. **50%**.
Scenario 7	20 + 20 = 40 sec. w. prob. 100%	○○	20 sec. w. prob. **55%**,	20 + 20 + 20 = 60 sec. w. prob. **45%**.
Scenario 8	20 + 20 = 40 sec. w. prob. 100%	○○	20 sec. w. prob. **60%**,	20 + 20 + 20 = 60 sec. w. prob. **40%**.
Scenario 9	20 + 20 = 40 sec. w. prob. 100%	○○	20 sec. w. prob. **65%**,	20 + 20 + 20 = 60 sec. w. prob. **35%**.
Scenario 10	20 + 20 = 40 sec. w. prob. 100%	○○	20 sec. w. prob. **70%**,	20 + 20 + 20 = 60 sec. w. prob. **30%**.
Scenario 11	20 + 20 = 40 sec. w. prob. 100%	○○	20 sec. w. prob. **75%**,	20 + 20 + 20 = 60 sec. w. prob. **25%**.

(b) Decision Screen Shot in Experiment 3

Scenario 1	20 + 20	**(40 sec. w. prob. 100%)**	○○	20 sec. w. prob. **25%**,	20 + 20 + 20 = 60 sec. w. prob. **75%**	**(50 sec. on average)**
Scenario 2	20 + 20	**(40 sec. w. prob. 100%)**	○○	20 sec. w. prob. **30%**,	20 + 20 + 20 = 60 sec. w. prob. **70%**	**(48 sec. on average)**
Scenario 3	20 + 20	**(40 sec. w. prob. 100%)**	○○	20 sec. w. prob. **35%**,	20 + 20 + 20 = 60 sec. w. prob. **65%**	**(46 sec. on average)**
Scenario 4	20 + 20	**(40 sec. w. prob. 100%)**	○○	20 sec. w. prob. **40%**,	20 + 20 + 20 = 60 sec. w. prob. **60%**	**(44 sec. on average)**
Scenario 5	20 + 20	**(40 sec. w. prob. 100%)**	○○	20 sec. w. prob. **45%**,	20 + 20 + 20 = 60 sec. w. prob. **55%**	**(42 sec. on average)**
Scenario 6	20 + 20	**(40 sec. w. prob. 100%)**	○○	20 sec. w. prob. **50%**,	20 + 20 + 20 = 60 sec. w. prob. **50%**	**(40 sec. on average)**
Scenario 7	20 + 20	**(40 sec. w. prob. 100%)**	○○	20 sec. w. prob. **55%**,	20 + 20 + 20 = 60 sec. w. prob. **45%**	**(38 sec. on average)**
Scenario 8	20 + 20	**(40 sec. w. prob. 100%)**	○○	20 sec. w. prob. **60%**,	20 + 20 + 20 = 60 sec. w. prob. **40%**	**(36 sec. on average)**
Scenario 9	20 + 20	**(40 sec. w. prob. 100%)**	○○	20 sec. w. prob. **65%**,	20 + 20 + 20 = 60 sec. w. prob. **35%**	**(34 sec. on average)**
Scenario 10	20 + 20	**(40 sec. w. prob. 100%)**	○○	20 sec. w. prob. **70%**,	20 + 20 + 20 = 60 sec. w. prob. **30%**	**(32 sec. on average)**
Scenario 11	20 + 20	**(40 sec. w. prob. 100%)**	○○	20 sec. w. prob. **75%**,	20 + 20 + 20 = 60 sec. w. prob. **25%**	**(30 sec. on average)**

away from the format differences between channels and encourage them to opt for the channel with the shorter expected waiting time.

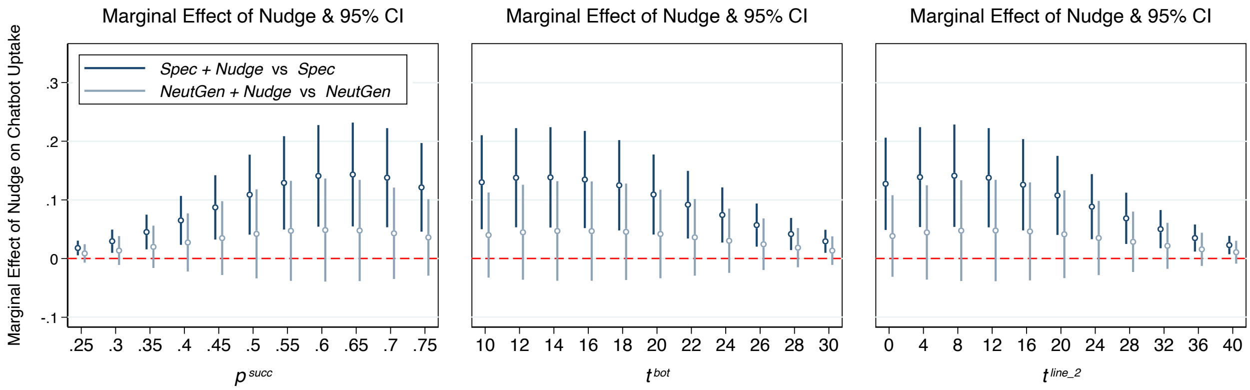
H5 (Nudge): *The expected waiting time nudge increases chatbot uptake.*

Note that H5 does not specify whether the effect of the nudge differs between the *Spec* or the *Neut-Gen* decision environment. However, in §5.4 we will provide some discussion about the mechanisms driving the changes in behavior.

5.3. Results

For brevity we omit the descriptive statistics of chatbot uptake and focus on the hypothesis tests. (The histograms of chatbot uptake are in Figures A1 and A2 in the Appendix.) As before, we use panel data logit regressions to test our hypotheses. The regression coefficients are reported in Table A3 in the Appendix; below we examine the relevant marginal effects.

Figure 7 The Effects of the Nudge on Chatbot Uptake



Note. Figure shows conditional marginal effects of the nudge for the parameters in each of the 33 decision scenarios. The regression specification used to produce these graphs is in column (4) of Table A3.

Figure 7 shows the marginal effects of the nudge, along with 95% confidence intervals. The figure suggests a robust increase of chatbot uptake in the *Spec* case, with effect sizes ranging between 2 and 14 percentage points. The average marginal effect of the nudge is significant at $p < 0.01$ in all four specifications in Table A3. In contrast, the effect size is at most 5 percentage points and not statistically significant if we compare chatbot uptake in the $NeutGen$ and $NeutGen + Nudge$ treatments ($p > 0.154$, see Table A3 bottom panel for parameter tests). Together, these results suggest that the nudge is effective primarily because it draws the participants' attention away from the format differences, and towards the operational efficiency and the resulting time savings offered by the chatbot. In contrast, the choices in a more neutral decision environment ($NeutGen$ treatment) are more stable and robust to the nudge intervention. Thus, the nudge helps mitigate algorithm aversion, but not gatekeeper aversion. These results are summarized below.

Result 5 (Nudge): *Hypothesis 5 is partially supported. Display of expected waiting times significantly increases chatbot uptake in the Spec treatment, but not in the NeutGen treatment.*

5.4. General discussion

Together, Experiments 1, 2 and 3 advance our understanding of the customer channel choice between an algorithmic chatbot and a live human agent. First, the observed choices are not consistent with expected time minimization; rather, customers are willing to wait longer (in expected value terms), in order to avoid being transferred from one provider to a second one. Second, chatbot uptake is further reduced when the chatbot communication patterns are different from the live agent. This behavior is not caused by the differences in service experience; rather, it appears to be caused by a more psychological aversion against interacting with an algorithmic agent, rooted in the users' prior experiences and beliefs. However, this aversion can be mitigated by directing customer attention to the operational benefits of the channel choice.

The result that decision-makers are averse to gatekeepers (in addition to being algorithm averse) may be caused by the feeling of disappointment, perceived lack of fairness, or by the inherent uncertainty of the chatbot relative to the live agent. While it is difficult to measure the experience of disappointment or fairness attitudes in our setting, the risk aversion measure with respect to money, elicited at the end of the experiment, is negatively correlated with the chatbot channel choice (Table A2 in the Appendix). However, the magnitude of the correlation ranges between 0.186 and 0.239 (all $p < 0.01$), suggesting that gatekeeper aversion is a richer phenomenon that includes factors beyond risk aversion. While we did not conduct further experiments to measure the role of specific psychological constructs driving gatekeeper aversion, the estimation of utility functions in §6 will provide some additional insight into these mechanisms.

While 18.65% of participants rejected chatbots regardless of their performance, 45.68% of participants were classified as moderately gatekeeper averse (see Table 4). This group of participants chose chatbots more frequently as chatbot performance went up, but they did so at a lower rate than expected time minimization would suggest. Improving the performance of the chatbot (increasing p^{succ} or decreasing t^{bot}) would increase their chatbot uptake. Such improvements, however, may require costly investment into the chatbot technology. Other, more innocuous interventions, such as a priority queue design, would shorten the wait t^{line_2} and allow chatbot users to jump the queue. Similar to priority queue design, the expected waiting time nudge is a simple remedy that can help make the chatbot more appealing, which would in turn lower the staffing costs for the service provider. In the next section we study the implications of these design decisions in more detail and evaluate their cost savings potential.

6. Implications for Service Design and Costs

We now turn to the implications of the behaviors documented in §3-§5 for the firm's service design and costs. We focus in particular on the cost implications of a "behavior-aware" service design that uses two operational innovations suggested by our results: a nudge that makes salient the expected waiting times, and a priority queue that prioritizes chatbot customers (who experienced chatbot failure) for faster access to a live agent. To evaluate potential cost savings we focus on the operational lever of right-sizing the staffing level – a key driver of controllable costs that motivated the development of chatbot technologies in the first place.

We proceed in three steps. We first formulate and estimate a random utility model of the customer choice between the live agent and chatbot channels. We then use the estimated utilities to predict aggregate customer demand (traffic) into each channel under a range of system parameters. We then perform counterfactual analyses by comparing the optimized staffing costs in an M/D/1 queuing regime under the "naïve" (pooled queue, no nudging) and the "behavior-aware" (priority queue, nudge) service designs, and evaluate the cost savings.

6.1. Random Utility Estimation

Equations (2.1) and (2.2) in §2 present a general version of a customer channel utility model that can accommodate different shapes of utility functions. In preliminary analysis (EC.2) we examined several common specifications, including power and exponential utility. A side-by-side comparison in Table EC.1 suggests that the linear specification achieves the best fit for our data (based on AIC); we will therefore use this specification for estimation. In contrast, concave-shaped utility functions often used to represent risk-aversion achieve a poorer fit with our data. Moreover, we tested several restricted versions of our linear model in Table EC.2 and again found that our full specification achieves a significantly better fit (based on AIC).

Our random utility model is based on the classic customer utility model from Allon and Kremer (2018) with two modifications. First, the waiting cost per unit time is allowed to vary by service provider: we denote by c^{agent} the cost per unit time of being served by the live agent and by c^{bot} the cost per unit time of being served by the chatbot. To account for the reduction in algorithm aversion from the nudge (studied in Experiment 3), c^{bot} is also allowed to differ by treatment. Second, we introduce a linear gatekeeper aversion factor, β, that is applied to the disutility incurred after the chatbot fails and the customer must be transferred to the live agent. This gatekeeper aversion factor captures several potential negative associations with being transferred to an additional stage of service, such as risk aversion, disappointment, perceived lack of fairness and the inconvenience of the transfer process. It can also capture the recency effect, i.e., the increased weighting of the component that occurred at the end of the service episode. In both channels we denote by c^{line} the cost per unit time spent waiting in line. Finally, the term ϵ_{ij}^{agent} (ϵ_{ij}^{bot}) is the idiosyncratic shock of choosing the live agent (chatbot) channel in decision j. Then, denoting by $\boldsymbol{\theta} = \{c^{line}, c^{agent}, c^{bot}, \beta\}$ the population vector of utility parameters, the respective utilities that customer i receives by choosing the live agent and chatbot channels in decision j are then given by:

$$U_{ij}^{agent}(\boldsymbol{\theta}) = r - c^{line} \cdot t_{ij}^{line\_1} - c^{agent} \cdot t_{ij}^{agent} + \epsilon_{ij}^{agent}, \tag{6.1}$$

$$U_{ij}^{bot}(\boldsymbol{\theta}) = r - c^{bot} \cdot t_{ij}^{bot} - (1 - p_{ij}^{succ}) \cdot \beta \cdot (c^{line} \cdot t_{ij}^{line\_2} - c^{agent} \cdot t_{ij}^{agent}) + \epsilon_{ij}^{bot}. \tag{6.2}$$

We estimate the model via the maximum likelihood approach. To focus on the contextualized version of the channel choice, we only use the data from the *Spec* and *Spec + Nudge* treatments. Since the customer receives the same reward (r) in either channel, it cannot be identified and is normalized to zero. The estimation results are in Table 6. We first note that while all costs are of a similar magnitude, customers experience the least discomfort while passively waiting for service ($c^{line} = 0.221$) and the most while interacting with the chatbot ($c^{bot}(no\ nudge) = 0.288, c^{bot}(nudge) = 0.248$). However, this cost is significantly lower in the nudge treatment ($p < .01$). The decrease in the unit cost of interacting with the chatbot increases the expected utility of the chatbot channel. This in turn increases the bot choice probability, and provides a plausible explanation for the positive effect of the nudge on chatbot adoption that we observed in Experiment 3. Finally, the gatekeeper aversion factor (β) is significantly greater than 1 ($p < .01$), indicating that customers experience additional disutility from moving to the live server after chatbot failure, above and beyond the extra time spent in line and with the server.

Table 6 Parameter Estimates for Random Utility Model

Parameter	Symbol	Estimate	Standard Error
Waiting Time Cost per Second	c^{line}	0.221***	(0.019)
Service Time Costs per Second			
Agent	c^{agent}	0.240***	(0.022)
Bot	$c^{bot}(no\ nudge)$	0.288***	(0.020)
Bot + Nudge	$c^{bot}(nudge)$	0.248***	(0.020)
Gatekeeper Aversion Factor	β	1.170***	(0.037)

Notes: Only includes subjects with consistent choices. *** $p < 0.01$.

6.2. Staffing Level and Cost Computation

We now turn to the characterization of the firm's live agent staffing levels and costs. We make the standard assumption that customers arrive according to a Poisson arrival process with demand intensity λ. The relative demand for each channel is formed according the logit choice probabilities $\rho^{agent}(\boldsymbol{\theta})$ and $\rho^{bot}(\boldsymbol{\theta})$, which can be derived from equations (6.1)-(6.2) and the estimates in Table 6. Consistent with the experimental setting, we model the live server as having deterministic service time. Despite these simplifications, characterizing the system entails solving for an equilibrium that takes into account the feedback effects arising from offering the waiting times (t^{line_1}, t^{line_2}), which lead to choice probabilities $(\rho^{agent}(\boldsymbol{\theta}), \rho^{bot}(\boldsymbol{\theta}))$, which in turn affect the waiting times t^{line_1} and t^{line_2}. The procedure to compute equilibrium staffing level (and costs) is as follows:

1. **Fix System Design:** Fixing system design parameters includes fixing t^{line_1}, t^{line_2}, the presence (or absence) of the nudge, as well as the remaining system parameters that affect the choice probabilities $(\rho^{agent}(\boldsymbol{\theta}), \rho^{bot}(\boldsymbol{\theta}))$.
2. **Calculate Channel Demand:** The demand for the chatbot is simply the portion of total demand intensity λ that is directed to the chatbot. The demand for the live agent is made up of two components: the customers who choose the live agent channel and the customers who choose the chatbot channel but experience chatbot failure and are redirected to the live agent:

$$\lambda^{bot} = \lambda \cdot \rho^{bot}(\boldsymbol{\theta}), \tag{6.3}$$

$$\lambda^{agent} = \lambda \cdot (\rho^{agent}(\boldsymbol{\theta}) + (1 - \rho^{agent}(\boldsymbol{\theta})) \cdot (1 - p^{succ})) \tag{6.4}$$

3. **Calculate Staffing Requirements and Costs:** Assuming that the chatbot development and training costs are sunk and that operating the chatbot is costless, we can focus on live agent staffing costs alone. To this end we first need to compute the average sojourn time in the live agent channel, $T^{agent}(\boldsymbol{\theta}) = \bar{t}^{line}(\boldsymbol{\theta}) + t^{agent}$, where $\bar{t}^{line}(\boldsymbol{\theta})$ is the average time that a customer spends in line waiting for the live agent, weighted by the proportion of the live agent demand coming from each channel. If we model the system as an $M/D/1$ queuing regime, then the live agent service rate μ required to deliver the announced waiting times, can be calculated as follows (derived from Tijms 2003, p. 59):

$$\mu(\lambda^{agent}(\boldsymbol{\theta}), T^{agent}(\boldsymbol{\theta})) = \frac{\lambda^{agent}(\boldsymbol{\theta}) + \sqrt{\lambda^{agent}(\boldsymbol{\theta})^2 + \frac{2\cdot\lambda^{agent}(\boldsymbol{\theta})}{T^{agent}(\boldsymbol{\theta})}}}{2}. \tag{6.5}$$

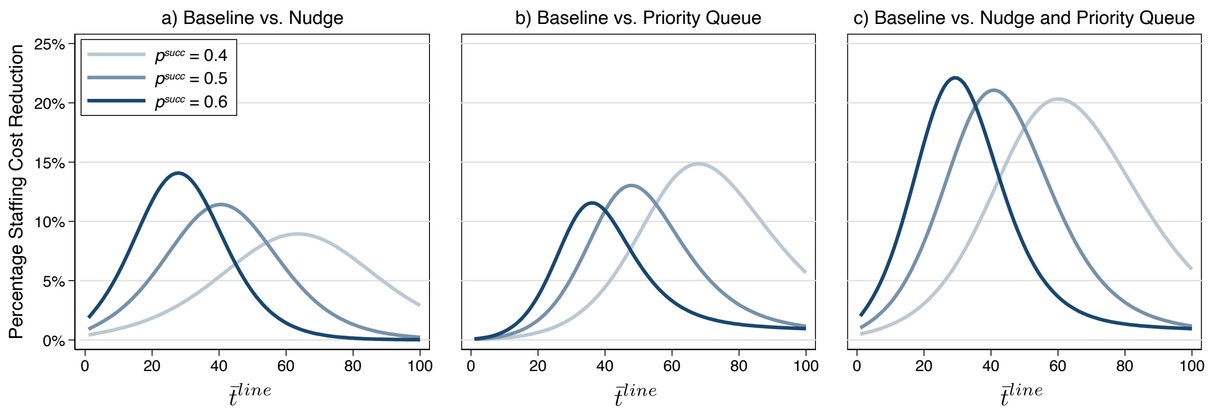
   If we assume that staffing costs increase linearly in the service rate μ (proxy for staffing level), we can use equation (6.5) to estimate staffing costs.

6.3. Naïve and Behavior-Aware Service Design: Cost Comparisons

We estimate staffing costs in four scenarios: a (baseline) naïve service design without the nudge and with a pooled queue (where $\bar{t}^{line} = t^{line\_1} = t^{line\_2}$), as well three scenarios that use either the nudge, a priority queue for chatbot users (where t^{line_2} is set lower than t^{line_1}), or a combination of the nudge and the priority queue. To compute costs for each scenario, we use the utility estimates from Table 6, the resultant demand for the live server, $\lambda^{agent}(\rho^{agent}(\boldsymbol{\theta}))$, and the corresponding staffing level given by (6.5). Further assumptions on system parameters are as follows. We set λ to 0.1, resulting in a system utilization between 75 and 80 % – a utilization level commonly used in queuing analysis of moderate-to-heavy traffic (see, for example, Hopp and Spearman 2011, for a discussion of common utilization ranges). We set t^{bot} to 20 and t^{agent} to 20. We vary the average waiting time to reach the live server, which we denote by $\bar{t}^{line}$, by increments of 1 from 1 to 90. We vary p^{succ} by setting it to 0.4, 0.5 and 0.6. Finally, we set the unit staffing cost to 1 (i.e., the staffing cost is simply given by equation (6.5)).

6.3.1. Cost Savings from Nudging Figure 8a shows the potential savings from implementing the nudge, relative to the benchmark case of a naïve service design that does not use the nudge. In all three scenarios ($p^{succ} = 0.4, p^{succ} = 0.5, p^{succ} = 0.6$) the cost savings peak at an interior value of $\bar{t}^{line}$. When $\bar{t}^{line}$ is sufficiently low, joining the live server queue for quick, guaranteed resolution is such an attractive option that nudging customers has little effect on demand, explaining the low cost savings. Likewise, when $\bar{t}^{line}$ is sufficiently high, the nudge has little effect as the bot is already perceived as an attractive option to avoid the long wait for the live server. It is only for intermediate values of $\bar{t}^{line}$ that the nudge has a significant enough effect on demand to substantially decrease staffing costs. Finally, because baseline staffing costs are higher under lower success probabilities, the peak savings amount is highest (at 14.1%) when $p^{succ} = 0.6$.

Figure 8 Staffing Cost Reductions Achieved by the Nudge and Prioritizing Chatbot Users



6.3.2. Cost Savings from Priority Queue Design To compare staffing costs with and without a priority queue for chatbot customers, we set $\bar{t}^{line}$ to be the same under each comparison between baseline and prioritization, but set t^{line_1} and t^{line_2} under prioritization such that the weighted average was equal to $\bar{t}^{line}$ and that $t^{line\_2} = 0.9 \cdot t^{line\_1}$.[9] Figure 8b shows that the impact on staffing costs is qualitatively similar to that of the nudge, with a maximum reduction of 14.9%. However, in this case the maximum is achieved when $p^{succ} = 0.4$, rather than at $p^{succ} = 0.6$ as it was in the case of the nudge. The reason for this contrast is that the nudge effectively lowers the disutility of interacting with the chatbot (by lowering c^{bot}), whereas queue priority lowers the disutility of experiencing a chatbot failure. Hence, this remedy is particularly effective for chatbots that exhibit frequent failure.

6.3.3. Combined Cost Savings Figure 8c shows the reduction in staffing costs under both remedies. Together, the two remedies reduce the disutility of the chatbot channel in both the first and (potential) second stage of service, resulting in cost savings across over a wide range of $\bar{t}^{line}$ values for all three p^{succ} values. Hence, together the two interventions could prove effective in reducing staffing costs for a variety of service settings.

7. Conclusions and Future Outlook

AI-powered chatbots are becoming an increasingly integral part of online customer service. To successfully leverage chatbot technology, firms need to understand both the relevant customer choice trade-offs, as well as their operational implications. In this paper we studied chatbot adoption by soliciting testimonies from chatbot users, by using these user stories to build a decision model of channel choice, and by testing (and refining) this model using online experiments.

[9] In other words, if the chatbot fails, the customer receives a 10 percentage point bump in the queue relative to a customer who immediately chose the live agent channel. Giving strict priority to chatbot users would result in factors even lower than our chosen factor of 0.9. We chose a fairly conservative priority factor and held it constant across all scenarios to make like comparisons.

In the first experiment we found that chatbot uptake increases with chatbot performance: chatbots were chosen more frequently when they required less waiting or had a higher request resolution rate. However, a large share of participants did not choose chatbots even when they offered significant time savings (in expectation) over live agents. Additionally, we found chatbot uptake to be further reduced when the chatbot used a distinctly robotic communication mode.

The second experiment revealed two mechanisms explaining these behaviors. First, it revealed an aversion to being served by a gatekeeper, i.e., a preference for a more seamless service experience and an aversion against a multi-stage experience that includes handovers from one server to a second one. Second, it revealed an ex ante bias towards chatbots that is not driven by the qualitative differences between the channels, but is more deeply ingrained in their experiences and beliefs about chatbot technology.

In the third experiment we examined the effects of a simple nudge that presents users with expected waiting times in each channel. We found that this nudge significantly increases chatbot uptake. The nudge works primarily by shifting the user's attention away from the qualitative differences between the channels and towards the decision-theoretic fundamentals of the choice.

Together, these results suggest that standard economic analysis of channel choice may oversimplify behaviors. Richer models of customer behavior that incorporate potential deviations from expected-time minimization can have nontrivial implications for service design and costs. We have examined the cost-reduction potential of two such design interventions: providing expected waiting time information as a nudge to divert some of the customer flow to the chatbots and a priority queue for customers whose requests were not resolved by the chatbot. Each of these interventions was found to reduce staffing costs by up to 14% (and together by up to 22%).

A question that we did not explore is *where* to set the level of chatbot investment (i.e., setting an "optimal" p^{succ}), and the underlying trade-off between the fixed cost of technology investment, and the variable costs of live agent staffing. Because chatbots often become better as they are trained on an existing customer base, there may also be interesting intertemporal aspects to this question. However, as natural language processing technologies reach maturity (as they have already done in some settings), improvements will need to be found in designing service systems in ways that make best use of existing technologies. To inform service design decisions, future research can benefit from integrating empirical and analytical tools, as we have done in this study. Customer choice data (both from the lab and field) can serve as input to build and refine customer choice models; queueing analysis can then use the choice model estimates to evaluate the implications of different service designs for system performance and costs.

Our results offer a glimpse into broader questions of technology-enabled multi-channel service design that are becoming increasingly relevant for service managers. As organizations become more customer-centric while also trying to keep costs low, efficient division of labor between AI-powered tools and human agents and procedural rules for their collaboration come to the fore. This includes enabling human workers to train AI (Mejia and Parker 2021), helping service workers be more productive with the help of AI-based tips and recommendations (Bastani et al. 2021, Balakrishnan et al. 2022), as well as developing guidelines for human-AI collaboration under varying system load (Snyder et al. 2022). The broader takeaway for service managers is to remain cognizant of the service design implications of new technologies, as new AI capabilities are developed.

References

Abdellaoui M, Kemel E (2014) Eliciting prospect theory when consequences are measured in time units: "time is not money". *Management Science* 60(7):1844–1859.

Aimone JA, Ball S, King-Casas B (2016a) It's not what you see but how you see it: Using eye-tracking to study the risky decision-making process. *Journal of Neuroscience, Psychology, and Economics* 9(3-4):137.

Aimone JA, Ball S, King-Casas B (2016b) 'nudging'risky decision-making: The causal influence of information order. *Economics Letters* 149:161–163.

Aksin OZ, Gencer B, Gunes ED (2020) How observed queue length and service times drive queue behavior in the lab.

Allon G, Kremer M (2018) Behavioral foundations of queueing systems. *The handbook of behavioral operations* 9325:325–366.

Arieli A, Ben-Ami Y, Rubinstein A (2011) Tracking decision makers under uncertainty. *American Economic Journal: Microeconomics* 3(4):68–76.

Balakrishnan M, Ferreira K, Tong J (2022) Improving human-algorithm collaboration: Causes and mitigation of over- and under-adherence.

Bastani H, Bastani O, Sinchaisri WP (2021) Learning best practices: Can machine learning improve human decision-making? *Academy of Management Proceedings*, volume 2021, 14006 (Academy of Management Briarcliff Manor, NY 10510).

Buell RW (2021) Last-place aversion in queues. *Management Science* 67(3):1430–1452.

Buell RW, Kim T, Tsay CJ (2017) Creating reciprocal value through operational transparency. *Management Science* 63(6):1673–1695.

Charness G, Gneezy U, Imas A (2013) Experimental methods: Eliciting risk preferences. *Journal of economic behavior & organization* 87:43–51.

Chen DL, Schonger M, Wickens C (2016) otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance* 9:88–97.

Das Gupta A, Karmarkar US, Roels G (2016) The design of experiential services with acclimation and memory decay: Optimal sequence and duration. *Management Science* 62(5):1278–1296.

Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144(1):114.

Eckel CC, Grossman PJ (2002) Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution Human Behav.* 23(4):281–295.

Eckel CC, Grossman PJ (2008) Men, women and risk aversion: Experimental evidence. *Handbook of experimental economics results* 1:1061–1073.

Flicker B, Hannigan C (2022) On people's utility over wait fundamentals and information.

Freeman M, Savva N, Scholtes S (2017) Gatekeepers at work: An empirical analysis of a maternity unit. *Management Science* 63(10):3147–3167.

Goot MJ, Hafkamp L, Dankfort Z (2020) Customer service chatbots: A qualitative interview study into the communication journey of customers. *International Workshop on Chatbot Research and Design*, 190–204 (Springer).

Goot MJ, Pilgrim T (2019) Exploring age differences in motivations for and acceptance of chatbot communication in a customer service context. *International Workshop on Chatbot Research and Design*, 173–186 (Springer).

Hathaway BA, Emadi SM, Deshpande V (2021) Don't call us, we'll call you: An empirical study of caller behavior under a callback option. *Management Science* 67(3):1508–1526.

Hathaway BA, Kagan E, Dada M (2022) The gatekeeper's dilemma:"when should i transfer this customer?". *Operations Research* .

Holt CA, Laury SK (2002) Risk aversion and incentive effects. *American economic review* 92(5):1644–1655.

Hopp WJ, Spearman ML (2011) *Factory physics* (Waveland Press).

Johannsen F, Leist S, Konadl D, Basche M (2018) Comparison of commercial chatbot solutions for supporting customer interaction .

Kajaria-Montag H, Freeman M, Scholtes S (2021) Continuity of care increases clinical productivity in primary care.

Kremer M, Debo L (2016) Inferring quality from wait time. *Management Science* 62(10):3023–3038.

Kumar P, Kalwani MU, Dada M (1997) The impact of waiting time guarantees on customers' waiting experiences. *Marketing science* 16(4):295–314.

Leclerc F, Schmitt BH, Dube L (1995) Waiting time and decision making: Is time like money? *Journal of consumer research* 22(1):110–119.

Lee YS, Seo YW, Siemsen E (2018) Running behavioral operations experiments using amazon's mechanical turk. *Production and Operations Management* 27(5):973–989.

Luo J, Valdés L, Linardi S (2022) Experienced and prospective wait in queues: A behavioral investigation. *Available at SSRN 4169028* .

Maister DH, et al. (1984) *The psychology of waiting lines* (Harvard Business School Boston).

Maynard N, Crabtree G (2020) Ai and automation in banking. Technical report, Juniper Research.

Mejia J, Parker C (2021) When systems fail: Remote worker accuracy and operational transparency.

Naor P (1969) The regulation of queue size by levying tolls. *Econometrica: journal of the Econometric Society* 15–24.

Paolacci G, Chandler J, Ipeirotis PG (2010) Running experiments on amazon mechanical turk. *Judgment and Decision making* 5(5):411–419.

Peer E, Rothschild DM, Evernden Z, Gordon A, Damer E (2021) Mturk, prolific or panels? choosing the right audience for online research. *SSRN Electronic Journal* .

Redelmeier DA, Kahneman D (1996) Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures. *pain* 66(1):3–8.

Schanke S, Burtch G, Ray G (2021) Estimating the impact of "humanizing" customer service chatbots. *Information Systems Research* 32(3):736–751.

Sheehan B, Jin HS, Gottlieb U (2020) Customer service chatbots: Anthropomorphism and adoption. *Journal of Business Research* 115:14–24.

Shumsky RA, Pinker EJ (2003) Gatekeepers and referrals in services. *Management Science* 49(7):839–856.

Snyder C, Keppler S, Leider S (2022) Algorithm reliance under pressure: The effect of customer load on service workers.

Tijms HC (2003) *A first course in stochastic models* (John Wiley and sons).

Ülkü S, Hydock C, Cui S (2020) Making the wait worthwhile: Experiments on the effect of queueing on consumption. *Management Science* 66(3):1149–1171.

Appendix A: Supporting Analysis: Individual Differences

Table A1 Participant Types

	Participant Type			
	(1) Never chooses chatbot channel (very gatekeeper-averse)	(2) Moderately gatekeeper-averse	(3) Expected time minimizer	(4) Other
Share of participants (%)				
Experiment 1				
Gen	14.12	52.94	8.24	24.71
Spec	29.63	39.51	11.11	19.75
Experiment 2				
NeutGen	17.28	45.68	18.52	18.52
NeutSpec	17.95	52.56	7.69	21.79
Experiment 3				
Spec + Nudge	20.99	38.27	23.46	17.28
NeutGen + Nudge	16.47	28.24	36.47	18.82
Exit survey responses/elicitation (averages by type)				
Female (0/1)	0.58	0.62	0.53	0.60
Age (years)	38.53	35.64	35.13	38.07
Risk Aversion w.r.t. money, gains domain, reverse coded (1-6, 1: most risk-averse)	3.29	3.70	4.38	4.47
Risk Aversion w.r.t. money, mixed domain, reverse coded (1-6, 1: most risk-averse)	2.92	3.00	3.93	3.95
Prefers chatbots over live agents outside of lab (0/1)	0.08	0.22	0.28	0.18
Has chosen chatbots over live agents outside of lab (0/1)	0.38	0.52	0.72	0.47

Notes: Type (1) participants always choose the live agent channel. Type (2) participants choose the live agent channel at least once for a decision where the chatbot channel yields a shorter expected time. Type (3) make strictly expected time minimizing choices. Type (4) participants are the remainder. Only consistent subjects are included. Note that the “Live Agent”/“Chatbot” labels are replaced by “Format A/B” in the *NeutGen* treatment).

Table A2 Correlations

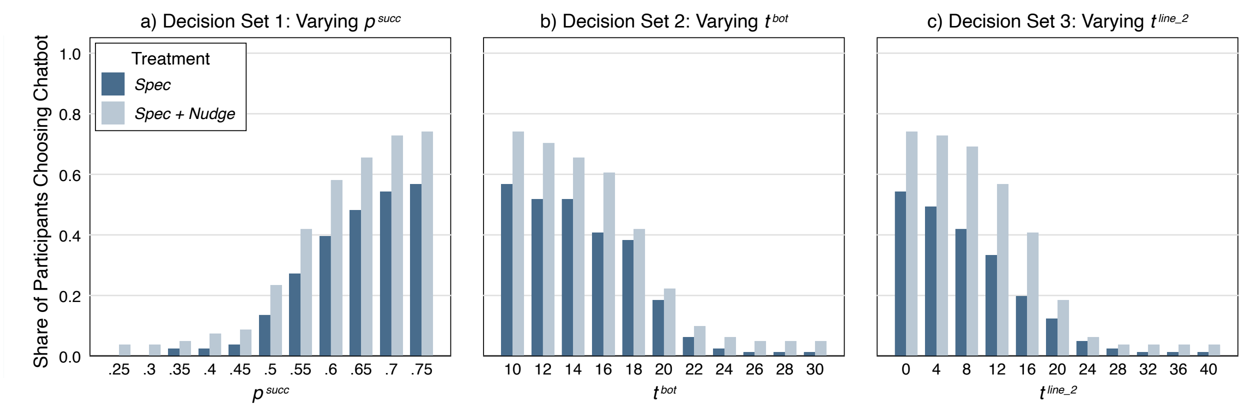
Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1) No. of chatbot choices in Decision Set 1	1								
(2) No. of chatbot choices in Decision Set 2	0.755***	1							
(3) No. of chatbot choices in Decision Set 3	0.715***	0.724***	1						
(4) Female (0/1)	-0.003	0.016	-0.035	1					
(5) Age (0/1)	-0.021	-0.051	-0.097**	-0.037	1				
(6) Risk Aversion w.r.t. money, gains domain, reverse coded	0.242***	0.221***	0.208***	-0.156***	-0.139***	1			
(7) Risk Aversion w.r.t. money, mixed domain, reverse coded	0.198***	0.222***	0.186***	-0.190***	-0.120***	0.530***	1		
(8) Prefers chatbots over live agents outside of lab	0.138***	0.143***	0.127***	-0.018	-0.122***	0.014	0.018	1	
(9) Has chosen chatbots over live agents outside of lab	0.117***	0.140***	0.136***	0.028	-0.056	0.010	0.000	0.344***	1

Notes: Pearson correlation coefficients and significance levels are reported for pooled data (Experiments 1 - 3), using only subjects with consistent choices. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Appendix B: Supporting Analysis for Experiment 3

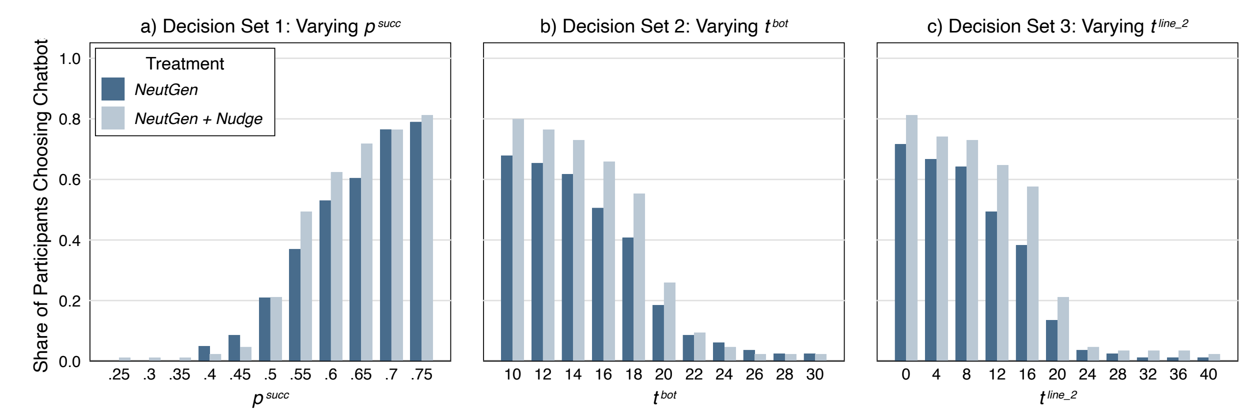
In this Section we present additional graphs and analyses related to the effects of nudges. Figures A1 and A2 show the chatbot choices by treatment. Table A3 reports the regression coefficients for a panel data logit regression of channel choice on the channel attributes and treatment dummies.

Figure A1 Experiment 3: Effect of Nudge on Spec



Note. Figure excludes participants with inconsistent choices (multiple switching decisions within the decision set).

Figure A2 Experiment 3: Effect of Nudge on NeutGen



Note. Figure excludes participants with inconsistent choices (multiple switching decisions within the decision set).

Table A3 Channel Preferences in Experiments 1, 2 and 3

Dependent Variable:	(1) Chatbot channel	(2) Chatbot channel	(3) Chatbot channel	(4) Chatbot channel
p^{succ}	13.42***	13.42***	30.30***	30.31***
	(0.340)	(0.340)	(0.807)	(0.806)
t^{bot}	-0.363***	-0.363***	-0.629***	-0.629***
	(0.009)	(0.009)	(0.017)	(0.017)
t^{line_2}	-0.188***	-0.188***	-0.350***	-0.350***
	(0.005)	(0.005)	(0.010)	(0.010)
Channel-specific communication mode	-0.937***	-0.785**	-1.689**	-1.461**
(*Spec* treatment)	(0.358)	(0.349)	(0.680)	(0.651)
Neutral context	-0.045	0.079	-0.014	0.321
(*NeutGen* treatment)	(0.354)	(0.343)	(0.676)	(0.647)
Channel-specific communication mode ×	0.834*	0.702	1.258	0.897
Neutral context (*NeutSpec* treatment)	(0.505)	(0.490)	(0.970)	(0.927)
Spec + Nudge treatment	1.240***	1.171***	2.132***	2.078***
	(0.364)	(0.353)	(0.692)	(0.660)
NeutGen + Nudge treatment	0.530	0.493	0.920	0.702
	(0.360)	(0.349)	(0.678)	(0.649)
Intercept	2.332***	2.247***	0.809	0.448
	(0.367)	(0.528)	(0.661)	(0.955)
Sample	All subjects	All subjects	Consistent subjects	Consistent subjects
Demographic Controls?	No	Yes	No	Yes
Observations	19303	19303	16203	16203
Subjects	583	583	491	491
		Parameter Tests:		
Spec treatment = *Spec + Nudge* treatment	$p = 0.001$	$p = 0.001$	$p = 0.002$	$p = 0.002$
NeutGen treatment = *NeutGen + Nudge* treatment	$p = 0.141$	$p = 0.158$	$p = 0.175$	$p = 0.280$

Notes: Random effect logit regression coefficients are reported. Dependent variable is the chatbot channel choice (yes = 1). Baseline is *Gen* treatment. Specifications (1) and (2) include all subjects that passed the screening questions. Specifications (3) and (4) include only the subjects with consistent choices throughout the task (no more than 1 switching point in each decision set). All specifications control for the decision set number (which serves as the time period variable in the panel data set). Specifications (2) and (4) control for the following demographic variables: age, gender, number of quiz errors and the Eckel-Grossman risk aversion measure (administered after the main task). *** $p < 0.01$ ** $p < 0.05$ * $p < 0.1$.

Electronic Companion

Appendix EC.1: Instructions and Experiment Details

Below we provide details and instructions for the user survey described in §1 of the manuscript, as well as the details and instructions for Experiments 1, 2 and 3.

EC.1.1. User Survey

The data were collected using the Prolific platform.[10] Participants were randomly assigned into either the live agent experience group or the chatbot experience group upon signing up for the study. Participants were first asked to recall an interaction with a live agent or a chatbot (depending on the treatment group) that had occurred in the 12 months prior to the study (Q1). Participants who reported not recalling such an interaction were directed to the exit survey. The remaining participants were asked to describe the interaction and its outcome (Q2) and were then asked 11 questions relating to the time spent waiting for the agent (chatbot) to become available, the information received prior to entering the interaction, the outcome of the interaction (i.e., whether their issue was resolved) and their overall satisfaction (Q3-Q13). Participants were compensated with a show up fee of $2. They also received an additional payment of $2 at the end of the study (i.e., a total of $4 for completing the entire study).

Below we reproduce the questions asked in the survey. Note that participants saw different questions depending on the treatment (live agent vs chatbot). Summary statistics of responses for multiple choice questions are provided after each questions.

Q1. *In the past 12 months, have you interacted with a customer support agent [chatbot]?*
Live Agent: No (10.00%) Yes (85.00%) Not Sure (5.00%).
Chatbot: No (16.33%) Yes (78.57%) Not Sure (5.10%).

[Note: If the answer to **Q1** is not "Yes", participant skips remaining questions and is redirected to the exit survey.]

[10] The Prolific platform was found to produce high quality data for individual decision-making tasks (Paolacci et al. 2010, Lee et al. 2018), and compared favorably in head-to-head comparisons with several other platforms (Peer et al. 2021). To further increase the quality of the data, we only used workers based in the United States (to avoid any country-specific effects) with an approval rating of at least 98%. The experiments were conducted in April and May 2022, on weekdays between 9am and 6pm Eastern Time.

Q2. *Please take a few minutes to describe, in as much detail as you can remember, a recent time when you had to contact customer support. Specifically, we are interested in a situation when you had to interact with a live (human) customer support agent [chatbot], either via phone or chat. Aiming for 1-2 sentences, please answer the following questions. What caused you to contact customer support? What type of service/issue did you need help with? What drove your decision to speak to a live agent [chatbot] (as opposed to, for example, looking at the FAQ)? How did the agent try to resolve your issue? How did your experience compare to your expectations? How did you feel about the decision to use live customer support?*

[Note: The six questions in **Q2** are split into three separate prompts with two questions each, with each prompt requiring a minimum of 50 characters for the response.]

Q3. *Were you given a choice between different options for customer support (e.g., a chatbot vs a live customer support agent)?*
Live Agent: No (52.94%) Yes (47.06%).
Chatbot: No (72.73%) Yes (27.27%).

Q4. *Were you given a time estimate for how much time it might take until you can use different support formats (e.g., waiting time for a chatbot vs waiting time for a live customer support* agent)?
Live Agent: No (69.41%) Yes (30.59%).
Chatbot: No (67.53%) Yes (32.47%).

Q5. *Of the two interaction types below* [Note: Figures EC.1/EC.2], *which one more closely resembles the customer support experience you described?*
Live Agent: Type A (49.41%) Type B (44.71%) Not Sure (5.88%).
Chatbot: Type A (46.75%) Type B (48.05%) Not Sure (5.19%).

Q6. *How long did you have to wait until the agent [chatbot] became available?*
Live Agent: Less than 1 minute (23.53%) 1-2 minutes (32.94 %) At least 3 minutes (43.53%).
Chatbot: Less than 1 minute (75.32%) 1-2 minutes (19.48 %) At least 3 minutes (5.19%).

Figure EC.1 Image Shown to Participants in the Live Agent Treatment

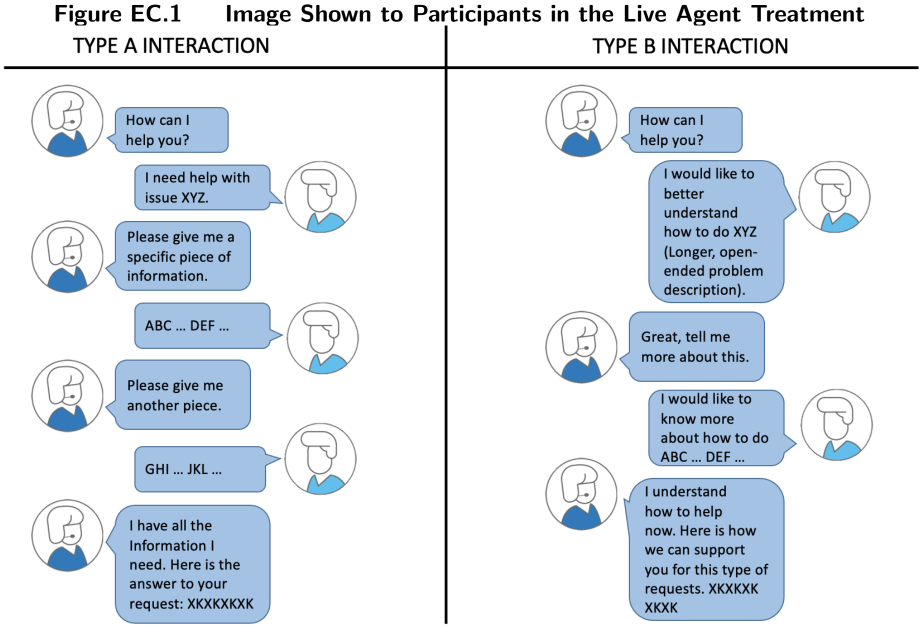
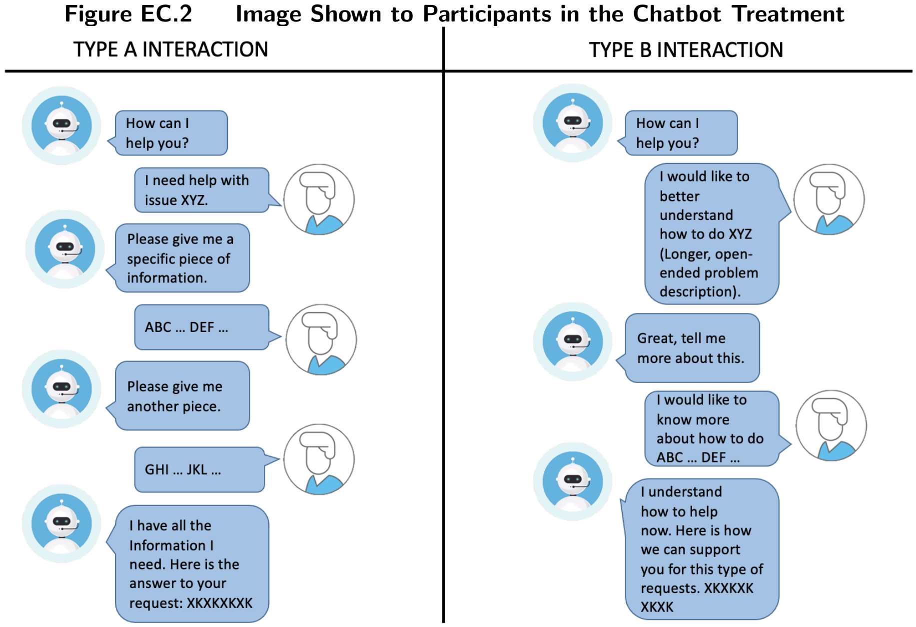


Figure EC.2 Image Shown to Participants in the Chatbot Treatment



Q7. *How long did you have to wait for the agent [chatbot] relative to your initial expectation?*

Live Agent: Less than expected (40.00%) Approximately as expected (47.06%) Longer than expected (12.94%).

Chatbot: Less than expected (29.87%) Approximately as expected (67.53%) Longer than expected (2.60%).

Q8. *How long did the interaction with the agent [chatbot] last?*

Live Agent: Less than 1 minute (0.00%) 1-2 minutes (8.24%) At least 3 minutes (91.76%).

Chatbot: Less than 1 minute (2.60%) 1-2 minutes (27.27%) At least 3 minutes (70.13%).

Q9. *How long did the interaction last relative to your initial expectation?*
Live Agent: Less than expected (22.35%) Approximately as expected (62.35%) Longer than expected (15.29%),
Chatbot: Less than expected (29.87%) Approximately as expected (55.84%) Longer than expected (14.29%).

Q10. *Did you have to share any information (e.g., order number, address, name, date of birth) with the agent [chatbot]?*
Live Agent: No, my details were not required (5.88%) No, the agent was able to retrieve most of the details from the system (23.53 %) Yes, I had to share those details (70.59%).
Chatbot: No, my details were not required (35.06%) No, the agent was able to retrieve most of the details from the system (23.38%) Yes, I had to share those details (41.56%).

Q11. *Approximately how many questions did the agent [chatbot] ask you during the interaction?*
Live Agent: 1-2 questions (28.24%) 3-4 questions (38.82%) 5-6 questions (17.65%) more than 6 questions (15.29%).
Chatbot: 1-2 questions (33.77%) 3-4 questions (48.05%) 5-6 questions (11.69%) more than 6 questions (6.49%).

Q12. *Was the agent [chatbot] able to resolve your request?*
Live Agent: No, I was transferred to another agent (5.88%) No, I had to call a different number to resolve the issue (1.18%) No, the issue remained unresolved (14.12%) Yes (78.82%).
Chatbot: No, I was transferred to another agent (23.38%) No, I had to call a different number to resolve the issue (29.87%) No, the issue remained unresolved (12.99%) Yes (33.77%).

Q13. *Overall, how satisfied were you with the customer support interaction? (1: very dissatisfied, 5: very satisfied)*
Live Agent: 3.01 on average.
Chatbot: 2.22 on average.

EC.1.2. Experiment 1: Details and Instructions

The data were collected on the Prolific platform.[11] The instructions are reproduced below.

Instructions

As part of this study you will experience several service "episodes". A service episode can be seeking customer support to resolve an issue with an online order you made, an online banking query, or passing through check-in at an airport. To represent the value from receiving a product or a service, at the end of each episode you will receive a fixed reward of 100 points.

Each service episode may include times that you spend in line and times that you will spend in service. We will use the word "server" to describe the service representative working on your request.

- *Time in line: Whenever you are "in line" you will spend time waiting for the server to become available.*
- *Time in service: Whenever you are "in service" you will also spend time while the server works on your request.*

In some episodes you will spend a fixed amount of time, for example, 15 seconds. In other episodes the amount of time you will spend will be uncertain. The specific amount of time you will spend depends on the service format in each episode. You will next experience two formats: The Live Agent Format and the Chatbot Format.

What will happen in the Live Agent Format? In this format you know the exact amount of time you will spend before receiving the reward. To be specific, you will spend 20 seconds in line and 20 seconds in service, i.e., 40 seconds total. After that, 100 points will be added to your account. Click "Next" to experience Format 1.

[...Subjects experience the Live Agent Format with 20 seconds in line and 20 seconds in service...]

[...Subjects answer comprehension questions about the details of the Live Agent Format...]

To summarize, in the Live Agent Format you will always spend 20 seconds in line, and then spend 20 seconds in service with the agent.

What will happen in the Chatbot Format? Unlike in the Live Agent Format, in the Chatbot Format you do not know the exact amount of time you will spend until you receive the reward. In this format

[11] We only used workers based in the United States (to avoid any country-specific effects) with an approval rating of at least 98%. The experiments were conducted in April and May 2022, on weekdays between 9am and 6pm Eastern Time. Participants were randomly assigned to a treatment upon signing up for the experiment.

you will first spend 20 with the chatbot. However, the chatbot is not always capable of resolving your request. If the chatbot fails to resolve your request you will need to interact with a live agent. Thus, different from Live Agent Format, in Format 2 there may be multiple service stages before service is completed.

You do not know ahead of time whether the chatbot will be able to resolve your request. However, you know that the total time (in line + in service) is either 20 or 60 seconds. To be more specific, in the Chatbot Format, there are two possible outcomes:

- *Chatbot succeeds: You spend 20 seconds with the chatbot. The chatbot succeeds in resolving your request and you receive 100 points, having spent 20 seconds total.*
- *Chatbot fails: You spend 20 seconds with the chatbot. The chatbot fails. To receive service from the live agent you need to wait in line until the live agent becomes available. This takes 20 seconds. After that, you spend another 20 seconds with the live agent. You then receive 100 points, having spent 20 + 20 + 20 = 60 seconds total.*

Click "Next" to experience the Chatbot Format.

[...Subjects experience the Chatbot Format with 20 seconds in service...]

The chatbot was successful. Total time spent: 20 seconds.

On the previous screen your service was completed by the chatbot. However, as mentioned previously, the chatbot may fail to resolve your request. In that case the live agent will be needed. On the next screens you will experience this scenario.

[...Subjects experience the Chatbot Format with 20 seconds in service, chatbot failure and spend additional 20 seconds in line, and 20 seconds with live agent...]

The live agent was successful. Total time spent: 60 seconds.

You are now ready to begin with the task. This task has two parts:

Part 1: You will be asked to make three sets of decisions; we call these the three "decision sets". In each decision set you will be presented several scenarios. For each scenario, you will be asked to choose whether you would rather experience the Live Agent Format or Chatbot Format. We will explain the details of each decision set on the next screens.

Part 2: Based on your choices in Part 1, you will experience three service encounters - one service encounter for each decision set. For each of the three service encounters you will wait the required amount of time to receive the reward.

Note that there are no "right" or "wrong" answers in this task - rather, we would like to know your personal preference for how to spend time.

[...Subjects complete all 33 decisions, then experience three randomly chosen decisions, then are directed to exit questionnaire...]

EC.1.3. Experiment 2: Details and Instructions

Instructions were kept similar to Experiment 1 with the difference that the "Live Agent Format" and the "Chatbot Format" were renamed to "Format 1" and "Format 2". Animations were removed. See below for a screen shot of the in service experience. See Figure EC.3 below for a sample screen shot of the in-service experience in both channels in the *NeutGen* treatment and in the (unlabeled) live agent channel in the *NeutSpec* treatment. No animations were displayed in either channel.

Figure EC.3 Experiment 2: Screen Shot

EC.1.4. Experiment 3: Details and Instructions

Instructions for *Spec + Nudge* treatment were unchanged relative to Experiment 1 and instructions for *NeutGen + Nudge* treatment were unchanged relative to Experiment 2. The sole difference were the decision screens shown in Figure 6.

Appendix EC.2: Alternative Utility Models

In this section we estimate alternative specifications of the structural model formulated in §6.1. Specifically, letting c denote a given waiting (service) time parameter and t denote the time spent waiting (in service) in a given stage, we estimated the following three specifications for the disutility the customer experiences by waiting (being served) in a given stage of the service delivery process: linear ($c \cdot t$), power (t^c), and exponential ($\frac{1-\exp(-c \cdot t)}{c}$). We present the estimates and the log likelihood of each specification in Table EC.1.

Table EC.1 Structural Estimates

Parameter	Linear	Power	Exponential
Waiting Time Parameter	0.221	0.629	0.103
Service Time Parameters			
Agent	0.240	0.536	0.136
Bot	0.288	0.685	0.080
Bot + Nudge	0.248	0.623	0.093
Gatekeeper Aversion Factor	1.170		
log likelihood	-2236.81	-2270.21	-2399.41

Only includes subjects with consistent choices.

We selected the linear specification based on best fit (as measured by AIC). Moreover, we estimated several restricted versions of the linear model above and present their estimates and log likelihoods in Table EC.2. The Linear 1 model is the full specification used in §6.1. In the Linear 2 model we set c^{agent} equal to $c^{bot}(no\ nudge)$. In the Linear 3 model we dropped the gatekeeper aversion factor (set β to 1). Finally, in the Linear 4 model we applied both restrictions. Based on AIC, the fully specified (Linear 1) model provides the best fit.

Table EC.2 Structural Estimates

Parameter	Linear 1	Linear 2	Linear 3	Linear 4
Waiting Time Cost per Second (c^{line})	0.221	0.198	0.256	0.177
Service Time Costs per Second				
Agent (c^{agent})	0.240	0.279	0.204	0.352
Bot ($c^{bot}(no\ nudge)$)	0.288	0.279	0.325	0.352
Bot + Nudge ($c^{bot}(nudge)$)	0.248	0.240	0.285	0.318
Gatekeeper Aversion Factor (β)	1.170	1.230	1.000	1.000
log likelihood	-2236.81	-2240.48	-2249.45	-2288.98

Only includes subjects with consistent choices.