AI Chatbots in Customer Service: Adoption Hurdles and Simple Remedies

Evgeny Kagan
Carey Business School, Johns Hopkins University

Maqbool Dada
Carey Business School, Johns Hopkins University

Brett Hathaway
Marriott School of Business, Brigham Young University

Despite recent advances in large language models, chatbot technology continues to face adoption hurdles. We report the results of six studies (two surveys and four experiments) that examine the choice between the chatbot channel and the live agent channel in customer service. We show that people respond positively to improvements in chatbot performance; however, the chatbot channel is utilized less frequently than expected time minimization would predict. This underutilization is caused by two separate behaviors: *gatekeeper aversion* (aversion to any service format that may involve a transfer to a different server) and *algorithm aversion* (aversion to an algorithmic service provider). Among the two factors, gatekeeper aversion is the more dominant one, accounting for 73% to 86% of behavior. Examining potential remedies, we find that chatbot uptake can be increased by making salient the expected time savings offered by the chatbot.

Key words: human-AI interfaces, technology management, experiments, service operations

1. Introduction

Chatbot technology has many uses, yet its commercial impact is perhaps most visible in online customer support. Recent technological advances 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 banking sector 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, Adam et al. 2021, Benke et al. 2022), 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 financial transactions. 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 channel choice.

Our investigation consists of two waves of a user survey and four incentivized experiments, summarized in Table 1. In our surveys (§2 and EC.1-EC.2) respondents retrospectively describe a recent customer service episode, either with a chatbot or with a live agent. Their testimonies suggest a key trade-off in channel choice: chatbots are faster to access but have a lower request resolution rate. This insight helps motivate the remainder of our studies, in which participants choose between two channels: one that is immediately available but frequently fails, and a second one that is slower to access but is more reliable.

In the first experiment (§3) we ask participants to join one of two waiting formats ("service channels"). The first format ("Channel A") represents the live agent and involves waiting in line to access service; the server resolves the request with probability 1, after which the participant receives a monetary reward. The second format ("Channel B") represents the chatbot and involves no waiting to access the first service stage; however the server fails with some known probability, requiring additional waiting in line and a second service stage. The two formats are described using neutral language, avoiding any contextual cues. Participants make a series of such choices, as we vary the operational parameters of Channel B, namely, the time spent in line and in service, and the probability of successful resolution in the first service stage.

[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.

Table 1 Study Overview

Questions & Objectives	Data	Key Results	Section
What are the key trade-offs in channel choice between chatbot and live agent?	Survey 1 + 2	Chatbots are faster to access but have lower request resolution rates.	§2, EC.1 - EC.2
Do customers minimize their expected waiting times when making channel choices?	Experiment 1	Customers will choose a channel that is slower in expectation, but does not involve a gatekeeper ("gatekeeper aversion").	§3
Are there further, non-operational drivers of channel preferences? ("algorithm aversion")	Experiment 2	Algorithm aversion may be present when chatbot experience is sufficiently different from live agent.	§4
Are channel choices different when the choice is between a real human and a chatbot?	Experiment 3	Algorithm aversion continues to be present with a real human (research assistant) playing the role of the server.	§5
What can managers do to increase chatbot uptake?	Experiment 4	Directing customer attention to expected time information helps increase chatbot uptake.	§6
What is the relative importance of operational vs. algorithmic factors in channel choice?	Experiments 1 - 4	Gatekeeper aversion (operational factors) accounts for 73% - 86% of observed behaviors.	§7

The results of the first experiment show that Channel B uptake is reduced as the first service stage duration gets longer, has a higher failure rate, and when the customer needs to wait longer for the second service stage after the first one fails. However, Channel B uptake remains far below what expected time minimization would predict. We term this behavior *"Gatekeeper aversion"* – the unwillingness to join a service channel with a server that may not be able to resolve the request resulting in a transfer to a different server, even when this channel saves time for the customer.

After examining the role of operational factors in channel choice, in §4 we turn our attention to non-operational factors, i.e., more qualitative features of chatbots and live agents. To examine chatbot uptake in a richer, more realistic setting, we vary the *content* of the wait. To do so, we present participants with a more continuous waiting experience of being served by a live agent, vis-a-vis a more fragmented waiting experience reflective of a back-and-forth message exchange with a chatbot. We first validate this manipulation using an attitudinal survey instrument. We then use this manipulation in our second experiment to examine chatbot channel uptake. We find that the manipulation significantly affects channel choices: chatbot uptake is further reduced (relative to Experiment 1) when the waiting experience differs between channels. However, this effect is only observed in the contextualized (and not in the neutrally-framed) treatment of the experiment, suggesting that it is caused not primarily by the experiential differences between channels, but by a psychological bias against algorithms.

To verify that the contextual cues presented in §4 are reflective of the channel choice in practice, we conduct a third experiment (§5), in which we employ real humans (research assistants, blind to the experimental hypotheses) who play the role of live agents and interact with participants using a live chat tool. We find that choices in this setting are largely consistent with our previous manipulations, further adding to the external validity of our study.

Lastly, to examine a way to increase chatbot uptake we conduct a fourth experiment (§6), in which participants see the expected waiting times for each channel, in addition to the waiting time distributions provided in the other treatments. Different from classic delay announcements (Jouini et al. 2011, Akşin et al. 2017, Ibrahim 2018), this does not add to the information set of the customer, but rather represents a more subtle nudge that directs the decision-maker's attention towards a specific aspect of the wait - the average time until completing service. We find that this nudge increases chatbot uptake, particularly in the contextualized version of the experiment.

Summarizing the results of our experiments, we identify two key barriers to chatbot adoption: a process-related adoption hurdle that is not tied to the algorithmic nature of the chatbot (gatekeeper aversion) and an attitudinal bias against chatbot technology (algorithm aversion). In §7 we use the data from all four experiments to evaluate the relative importance of gateekeper aversion and algorithm aversion and find that gatekeeper aversion accounts for 73% to 86% (depending on the study) of chatbot underutilization. This suggests that the operational aspects of the chatbot channel – which include not only the performance of the chatbot technology, but also the broader service ecosystem around the chatbot – are the key determinant of channel preferences.

Our contributions help support future research on algorithmic automation in service systems. First, we extend the conversation on human-AI interfaces, which has traditionally focused on AI adoption among workers (Dietvorst et al. 2015), to customer decisions in service channel selection. Our findings support the presence of algorithm aversion in this setting but suggest that it is overshadowed by operational factors unrelated to the human or algorithmic nature of the channel. Second, our study contributes to the growing experimental literature on queue joining (Kremer and Debo 2016, Flicker and Hannigan 2022), and queue switching and reneging behaviors (Akşin 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). Our contribution to this literature is that we study different waiting modalities, such as in-line and in-service waits. Our experimental approach presents a novel way to systematically evaluate responses to the content of the wait, while carefully controlling for its duration.

2. Surveys 1 and 2

We begin by reporting the results of two waves of a retrospective survey in which we asked online users to describe a recent customer service encounter with a live agent or with a chatbot. The complete lists of survey questions and detailed data analysis are in EC.1 and EC.2.

2.1. Methodology

We conducted the survey in two waves - an exploratory survey in June 2022 ($N = 198$) and a confirmatory one in July 2023 ($N = 202$).

2.1.1. Survey 1 (exploratory) The survey was conducted on the Prolific platform and consisted of two versions or "treatments", assigned at random. In one version ($N = 100$) 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. Respondents were US-based (55% female, average age: 33). All respondents received a show up payment of $2.00 and an additional payment of $2.00 at the end of the study.

2.1.2. Survey 2 (confirmatory) This survey was also conducted on the Prolific platform and also consisted of two between-subject "treatments" ($N = 106$ and $N = 96$), analogous to Survey 1. Respondents were US-based (49% female, average age: 38). They received a show up payment of $3.00 and an additional payment of $2.00 at the end of the study. The questions were similar to Survey 1 with three key differences. First, we asked a more specific, free-form question about what had driven the respondent's decision to choose the chatbot (a live agent) channel. Second, we included a more structured question regarding the respondents' attitudes towards algorithms. Third, we included an attitudinal test to validate one of our experimental manipulations (The discussion and analysis of this manipulation check are deferred to §4.1.2.)

2.2. Results

2.2.1. Survey 1 The survey responses reveal a wide range of problem types and settings in which both live agents and chatbots are used, from e-commerce and technology-related issues, to travel and transportation, food and groceries, and finance and banking. The descriptions of these interactions further suggest 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:

- *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 to 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)*

2.2.2. Survey 2 Based on the patterns of responses in Survey 1, we conducted a second survey, in which we added a more specific question related to the decision to choose a particular channel. Specifically, we separately asked the following question: *What drove your decision to speak to a chatbot?* in the Chatbot treatment. Analogously, we asked *What drove your decision to speak to a live agent?* in the Live Agent treatment. We then applied standard textual analysis tools (Krippendorff 2018) to identify keywords related to speed and performance, and compared their occurrence and frequency across the two channels.

Table 2 summarizes the results. The difference in speed-related keywords is quite large between the two treatments, with speed mentioned more frequently in the descriptions of chatbot interactions (15.22% vs. 43.01%, proportion test $p \ll 0.01$). In contrast, performance-related keywords were used more frequently in the descriptions of live agent interactions (38.04% vs. 23.66%, $p = 0.034$). To confirm that the textual analysis was representative of real channel experiences, 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 77.42% of respondents reporting in-line waits of "less than one minute", compared to 26.09% for live agents (proportion 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 41.94% of requests, relative to 86.96% for the live agents ($p \ll 0.01$).[2] Together, these comparisons suggest a speed-performance trade-off in channel

[2] 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. See Appendix EC.2 for detailed data.

Table 2 Survey Results: Speed-Performance Trade-off

Metric	Live Agent Treatment	Chatbot Treatment	$p-$value
Differences in time to access the server:			
% of respondents using speed-related key words (Q2)	15.22%	43.01%	$\ll 0.001$
% of respondents reporting no or minimal waiting (Q6)	26.09%	77.42%	$\ll 0.001$
Differences in performance:			
% of respondents using performance-related key words (Q2)	38.04%	23.66%	0.034
% of respondents reporting successful request resolution (Q12)	86.96%	41.94%	$\ll 0.001$

Q2 ("*What drove your decision to speak to a chatbot/live agent...*") responses were analyzed through keyword textual analysis using the *tm* package (Feinerer et al. 2008) in R. The following speed-related keywords were obtained from the WordNet synonym database: *"easy", "speed", "fast", "quick", "rapid", "efficient", "timely", "prompt", "immediate", "instant", "right away", "ASAP"*. The following performance-related keywords were obtained: *"capability", "competence", "proficiency", "skill", "accurate", "understand", "complexity", "ability", "difficult", "challenging", "hard", "tough", "resolve", "help", "succeed", "solve"*. The keywords were stemmed to their root forms using the Porter algorithm (Porter 1980). Statistical comparisons are based on non-parametric tests of proportions.

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. Our subsequent studies in §3-§6 will delve deeper into this trade-off.

3. Experiment 1: Gatekeeper Aversion

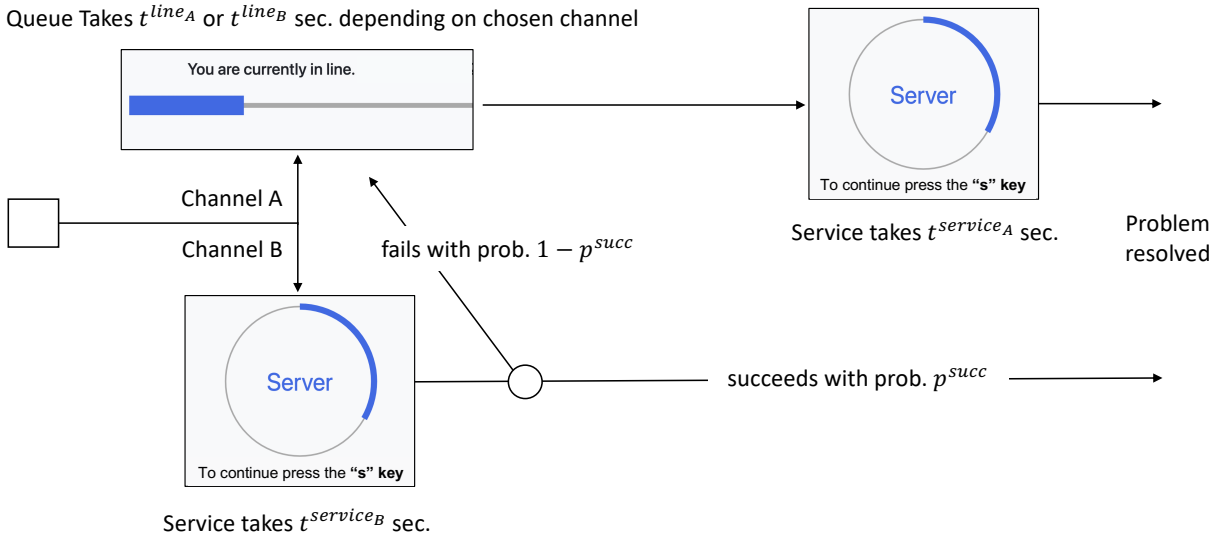
In Experiment 1 we focus on identifying *gatekeeper aversion* – an aversion to a service format where the first server may not be able to resolve the request, prompting the customer to restart service with a different server. To do so we will examine channel choices in a setting with a neutral framing, asking participants to choose among two different formats for how to spend time: one with a gatekeeper and one without.

3.1. Methodology

3.1.1. Channel choice We begin by describing the fundamentals of the channel choice. Note that in Experiment 1 the channel choice is framed neutrally; in particular, the live agent channel is referred to as "Channel A" and the chatbot channel as "Channel B". The instructions do not mention live agents or chatbots, but rather refer to them as "servers".

The channel choice is represented in Figure 1. Consider first Channel A, which represents the live agent channel. To access the server, the customer first needs to wait in line. The wait in line takes t^{line_A} seconds, and the wait in service takes $t^{service_A}$ seconds. The server always succeeds in resolving the request and the customer exits the system. Next consider Channel B. There is no line, so that the customer immediately proceeds to interacting with the chatbot and spends $t^{service_B}$ 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_B} seconds and then in

Figure 1 Channel Choice



service with a live agent agent for $t^{service_A}$ seconds. After that, the customer exits the system. To focus on the main decision trade-offs we assume that $p^{succ}, t^{service_A}, t^{service_B}, t^{line_A}$ and t^{line_B} are constant parameters known to the customer.

3.1.2. Experiment protocol Figure 2 presents the experiment protocol. After being randomly assigned to a treatment, participants first experience a demo of both channels. The Channel A demo begins with a wait in line (screen shot in upper left of Figure 1). While waiting in line, participants see a horizontal progress bar fill from left to right. After that, participants proceed to service, which involves a circular progress bar (to distinguish it from waiting in line). During service, participants are prompted to click on pre-specified keys for the progress bar to advance. They receive three such prompts, equally spaced throughout the service time. The Channel B demo is analogous and includes both the successful and failed chatbot resolution scenarios.

During the demo, all stages last 20 seconds. After the demo, participants make a total of 33 decisions, subdivided into three decision sets of 11 decisions with varying parameters. The sequence of decision sets was randomized to control for any order effects. Each of the 33 decisions is a binary choice between Channel A and Channel B. The parameters, p^{succ}, t^{line_B} and $t^{service_B}$ vary for each of the 33 decisions in ways that will be discussed next.

3.1.3. Elicitation Table 3 presents the parameters used in the experiment. 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 steps of 0.05 from 0.25 to 0.75 percent, while all

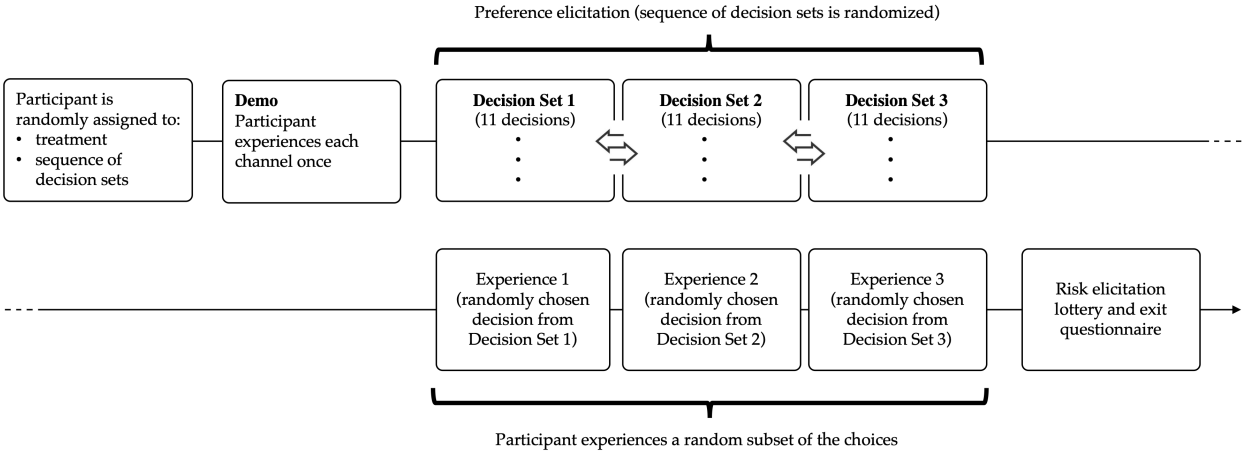
other parameters were held constant. In Decision Set 2 we varied the service time of the chatbot ($t^{service_B}$) in steps of 2 from 30 to 10 seconds. In Decision Set 3 we varied the time spent in line if the chatbot fails (t^{line_B}) in steps of 4 from 40 to 0 seconds. Across all three decision sets, 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).

3.1.4. Incentives After 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, resulting in each participant experiencing three of their 33 choices prior to receiving their (fixed) dollar payment and exiting the experiment. Thus, participants were incentivized to report their true preferences. 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.[3]

3.1.5. Theory and Hypotheses Research on decision-making under uncertainty in the time domain (as opposed to money) is quite limited; however, the extant work indicates minimal risk-aversion for the short time durations considered in our study (Leclerc et al. 1995), with some research suggesting close to risk-neutral behavior (Kroll and Vogt 2008, Festjens et al. 2015). We therefore use expected time minimization as our theoretical benchmark against which we compare behaviors. In other words, we hypothesize that participants will choose Channel B (the gatekeeper channel) when it offers a lower expected time than that of Channel A.

[3] 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.

Figure 2 Experiment Protocol



H1 (No Gatekeeper Aversion): *Participants choose the channel that minimizes their total expected time in the system.*

Expected time minimization would predict that participants choose Channel A in the first five decisions within each decision set, choose Channel B in the last five decisions, and are indifferent in the middle decision (see Table 3). Rejecting H1 would require that the count of Channel A choices

Table 3 Experimental Decisions

Decision Set 1: Varying Gatekeeper Success Rate (p^{succ})

	Channel A		Channel B				Expected time minimizing choice
	t^{line_A}	$t^{service_A}$	p^{succ}	$t^{service_B}$	t^{line_B}	$t^{service_A}$	
Decision 1	20	20	0.25	20	20	20	Channel A
Decision 2	20	20	0.3	20	20	20	Channel A
Decision 3	20	20	0.35	20	20	20	Channel A
Decision 4	20	20	0.4	20	20	20	Channel A
Decision 5	20	20	0.45	20	20	20	Channel A
Decision 6	20	20	0.5	20	20	20	Indifferent
Decision 7	20	20	0.55	20	20	20	Channel B
Decision 8	20	20	0.6	20	20	20	Channel B
Decision 9	20	20	0.65	20	20	20	Channel B
Decision 10	20	20	0.7	20	20	20	Channel B
Decision 11	20	20	0.75	20	20	20	Channel B

Decision Set 2: Varying Gatekeeper Service Time ($t^{service_B}$)

	Channel A		Channel B				Expected time minimizing choice
	t^{line_A}	$t^{service_A}$	p^{succ}	$t^{service_B}$	t^{line_B}	$t^{service_A}$	
Decision 1	20	20	0.5	30	20	20	Channel A
Decision 2	20	20	0.5	28	20	20	Channel A
Decision 3	20	20	0.5	26	20	20	Channel A
Decision 4	20	20	0.5	24	20	20	Channel A
Decision 5	20	20	0.5	22	20	20	Channel A
Decision 6	20	20	0.5	20	20	20	Indifferent
Decision 7	20	20	0.5	18	20	20	Channel B
Decision 8	20	20	0.5	16	20	20	Channel B
Decision 9	20	20	0.5	14	20	20	Channel B
Decision 10	20	20	0.5	12	20	20	Channel B
Decision 11	20	20	0.5	10	20	20	Channel B

Decision Set 3: Varying Line Duration After Gatekeeper Failure (t^{line_B})

	Channel A		Channel B				Expected time minimizing choice
	t^{line_A}	$t^{service_A}$	p^{succ}	$t^{service_B}$	t^{line_B}	$t^{service_A}$	
Decision 1	20	20	0.5	20	40	20	Channel A
Decision 2	20	20	0.5	20	36	20	Channel A
Decision 3	20	20	0.5	20	32	20	Channel A
Decision 4	20	20	0.5	20	28	20	Channel A
Decision 5	20	20	0.5	20	24	20	Channel A
Decision 6	20	20	0.5	20	20	20	Indifferent
Decision 7	20	20	0.5	20	16	20	Channel B
Decision 8	20	20	0.5	20	12	20	Channel B
Decision 9	20	20	0.5	20	8	20	Channel B
Decision 10	20	20	0.5	20	4	20	Channel B
Decision 11	20	20	0.5	20	0	20	Channel B

Notes: The sequence of decision sets was chosen at random for each participant. All time parameters ($t^{line_A}, t^{service_A}, t^{line_B}, t^{service_B}$) are in seconds.

is significantly different from 5.5 (assuming randomly breaking ties for the sixth decision in which both channels yield the same expected time). We will use one-sample $t-$tests to test H1.

3.1.6. Treatments and Participants Experiment 1 consisted of one treatment, which we will refer to as *BASELINE*. A total of 107 participants were recruited on Prolific; 97 of them passed the screening questions and attention checks and were admitted to the experiment. Each Prolific worker was restricted to participating in one session only. US-based workers with approval rating of at least 98% were recruited. All experiments were programmed in oTree (Chen et al. 2016).

3.2. Results

Before presenting our 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 3, they switch at most once from left (Channel A) to right (Channel B). Multiple switching points within a decision set violate basic choice axioms (Charness et al. 2013). In our data, 81 out of 97 participants (83.5%) are consistent (i.e., their Channel B uptake is always weakly increasing in p^{succ} and weakly decreasing in $t^{service_B}$ and t^{line_B}). This compares favorably to the consistency numbers reported in the prior literature using our elicitation method (Holt and Laury 2002, Charness et al. 2013), suggesting high levels of attention and engagement with the experimental stimuli.

3.2.1. Descriptive Statistics Figure 3 shows the share of participants choosing Channel B in each of the 33 decisions. As one may expect, the proportion of participants choosing Channel B increases with p^{succ} and decreases with $t^{service_B}$ and t^{line_B}. Consistent with expected time minimization, very few participants (between 0% and 10%) choose Channel B when the parameters are such that the expected time in Channel B is higher than the expected time in Channel A ($p^{succ} \leq 0.45$ in panel a, $t^{service_B} \geq 22$ in panel b, $t^{line_B} \geq 24$ in panel c). However, for decisions where both channels have the same expected waiting time ($p^{succ} = 0.5$ in panel a, $t^{service_B} = 20$ in panel b, $t^{line_B} = 20$ in panel c) there is a preference for Channel A – over 75% of participants

Figure 3 Channel B uptake in Experiment 1

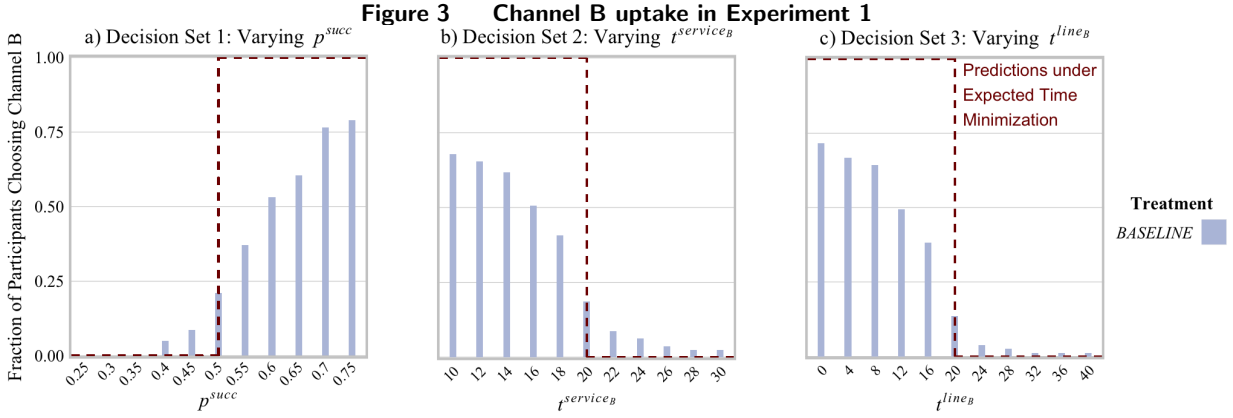


Table 4 Participant Types in Experiment 1

	Participant Type				
	(1) Never chooses Channel B	(2) Underutilizes Channel B	(3) Expected time minimizer	(4) Other	Total
Share of participants	17.28%	45.68%	18.52%	18.52%	100%
# Channel A choices per decision set	11	8.41	5.71	4.98	7.72
# Channel B choices per decision set	0	2.59	5.29	6.02	3.28

Type (1) participants choose the live agent format in all 33 decisions. Type (2) participants choose Channel A at least once for a decision where Channel B yields a shorter expected time and minimize expected waiting time in the remaining. Type (3) make strictly expected time-minimizing choices (either five or six Channel B choices in each decision set). Type (4) collects the remaining participants. Only consistent subjects are included.

choose Channel A in those decisions. Finally, a large group of participants (between 24% and 30%, depending on decision set) chooses Channel A for all 11 decisions.

To better understand channel choices we next divide participants into discrete types. The prevalence of each type, and a summary of their decisions are shown in Table 4. The data 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 in at least one decision set and are expected-time minimizers in the remaining decisions. The third group minimizes expected wait across all three decision sets and the fourth group collects remaining participants, with each group containing fewer than 20% of participants. Thus, the majority of participants (64.20%) fall into categories (1) and (2), both of which avoid the Channel B (gatekeeper channel) even in situations when it has a shorter expected time relative to Channel A.

3.2.2. Hypothesis Test We next test H1, i.e., examine whether channel choices are consistent with expected time minimization. To do so we use $t-$tests to compare observed channel uptake with 5.5, the expected time minimization benchmark. Our data strongly reject H1: average Channel A uptake is 7.59 in Decision Set 1, 7.71 in Decision Set 2 and 7.86 in Decision Set 3 (significantly different from 5.5, with each $p \ll 0.01$).[4] Thus, our first result is as follows:

Result 1 (Gatekeeper Aversion): *H1 is rejected. Channel B (the gatekeeper channel) uptake is lower than expected time minimization would predict.*

3.3. Discussion of Mechanisms and Utility Estimation

Experiment 1 suggests that Channel B (gatekeeper channel) uptake goes up as p^{succ} increases and as $t^{service_B}$ and t^{line_B} decrease. This aligns with intuition: a more reliable and faster service format generates a lower expected wait and should therefore be more preferred. However, approximately

[4] H1 continues to be rejected if we assume 6.0 instead of 5.5 as the benchmark for expected time minimization, which would be consistent with choosing the riskless Channel A for decision 6, in which both channels yield he same expected time ($t-$tests, each $p \ll 0.01$).

two thirds of participants continue to choose Channel A even when the expected time in that channel is the same or lower than in Channel B. This suggests that expected time minimization alone does not fully explain behavior.

The extant literature offers several plausible explanations for the observed behavior. First, participants might struggle to calculate expected values (Aimone et al. 2016a,b), leading to incorrect beliefs about wait duration. We will address this issue in §6, in which we will explicitly present participants with the expected time information for both channels. Second, participants may attach value not only to the durations but also to the content of the wait (Maister et al. 1984). Our experiment includes two different types of waits – waiting in line and waiting in service, which may be perceived and evaluated differently. Between the two waiting experiences, waiting while service is being performed may be preferred (Buell et al. 2017). Relatedly, Hathaway et al. (2021) show that waiting for a callback is associated with different waiting costs than holding on the line while waiting. Third, the gatekeeper channel is inherently risky. Less risky waits are typically found to be preferred to more risky ones (Leclerc et al. 1995, Flicker and Hannigan 2022), although risk-neutrality or even risk-seeking behavior is often observed for waiting episodes with short durations (Kroll and Vogt 2008, Festjens et al. 2015).[5] Finally, interrupting service experiences with waits in line, as in our Channel B, may be perceived as a psychological setback (Soman and Shi 2003), introducing a setup cost for each waiting episode and potentially resulting in lower satisfaction (Kumar and Dada 2021). Yu et al. (2017) and Althenayyan et al. (2022) show that such delays can lead to dissatisfaction and elevate waiting costs.

To evaluate these explanations more systematically, we estimate channel waiting costs via Maximum Likelihood Estimation (see EC.5 for estimation details). We assume that participants maximize their expected utility, consisting of the reward (value of service) net the cost of waiting (Naor 1969, Hassin and Haviv 1995). The rewards are identical across channels in our setting and cannot be separately identified in the estimation. However, the costs of waiting differ by channel. In Channel A the cost of waiting is the combined cost of waiting in line and with the server. In Channel B the cost of waiting is probabilistic and includes a possible cost of spending additional time if the first service stage fails. Denoting by $c(\cdot)$ the mapping between time and waiting cost, we can estimate the sensitivity parameters for several specifications $c(\cdot)$, and find the specification that best fits the experimental data. For example, if we assume that $c(\cdot)$ is linear in time, $c(t) = \alpha t$,

[5] To examine potential effects of risk preferences on channel choices we used a lottery task (Eckel and Grossman 2002, 2008), administered after the main task. In this task participants were asked to choose between a certain, but small, amount of money and larger, but riskier, amounts. Indeed, we found risk aversion with respect to money to be correlated with the number of Channel B choices (correlation coefficients between 0.253 and 0.262 depending on the lottery, $p = 0.018$ and $p = 0.023$). However, in our extended regression analysis (columns 2 and 4 of Table EC.2), adding risk aversion controls has minimal effects on results. Thus, to the extent that risk preferences are consistent between money and time domains, risk aversion alone does not fully explain behavior.

Table 5 Waiting Cost Estimates

	(1) **Linear** $c(t)=\alpha t$	(2) **Content of wait** $c(t^{line})=\alpha^{line}t^{line}$ $c(t^{service})=\alpha^{service}t^{service}$	(3) **Risk** $c(t)=(\alpha t)^{\beta}$	(4) **Delay** w/o delay: $c(t)=\alpha t$ w delay: $c(t)=\alpha t+\gamma\alpha t^{delay}$
α	0.212***		0.158***	0.249***
α^{line}		0.224***		
$\alpha^{service}$		0.359***		
β			1.223***	
γ				1.290***
LL	-1422.98	-1101.64	-1097.36	-1075.91
AIC	2847.95	2207.28	2198.72	2155.81

Estimates are obtained using Maximum Likelihood Estimation. (Full specification with bootstrapped standard errors in Appendix EC.5). $^{***}p<0.01$

then α is the sensitivity parameter that measures the strength of the response to duration t. We will also examine specifications in which $c(\cdot)$ is allowed to vary with the type of wait (in line vs. in service) and specifications in which $c(\cdot)$ is nonlinear.[6]

Table 5 shows the estimates. Column 1 examines a linear model, which is equivalent to assuming that decision-makers are "boundedly rational" and simply minimize expected time, but make random, symmetric errors. Perhaps unsurprisingly, this specification performs quite poorly in terms of fit ($LL=-1422.98$). Allowing the sensitivity parameters to vary between waiting in line and in service achieves a much better fit (col. 2, $LL=-1101.64$). This specification also shows that participants perceive the wait with the server to be more costly than the wait in line. This makes intuitive sense, given that participants are asked to use keystrokes during the interaction with the server but do not need to interact with the system while waiting in line. A somewhat better fit ($LL=-1097.36$) is achieved by the specification in column 3, which allows for nonlinearities and captures risk aversion. However, the specification with the best fit is in column 4 ($LL=-1075.91$). This specification introduces an explicit cost of delay that occurs whenever the gatekeeper fails.

Summarizing our estimation results, we find that models that incorporate the content of the wait, or account for risk aversion, achieve a better fit than the linear model, but do not fully explain behavior. Among the two-parameter models, explicitly modeling the disutility of a delay achieves the best fit. This explanation is most consistent with an increase in waiting costs after experiencing a delay (Soman and Shi 2003, Yu et al. 2017, Althenayyan et al. 2022). In EC.5 we examine more complex models that account for multiple factors and include more than two parameters; however, their fit is only minimally better than the fit of the models presented here.

[6] Early queueing models typically assumed that costs increase linearly in time (Naor 1969, Hassin and Haviv 1995). However, convex waiting costs have also been examined (Van Mieghem 2000, Shimkin and Mandelbaum 2004), see also Dewan and Mendelson (1990) for more general waiting cost functions.

4. Experiment 2: Algorithm Aversion

We have so far examined choices in a setting where the only differences between channels were operational, i.e., related to the sequence and durations of waiting episodes. However, in practice, channels may also differ in the content of the wait. For example, chatbot interactions are often quite structured, requiring the customer to repeatedly choose between alternatives to diagnose and resolve their request, while live agent interactions are typically more seamless. To examine choices in this richer, more realistic setting, in Experiment 2 we retain the channel choice structure of Experiment 1, but vary channel presentation and experience.

4.1. Methodology

4.1.1. Experiment Design We recruited 331 participants on Prolific; a total of 296 passed the comprehension checks. Participation was restricted to workers who did not participate in prior studies. The experiment protocol was identical to Experiment 1 (§3), shown in Figure 2 and Table 2. However, the presentation of the choices and channel experience differed by treatment:

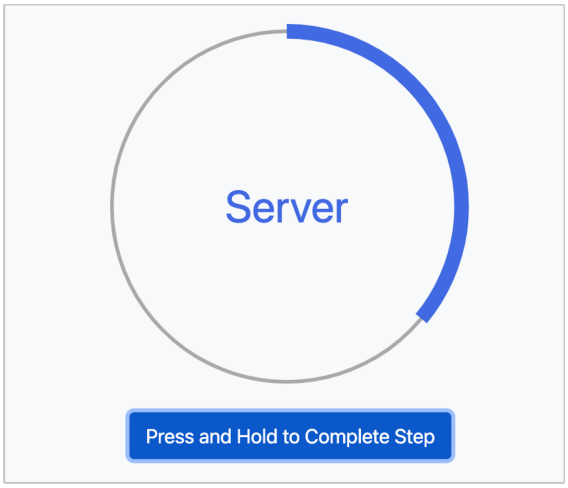
***FEEL* Treatment**. The *FEEL* treatment closely follows the *BASELINE* treatment in Experiment 1 and uses neutral channel labels for channels. However, the nature of the waiting experience is different between channels. Channel A, representing the live agent, requires the participant to hold down a button, as shown in Figure 2(a). Channel B, representing the chatbot, uses the keystrokes manipulation shown in Figure 2(b) during the first service stage. This is to emulate the continuous experience of receiving service from a live agent vs. the more fragmented experience of a message exchange with the chatbot.

***CONTEXT* Treatment**. The *CONTEXT* treatment replicates the *BASELINE* treatment from Experiment 1 but replaces the neutral choice between "Channel A" and "Channel B" with a contextualized frame. This includes naming the channels ("Live Agent Channel" and "Chatbot Channel"), as well as using the service process animations shown in Figure 2(c) and (d). However, as in the *BASELINE* treatment, the waiting experience involves keystrokes for both channels. This treatment helps isolate whether channel uptakes is affected by how the choice is presented, while keeping constant how participants interact with the system.

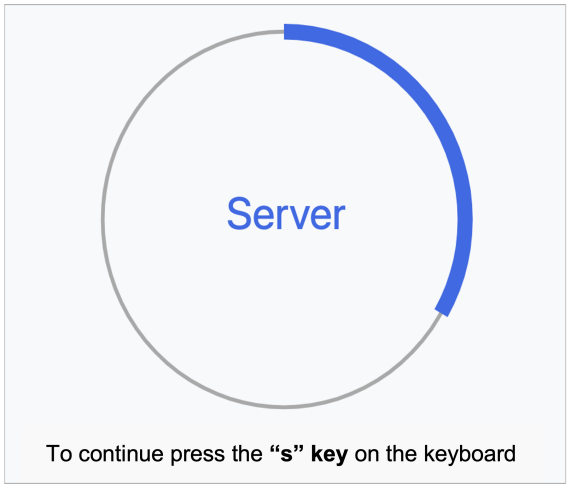
***FEEL+CONTEXT* Treatment**. In the *FEEL+CONTEXT* treatment we vary both the experiential differences between the two waiting experiences (*FEEL*), and the way channels are presented (*CONTEXT*). That is, the *FEEL+CONTEXT* treatment presents participants with both the qualitative differences between channels and with the contextual cues. The service experience is shown in Figure 2(e) and (f).

Figure 4 Waiting Experiences (Screenshots)

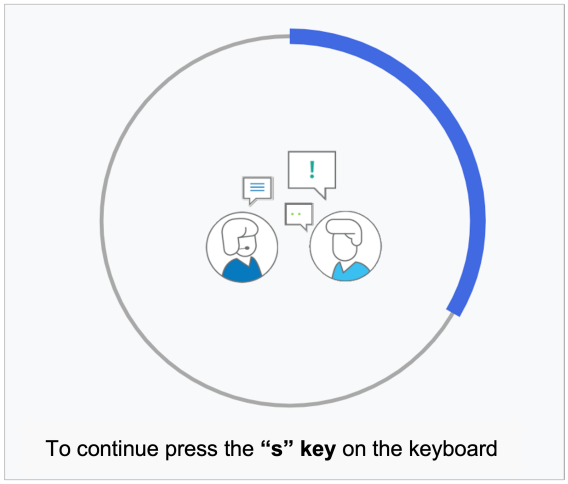
(a) *FEEL* Treatment, Channel A (Service Stage)



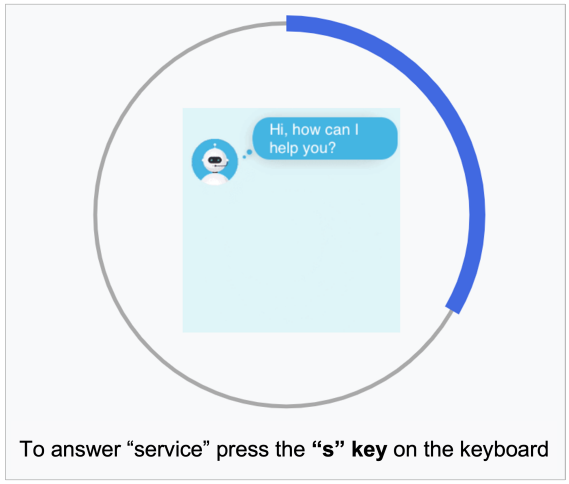
(b) *FEEL* Treatment, Channel B (First Service Stage)



(c) *CONTEXT* Treatment, Channel A (Service Stage)

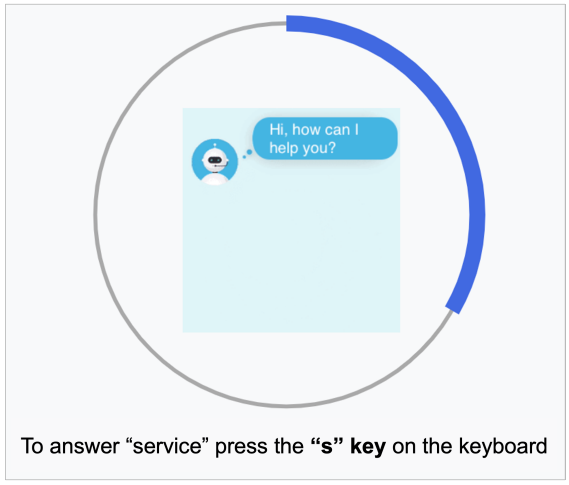


(d) *CONTEXT* Treatment, Channel B (First Service Stage)



(e) *FEEL+CONTEXT* Treatment, Channel A (Service Stage)

(f) *FEEL+CONTEXT* Treatment, Channel B (First Service Stage)



4.1.2. Manipulation Check We used our survey instrument (specifically, Survey 2 described in §2 and EC.2) to validate our experimental manipulations. In particular, in one of the survey questions we asked respondents to rate different modes of interaction – keystrokes vs. holding down a button – as more human-like or more chatbot-like. The analysis of their responses confirmed that both the neutral manipulation (Figures 4(a) and (b)), and the contextualized manipulation (Figures 4(e) and (f)) prompted the intended associations. In particular, the proportion of respondents associating the keystrokes manipulation with chatbots was 60.38% (vs. 42.45% for holding down the button, $p = 0.009$). Similarly, for the contextualized version of the manipulation, the proportion of respondents associating the keystrokes manipulation with live agents was 25.00% (vs. 79.17% for holding down the button, $p \leq 0.001$). In contrast, the holding down manipulation in panels a and e was less strongly associated with chatbots and more strongly associated with live agents (for details see EC.2). Thus, the differences were present in both the neutral and the contextualized version of the manipulations, but the gap was somewhat stronger in the latter.

4.1.3. Theory and Hypotheses To develop hypotheses in this richer setting, we build on the concept of "algorithm aversion" (Dietvorst et al. 2015). Most of the algorithm aversion literature focuses on augmenting worker performance in different operational domains, including forecasting (Dietvorst et al. 2015, Prahl and Van Swol 2017, Balakrishnan et al. 2022), service delivery (Bastani et al. 2021, Mejia and Parker 2021, Snyder et al. 2022) or order picking (Sun et al. 2022). Different from these studies, we focus on customer (rather than on worker) behavior. Most relevant to our study is Castelo et al. (2019), who find that the framing of the task as more objective/subjective can affect the degree of algorithm aversion, even when the task is unchanged. Also of relevance is the operations and marketing literature that finds that self-service technology adoption may have important effects on sales, customer loyalty and satisfaction (Curran and Meuter 2005, Buell et al. 2010, Tan and Netessine 2020, Castelo et al. 2023), but does not separately identify the effect of operational improvements caused by the technology from the effect of algorithmic perceptions. However, there is some evidence that failures of self-service machines trigger algorithm aversion (Chen et al. 2021), and that disclosing the algorithmic nature of sales agents reduces purchase rates (Luo et al. 2019). With our chatbot channel (Channel B) failing between 25% and 75% of the time, we can expect that making Channel B identifiable as a chatbot would therefore lead to a lower chatbot uptake.

Our treatments examine three possible sources of algorithm aversion. Given the extant literature, we expect all three manipulations to reduce channel uptake. The hypotheses are as follows:

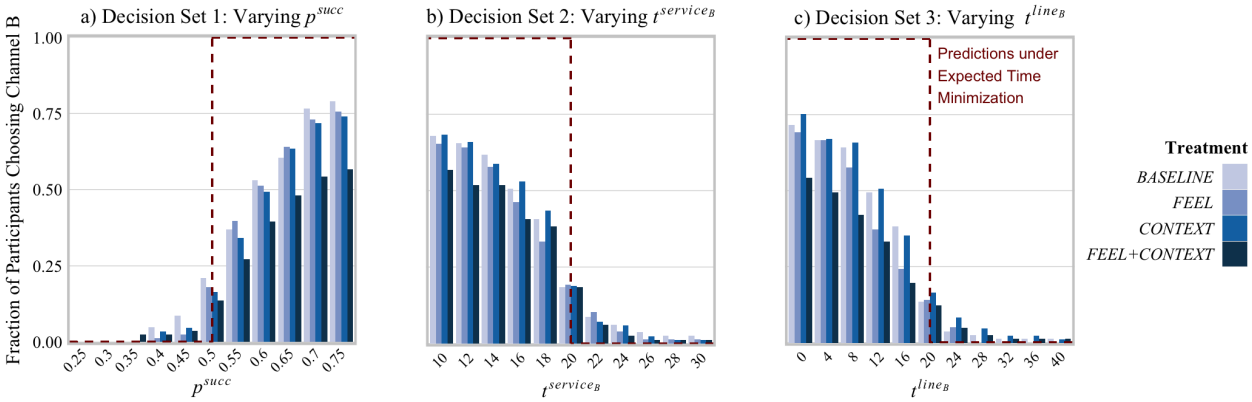
H2 (Algorithm Aversion): *Channel B uptake is reduced (relative to BASELINE) when ...*

...the waiting experience differs between channels. ***(H2A)***

...the choice is contextualized. ***(H2B)***

...the waiting experience differs between channels and the choice is contextualized. ***(H2AB)***

Figure 5 Channel B Uptake in Experiment 2



4.2. Results

4.2.1. Descriptive Statistics Figure 5 shows the Experiment 2 channel choices by treatment, decision set, and decision. As before, we highlight the hypothetical decisions that would result from all participants minimizing expected time (dashed lines). Several observations are in order. First, the increasing pattern in panel a and the decreasing patterns in panels b and c mirror the patterns observed in Experiment 1: Channel B (i.e., chatbot channel) uptake increases with chatbot performance. Second, as before, Channel B uptake remains well below the expected time minimization benchmark. Third, there appear to be minimal treatment effects of the *FEEL* and *CONTEXT* manipulations. However, there appears to be a clear drop-off in Channel B uptake in the *FEEL+CONTEXT* treatment, for $p^{succ} > 0.5$ in panel a, for $t^{service_B} < 16$ in panel b and for $t^{line_B} < 20$ in panel c (parameters at which Channel B offers expected time savings relative to Channel A).

Table 6 shows the prevalence of each participant type in the data. We observe that the number of type (1) participants, i.e., participants who never choose Channel B, remains relatively constant if we compare *BASELINE* (17.28%) with *FEEL* (17.95%, proportion test $p = 0.912$) and with *CONTEXT* (14.12%, proportion test $p = 0.575$). However, this number jumps to 29.63% in *FEEL+CONTEXT* (proportion test $p = 0.064$). Thus, there is some indication that the combination of a different waiting experience and contextual cues prompts a larger share of participants to reject chatbots, regardless of time savings offered by using them.

4.2.2. Hypothesis Tests Next, we examine the treatment effects more formally, using regression analysis. Table 7 shows the estimates (more detailed regression results are in EC.3). The treatment coefficients show that the *FEEL* and *CONTEXT* manipulations alone do not affect channel choice ($p-$values between 0.989 and 0.989 depending on the specification). However, both manipulations together significantly reduce Channel B uptake in the *FEEL+CONTEXT* treatment

($p-$values between 0.005 and 0.011, average marginal effect between -7.35 and -8.88 percentage points). Thus, we conclude that algorithm aversion is only observed when both the waiting experience is different between channels, and contextual cues are present.

Result 2 (Algorithm Aversion): *Hypothesis 2A and 2B are rejected. Hypothesis 2AB is supported. Channel B uptake is reduced when the waiting experience differs between channels and the choice is contextualized.*

4.3. Discussion

In addition to gatekeeper aversion, i.e., the unwillingness to use a service channel where the server may not be able to resolve the request (resulting in a second service stage with a different server), Experiment 2 reveals a second source of chatbot underutilization – algorithm aversion. Further, the presence of algorithm aversion depends on how the chatbot is presented. Embedding the channel choice within a more contextualized frame does not affect decisions. Similarly, varying the content of the wait alone does not affect decisions. However, when combined, the two manipulations significantly reduce chatbot uptake, suggesting that algorithm aversion can affect decisions, but only if the channels "feel" sufficiently different from each other.

Collectively, these results suggest that algorithm aversion in customer service is not a universal phenomenon, but is more situational and depends on the specifics of AI design and appearance. In particular, the result that the qualitative differences between channels matter *only* when the choice is contextualized, suggests that chatbots are avoided in part due to the users' preexisting biases and beliefs and not because of the channel differences in interaction modes. The implication of this result in practice is that de-emphasizing the algorithmic nature of the chatbot may be an effective tool for increasing adoption. We will further discuss this and other managerial levers for increasing chatbot uptake in §6 and §7.

Table 6 Participant Types in Experiment 1 and Experiment 2

	Participant Type			
	(1) Never chooses Channel B	(2) Underutilizes Channel B	(3) Expected time minimizer	(4) Other
Share of participants (%)				
Experiment 1				
BASELINE	17.28	45.68	18.52	18.52
Experiment 2				
FEEL	17.95	52.56	7.69	21.79
CONTEXT	14.12	52.94	8.24	24.71
FEEL+CONTEXT	29.63	39.51	11.11	19.75

Notes: Type (1) participants choose the live agent format in all 33 decisions. Type (2) participants choose Channel A at least once for a decision where Channel B yields a shorter expected time and minimize expected waiting time in the remaining. Type (3) make strictly expected time-minimizing choices (either five or six Channel B choices in each decision set). Type (4) collects the remaining participants. Only consistent subjects are included.

Table 7 Channel Preferences in Experiment 2

Dependent Variable:	(1) Channel B	(2) Channel B
BASELINE	-	-
FEEL	-0.073	-0.510
	(0.313)	(0.580)
CONTEXT	-0.108	-0.340
	(0.309)	(0.563)
FEEL + CONTEXT	-0.806**	-1.602***
	(0.317)	(0.579)
Intercept	2.238***	-0.021
	(0.573)	(1.053)
Channel B Performance Controls? ($p^{succ}, t^{service_B}, t^{line_B}$)	Yes	Yes
Demographic Controls?	Yes	Yes
Sample	All subjects	Consistent subjects
Observations	13002	10725
Subjects	394	325

Notes: Random effect logit regression coefficients are reported. Dependent variable is the channel choice (Channel B = 1). All specifications control for the decision set number (which serves as the time period variable in the panel data set) and for the following demographic variables: age, gender, number of quiz errors and the Eckel-Grossman risk aversion measure (administered after the main task). *** $p < 0.01$ ** $p < 0.05$ * $p < 0.1$.

5. Experiment 3: Channel Choice with Live Servers

We have so far examined channel choices using controlled manipulations of the waiting experience. While capturing one aspect of channel experience, these manipulations may not fully mimic the differences of interacting with a human vis-a-vis a machine. To add some external validity to our results, we next examine a setting where Channel A, representing the live agent, is staffed by actual live servers (research assistants).

5.1. Methodology

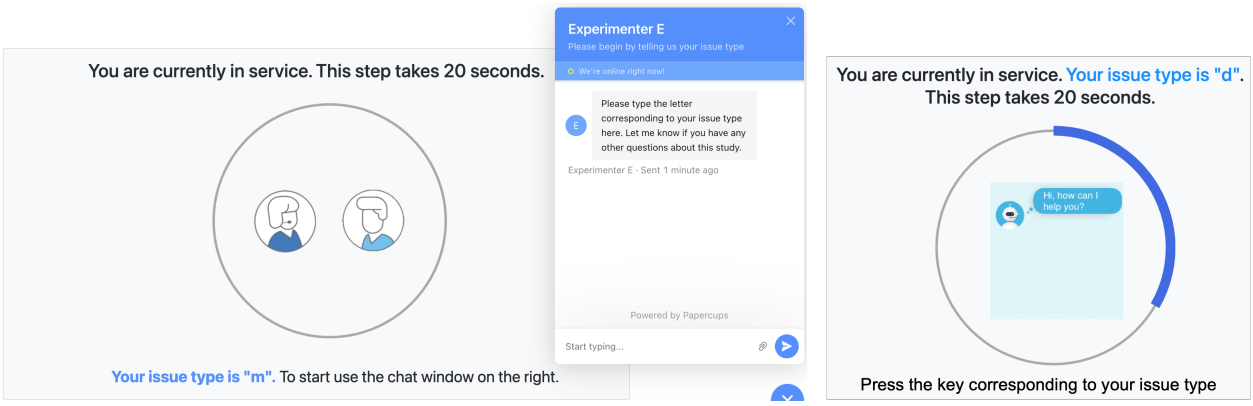
5.1.1. Experiment Design We recruited 116 participants on Prolific; a total of 108 passed the comprehension checks. Participation was restricted to Prolific workers who did not participate in prior studies. The experiment protocol was identical to Experiments 1 and 2 (Figure 1 and Table 3), with the difference being that the role of the live agent was now played by an experimenter. To this end we recruited two research assistants who had no prior knowledge of the hypotheses. We followed common practices for deploying confederates (research assistants) in experimental research (Kuhlen and Brennan 2013) and trained the assistants using a script, to control for potential differences in communication patterns. The training materials are in EC.3.

The experiment consisted of a single treatment, which we will refer to as the *LIVE* treatment. As in Experiments 1 and 2, participants made 33 decisions, with three of these decisions being implemented after the decisions were submitted.[7] To provide some context for the customer support

[7] As in Experiments 1 and 2, participants were shown a demo of each channel prior to making their decisions.

Figure 6 Screenshots of the Experiment

(a) Channel A Interaction (b) Channel B Interaction



interactions, participants were assigned an issue type at the beginning of each interaction. The issue type was represented through a letter of the alphabet. At the beginning of each service stage, participants were asked to enter that letter to begin service.

The chat interface used in Channel A is shown in Figure 6(a). The chat window popped up as soon as participants entered the first service stage in Channel A, or the second stage in Channel B (if the first stage had failed). The experimenters were instructed to initiate service (the 20 seconds in the example in the Figure 6(a) screenshot) as soon as they received the correct letter from the participant. The interface for Channel B was kept identical to the *CONTEXT* and *FEEL+CONTEXT* treatments in Experiment 2; however, to keep the framing constant across channels, we also required participants to type their issue type when interacting with Channel B (using keystrokes; see Figure 6(b)).

5.1.2. Theory and Hypotheses The algorithm aversion arguments presented in §4.1.3 suggest that Channel B uptake will be lower in the *LIVE* treatment, compared to the *BASELINE* treatment:

H2C (Algorithm Aversion, cont.): *Channel B uptake is reduced (relative to BASELINE) when the choice is between a real human server and a machine.*

5.2. Results

5.2.1. Descriptive Statistics Figure 7 summarizes the choices. As before, we highlight the hypothetical decisions that would result from all participants minimizing expected time (dashed lines). Several observations are in order. First, the patterns of choices are similar between the *BASELINE* and *LIVE* treatments. Second, there are some differences between treatments. In particular, there appears to be a consistent gap in chatbot uptake, especially for decisions where the chatbot channel results in expected time savings.

Table 8 shows the prevalence of each participant type in the data. We observe that the number of participants never choosing Channel B jumps from 17.28% in *BASELINE* to 28.09% in the *LIVE* treatment (proportion test $p = 0.047$), while the share of expected time minimizers drops from 18.52% to 7.87% (proportion test $p = 0.039$). This indicates that the human nature of the server prompts a significant portion of decision-makers to *never* consider the chatbot channel.

5.2.2. Hypothesis Tests Next, we examine the treatment effects more formally using regression analysis. Table 9 shows the estimates. To examine the hypothesized effects of the human live agent relative to a neutral benchmark we include the *BASELINE* treatment. We also include the *FEEL+CONTEXT* treatment (from Experiment 2), which presents participants with a similarly rich, contextualized decision, yet does not include a real human server. The results suggest that the *LIVE* treatment indeed leads to a drop in Channel B uptake. Depending on the specification, the size of the average marginal effect ranges between 4.88 and 5.83 percentage points. The effect is marginally significant ($p = 0.084$ and $p = 0.088$). Further, looking at the demographic data, we find participant age to be a significant positive moderator of the *LIVE* treatment effect.[8] (See Table EC.4 for details.) Finally, the differences between *FEEL+CONTEXT* and *LIVE* treatment dummies are not significant ($p-$values between 0.370 and 0.596), suggesting similar levels of algorithm aversion resulting from these manipulations.

Result 2 (Algorithm Aversion, cont.): *Hypothesis 3C is (weakly) supported. Channel B uptake is reduced under the LIVE manipulation.*

5.3. Discussion

The results of Experiment 3 reinforce our previous findings. First, Channel B (chatbot channel) uptake continues to be below the expected time minimization benchmark. Second, there is a

[8] In particular, after including an interaction term between age and the *LIVE* treatment dummy, the *LIVE* treatment effect is significant at $p < 0.05$ for participants aged 40 and above.

Figure 7 Channel B Uptake in Experiment 3

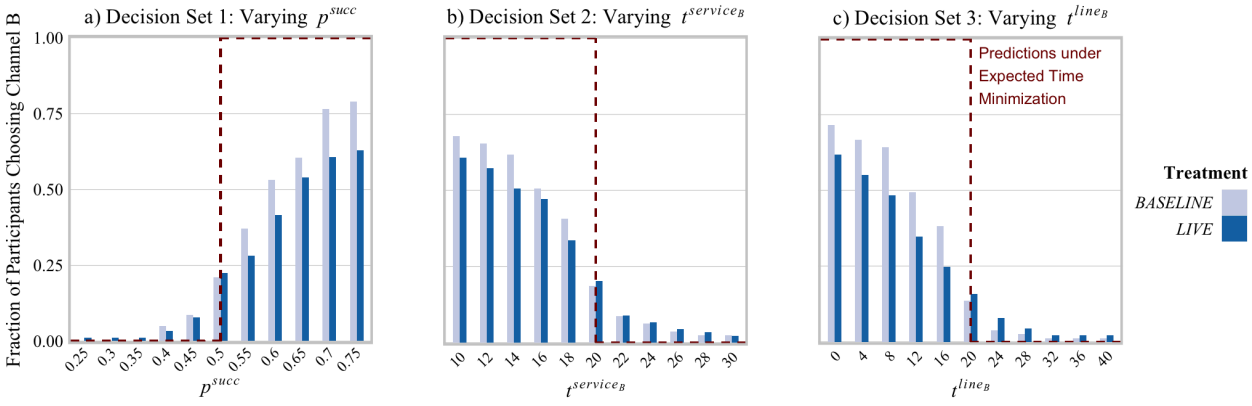


Table 8 Participant Types in Experiment 1 and Experiment 3

	Participant Type			
	(1) Never chooses Channel B	(2) Underutilizes Channel B	(3) Expected time minimizer	(4) Other
Share of participants (%)				
Experiment 1				
BASELINE	17.28	45.68	18.52	18.52
Experiment 3				
LIVE	28.09	40.45	7.87	23.60

Notes: Type (1) participants choose the live agent format in all 33 decisions. Type (2) participants choose Channel A at least once for a decision where Channel B yields a shorter expected time and minimize expected waiting time in the remaining. Type (3) make strictly expected time-minimizing choices (either five or six Channel B choices in each decision set). Type (4) collects the remaining participants. Only consistent subjects are included.

Table 9 Channel Preferences in Experiment 3

Dependent Variable:	(2) Channel B	(4) Channel B
BASELINE	-	-
FEEL+CONTEXT	-0.806**	-1.493**
	(0.368)	(0.652)
LIVE	-0.613*	-1.094*
	(0.360)	(0.633)
Intercept	3.521***	1.484
	(0.748)	(1.328)
Channel B Performance Controls? ($p^{succ}, t^{service_B}, t^{line_B}$)	Yes	Yes
Demographic Controls?	Yes	Yes
Sample	All subjects	Consistent subjects
Observations	9000	8283
Subjects	300	251

Notes: Random effect logit regression coefficients are reported. Dependent variable is the channel choice (Channel B = 1). Specifications (1) includes all subjects; specification (2) only includes the subjects with consistent choices throughout the task, respectively. All specifications control for the decision set number, channel B performance controls and the following demographic variables: age, gender, number of quiz errors and risk aversion. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.1$.

marginally significant drop of Channel B uptake relative to the *BASELINE* case, driven primarily by participants who reject Channel B altogether and do not respond to operational improvements. Further, the differences between the *FEEL+CONTEXT* treatment in Experiment 2 and the *LIVE* treatment in Experiment 3 are minimal, suggesting that both manipulations prompt a similar level of algorithm aversion. In the next sections we will further discuss the relative importance of gatekeeper aversion and algorithm aversion, as well as examine a way to mitigate these behaviors and increase chatbot uptake.

6. Experiment 4: Nudge

To examine potential ways to increase chatbot adoption we replicated the *BASELINE* treatment from Experiment 1 and the *FEEL+CONTEXT* treatment from Experiment 2, but added a nudge that presents participants with the expected waiting time for Channel B (the gatekeeper/chatbot

channel). A total of 208 participants were recruited on Prolific. Participation was restricted to workers who did not participate in prior studies.

6.1. Methodology

6.1.1. Experiment Design The nudge consisted of displaying the total expected waiting time (line + service) for the chatbot channel. Figure 8 shows a screenshot of both the original decision screen (Experiments 1-3) in panel a for Decision Set 1 and the decision screen for the same decision set with a nudge (Experiment 4) in panel b. The average waiting time information was added in all 33 decisions. To better understand how the nudge operates, we conducted two between-subject treatments.

***BASELINE+NUDGE* Treatment.** This treatment is identical to the *BASELINE* treatment, plus the expected waiting time information.

Figure 8 Screenshots of the Experiment

(a) Decision Set 1 Screen Shot in Experiments 1-3

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 Set 1 Screen Shot in Experiment 4

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)**

***FEEL+CONTEXT+NUDGE* Treatment.** This treatment is identical to the *FEEL+CONTEXT* treatment, plus the expected waiting time information.

6.1.2. Theory and 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 expected utility maximization, but rather focus on some salient set of choice components (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 away from the format differences between channels (both operational and non-operational/algorithmic) and encourage them to opt for the channel with the shorter expected waiting time.

Our nudge intervention is also related to the literature on delay announcements (Ibrahim 2018, and references there). This literature shows that providing waiting time information can help lower abandonment rates, but the optimal level and granularity of information may vary across contexts (Guo and Zipkin 2007, Jouini et al. 2011, Yu et al. 2017, Akşin et al. 2017). Our study is different

Figure 9 Channel B Uptake in Experiment 4

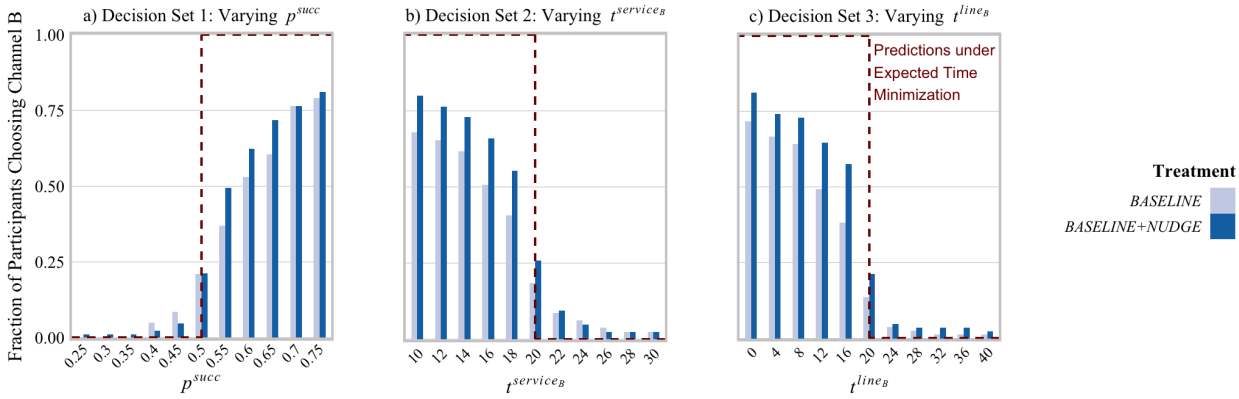
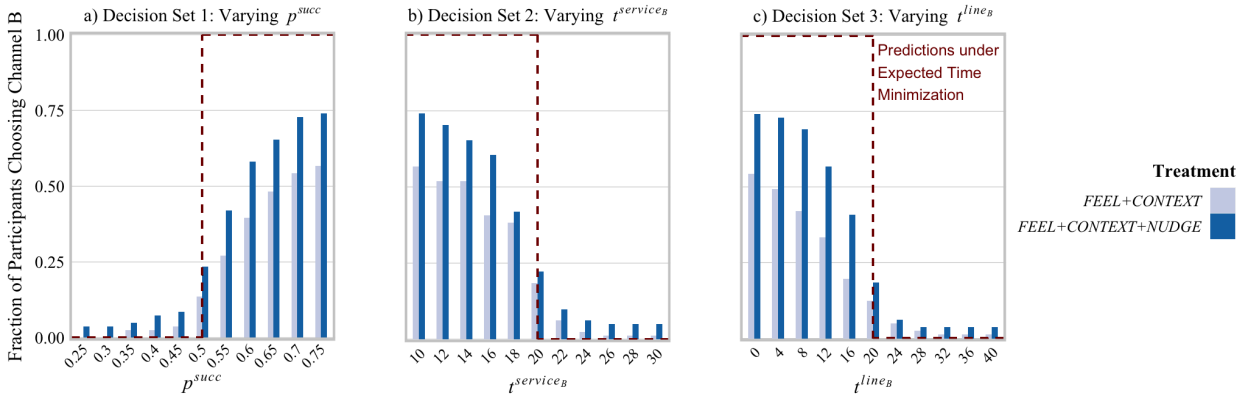


Figure 10 Channel B Uptake in Experiment 4 (continued)



from this literature in two ways. First, the literature typically assumes that the customer has limited information. This is not true in our setting, where the customer has full distributional information on the duration of each transaction – our nudge simply directs customer attention to a specific aspect of the waiting time distribution. Second, rather than choosing to join or to abandon a queue, the customer in our case makes a choice between two service channels. While Channel A may be interpreted as an "outside" option of abandoning a queue, the trade-offs are different in our setting, since each channel involves its own progression of waiting stages. Despite these differences, Yu et al. (2017) provide the closest comparison and suggest that delay information can lower waiting costs and make a variable channel, such as our Channel B, more appealing.

H3 (Nudge): *The expected waiting time nudge increases Channel B uptake.*

6.2. Results

6.2.1. Descriptive Statistics Figures 9 and 10 show the Experiment 4 channel choices by treatment, decision set, and decision. The figures suggest that the nudge indeed increases Channel B uptake. The effect is somewhat stronger if we compare the *FEEL+CONTEXT* and *FEEL+CONTEXT+NUDGE* treatments, reaching up to 25 percentage points for some decisions, but is still visibly present in the *BASELINE* and *BASELINE+NUDGE* treatments. As in the previous experiments, the treatment differences are concentrated in the decisions in which Channel B offers some time savings relative to Channel A.

Table 10 shows the split of types in the data. We observe that in the *BASELINE* scenario the nudge increases the share of participants of type (3), i.e, those who consistently minimize expected time, from 18.52% to 36.47%, and decreases the share of type (2) (proportion tests $p = 0.010$ and $p = 0.020$). Similarly, in the *FEEL+CONTEXT* scenario the largest increase is in the share of type (3) (proportion tests $p = 0.038$). This change is driven mainly by a decrease in the share of type (1) participants, i.e., those who never choose Channel B. This provides some indication that the nudge makes expected time minimization more prevalent as a decision strategy.

6.2.2. Hypothesis Tests As before, we use panel data logit regressions to test our hypotheses. The regression coefficients are reported in Table 11. The bottom panel also reports pairwise comparisons between the relevant treatments. The results suggest a significant effect of nudge on Channel B uptake in the *FEEL+CONTEXT* case, with effect sizes ranging between 12 and 14 percentage points. The average marginal effect of the nudge is significant at $p = 0.003$ in both specifications. In contrast, the effect size is quite small and not statistically significant if we compare the *BASELINE* and *BASELINE+NUDGE* treatments ($p = 0.199$ and $p = 0.272$). Further, the difference between *BASELINE+NUDGE* and *FEEL+CONTEXT+NUDGE* treatments is minimal

Table 10 Participant Types in Experiment 1 and Experiment 3

	Participant Type			
	(1) Never chooses Channel B	(2) Underutilizes Channel B	(3) Expected time minimizer	(4) Other
Share of participants (%)				
Experiment 1 & Experiment 4				
BASELINE	17.28	45.68	18.52	18.52
BASELINE+NUDGE	16.47	28.24	36.47	18.82
Experiment 2 & Experiment 4				
FEEL+CONTEXT	29.63	39.51	11.11	19.75
FEEL+CONTEXT+NUDGE	20.99	38.27	23.46	17.28

Notes: Type (1) participants choose the live agent format in all 33 decisions. Type (2) participants choose Channel A at least once for a decision where Channel B yields a shorter expected time and minimize expected waiting time in the remaining. Type (3) make strictly expected time-minimizing choices (either five or six Channel B choices in each decision set). Type (4) collects the remaining participants. Only consistent subjects are included.

Table 11 Channel Preferences in Experiment 4

Dependent Variable:	(1) Channel B	(2) Channel B
BASELINE	-	-
BASELINE+NUDGE	0.524	0.809
	(0.380)	(0.736)
FEEL+CONTEXT	-0.915**	-1.890**
	(0.383)	(0.744)
FEEL+CONTEXT+NUDGE	0.317	0.351
	(0.381)	(0.745)
Channel B Performance Controls? ($p^{succ}, t^{service_B}, t^{line_B}$)	Yes	Yes
Demographic Controls?	Yes	Yes
Sample	All subjects	Consistent subjects
N	12672	10824
Subjects	384	328
Pairwise Comparisons of Treatment Effects ($p-$value)		
BASELINE = BASELINE+NUDGE	0.199	0.272
FEEL+CONTEXT = FEEL+CONTEXT+ NUDGE	0.003	0.003

Notes: Random effect logit regression coefficients are reported. Dependent variable is the channel choice (Channel B = 1). Specifications (1) includes all subjects; specification (2) only includes the subjects with consistent choices throughout the task, respectively. All specifications control for the decision set number, channel B performance controls and the following demographic variables: age, gender, number of quiz errors and risk aversion. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.1$.

and not statistically significant. This suggests that the effects of non-operational factors on chatbot uptake (algorithm aversion) are largely muted when average waiting times are presented. These results are summarized below.

Result 3 (Nudge): *H3 is partially supported. The nudge significantly increases Channel B uptake in the FEEL+CONTEXT treatment, but not in the BASELINE treatment.*

6.2.3. Discussion Experiment 4 shows that the hurdles to chatbot adoption can be mitigated by explicitly presenting participants with average waiting time information. Notably, increased adoption is statistically significant only in the contextualized *FEEL+CONTEXT* treatment and

not in the *BASELINE* treatment. Thus, the nudge works primarily by redirecting attention away from the qualitative channel differences and towards the time savings offered by the chatbot. In addition to suggesting a useful tool for managing service systems, this result adds to our discussion of the mechanisms behind gatekeeper aversion (§3.3), confirming that gatekeeper aversion is not driven primarily by participants' inability to compute expected times, but rather is a response to the delays that lower the quality of the waiting experience in the gatekeeper channel. We will discuss further implications of the nudge for the firm's service design and costs in the next section.

7. General Discussion and Implications

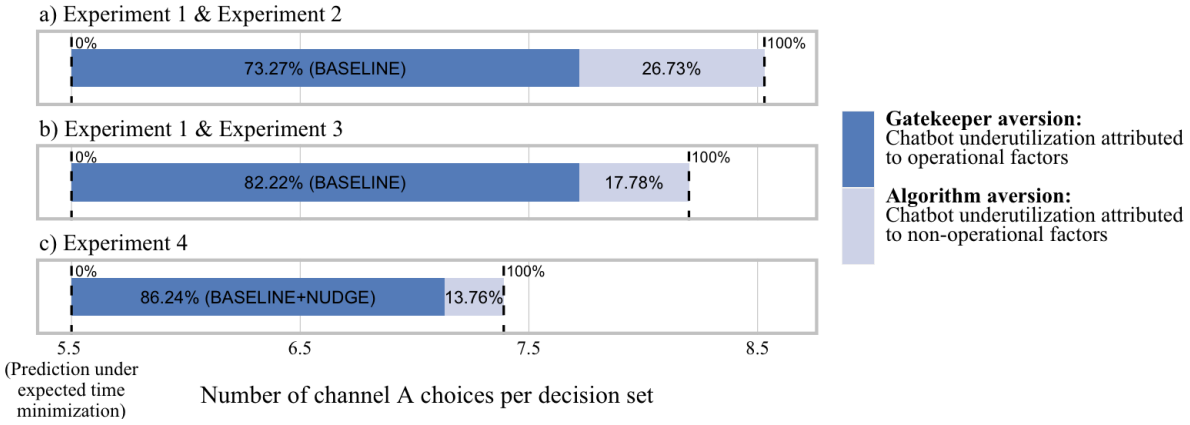
This section offers an integrated discussion of our results and explores their practical implications. We proceed in three steps. We first present an integrative analysis of all four experiments and evaluate the relative importance of the operational (gatekeeper aversion) and non-operational (algorithm aversion) factors in chatbot uptake. We then revisit our retrospective survey from §2 and discuss an external validity check on our main result. Finally, we discuss the implications of our results for service design and staffing costs.

7.1. Gatekeeper Aversion and Algorithm Aversion

Experiments 1-4 advance our understanding of the customer channel choice between an algorithmic chatbot and a live agent. First, the choice data in the neutral, non-contextualized setting 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 (Experiment 1). Second, chatbot uptake is further reduced when the service process is sufficiently different between channels, whether this difference is caused by a manipulation of the waiting experience (Experiment 2), or by introducing an actual human into the process (Experiment 3). Further, the result that neither experiential differences, nor the presence of context are sufficient alone, but the combination of the two produces a significant effect (Experiment 2), suggests an aversive attitude (bias) against chatbot technology that is reinforced when the chatbot feels sufficiently different than the live agent. Finally, some of these behaviors can be mitigated by directing customer attention towards the operational benefits of the channel choice (Experiment 4).

Figure 11 presents an integrated summary of the above results. This figure shows the number of Channel A choices in a subset of our treatments. Recall H1, which stated that under expected time minimization, Channel A should be chosen in half of the decisions (i.e., in 5.5 out of 11 decisions per decision set). Any number above 5.5 suggests gatekeeper and/or algorithm aversion. Our test of H1 in Experiment 1 showed that the number of Channel A choices was between 7.59 and 7.86 (depending on decision set), with an average of 7.72. This is plotted in darker blue in panels a and b of Figure 11. In both these scenarios, there is a noticeable effect of introducing qualitative

Figure 11 Gatekeeper Aversion and Algorithm Aversion



differences between channels (algorithm aversion), which leads to a further drop in Channel B utilization, or conversely to an increase in Channel A utilization. In particular, the number of Channel A choices increases from 7.72 to 8.53 in the *FEEL+CONTEXT* treatment (plotted in panel a) and to 8.20 in the *LIVE* treatment (plotted in panel b). These increases are relatively small – in fact, the share of algorithm aversion remains below 30% across all comparisons. Finally, the share of algorithm aversion is even smaller (at 13.76%) once we introduce the nudge in panel c. Taken together, these comparisons suggest that gatekeeper aversion is the dominant factor, while algorithm aversion is a secondary hurdle that may further reduce adoption in certain settings.

7.2. Additional Survey Question

To validate the above result (that gatekeeper aversion is the primary factor driving channel choices) we return to the survey of §2 and re-examine one of the questions asked in that survey. In particular, one of the questions in Survey 2 (Q15) asked the respondents to choose a response category that most accurately describes their attitudes towards choosing a service channel. Based on the themes identified in Survey 1, we included an operational and a non-operational aspect of each channel and examined how frequently each category was chosen. The response categories are listed below (% of respondents choosing each option is in parentheses):

1. *I prefer live agents because I like interacting with a human.* (15.84%)
2. *I prefer chatbots because I do not like interacting with a human.* (12.38%)
3. *I prefer live agents because they can handle more complicated requests.* (54.46%)
4. *I prefer chatbots because they are faster to access.* (17.33%)

The responses indicate that channel choices are more closely related to operational factors (speed and performance) than to the algorithmic/human nature of the server. In particular, a total of 71.79% of respondents invoke operational factors as the key determinant of their choice, with only 28.19% prioritizing the algorithmic/human nature of the server. Thus, the incentivized decisions made in our experiments are broadly consistent with the self-reported attitudes in the survey.

7.3. Service Design Implications

In addition to contributing to the academic discourse on the role of technology in service design, our results suggest an important managerial lever that can be effective in increasing chatbot uptake and reducing costs. In particular, in customer support settings where chatbots are faster than live agents, managers may be able to increase chatbot uptake by explicitly presenting customers with expected time savings. To evaluate the benefits of such interventions we developed an analytic approach that feeds our experimental data into a structural model of channel demand and estimates the resultant staffing level and costs.

Our approach and results are described in detail in Appendix EC.6. In the first step we use the utility functions estimated from our experimental data (using the approach of §3.3) to compute the relative demand for each channel for a range of system parameters. In the second step we compute aggregate demand for live agents, using the direct demand for the live agent channel and the indirect demand from failed chatbot interactions. In the third step we compute the required staffing level that satisfies aggregate demand given a desired service level. Our results in Appendix EC.6 suggest that the nudge improves adoption over a wide range of system parameters and leads to cost savings of up to 20.2%.

8. Concluding Remarks

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 and analyzing testimonies from chatbot users, by using these user stories to formulate a key trade-off in channel choice, and by studying how online users navigate this trade-off in incentivized experiments.

We first studied choices in a neutral setting, focusing on the response to the operational differences between channels. We found that many participants had a strong preference for the more seamless experience of interacting with a single server, even when the alternative would have resulted in substantial expected time savings – a behavior that we termed "gatekeeper aversion". In subsequent experiments we studied choices in a richer setting and found chatbot uptake to be further reduced when the chatbot used a distinctly robotic communication mode. This suggests an ex ante bias against chatbots that is not driven by the qualitative differences between the channels, but is more deeply ingrained in people's experiences and beliefs about chatbot technology. Finally, we examined the effects of a simple nudge that presents users with expected waiting times in each channel and found that it can significantly increase chatbot uptake.

Together, our results suggest that standard economic analysis of channel choice (based solely on the analysis of waiting times) 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 identified two such deviations: one based on the operational consequences of choosing a particular channel ("gatekeeper aversion") and a second one, driven by algorithmic attitudes of the customer ("algorithm aversion"). Between gatekeeper aversion and algorithm aversion, we found the former to account for 73% to 86% of chatbot underutilization.

Our contributions help support future work on technology adoption in service systems. First, we contribute to the algorithm aversion literature, which focuses primarily on algorithmic attitudes in human-AI collaboration, by demonstrating that algorithm aversion also plays a role for customers. We identify contingencies when algorithm aversion is present and measure its relative importance. Second, we use a novel experimental framework for examining queue joining behaviors when service providers differ not only in duration, but also in the content of the wait. This framework can serve as a template for identifying operational and non-operational effects of other algorithmic technologies on customer attitudes and adoption behavior.

As chatbot and other algorithmic technologies continue to evolve, new questions related to their usage come to the fore. These include the role of algorithmic preferences when interacting with industry and firm-specific vs. general-purpose chatbots (enabled by the latest generations of large language models), the role of privacy and the use of customer data in these service interactions, as well as richer interaction environments, such as voice, video or virtual reality. Controlled experiments can greatly add to our understanding of the relevant trade-offs in these environments, and can help identify pain points and solutions. The broader takeaway for service managers is to remain cognizant of behavioral implications of service technologies as new AI capabilities are developed.

References

Adam M, Wessel M, Benlian A (2021) Ai-based chatbots in customer service and their effects on user compliance. *Electronic Markets* 31(2):427–445.

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.

Akşin Z, Ata B, Emadi SM, Su CL (2017) Impact of delay announcements in call centers: An empirical approach. *Operations Research* 65(1):242–265.

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

Althenayyan A, Cui S, Ulku S, Yang L (2022) Not all lines are skipped equally: an experimental investigation of line-sitting and express lines. *Available at SSRN 4179751* .

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).

Benke I, Gnewuch U, Maedche A (2022) Understanding the impact of control levels over emotion-aware chatbots. *Computers in Human Behavior* 129:107122.

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

Buell RW, Campbell D, Frei FX (2010) Are self-service customers satisfied or stuck? *Production and Operations Management* 19(6):679–697.

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

Castelo N, Boegershausen J, Hildebrand C, Henkel AP (2023) Understanding and improving consumer reactions to service bots. *Journal of Consumer Research* ucad023.

Castelo N, Bos MW, Lehmann DR (2019) Task-dependent algorithm aversion. *Journal of Marketing Research* 56(5):809–825.

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.

Chen N, Mohanty S, Jiao J, Fan X (2021) To err is human: Tolerate humans instead of machines in service failure. *Journal of Retailing and Consumer Services* 59:102363.

Curran JM, Meuter ML (2005) Self-service technology adoption: comparing three technologies. *Journal of services marketing* 19(2):103–113.

Dewan S, Mendelson H (1990) User delay costs and internal pricing for a service facility. *Management Science* 36(12):1502–1517.

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.

Feinerer I, Hornik K, Meyer D (2008) Text mining infrastructure in r. *Journal of statistical software* 25:1–54.

Festjens A, Bruyneel S, Diecidue E, Dewitte S (2015) Time-based versus money-based decision making under risk: An experimental investigation. *Journal of Economic Psychology* 50:52–72.

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).

Guo P, Zipkin P (2007) Analysis and comparison of queues with different levels of delay information. *Management Science* 53(6):962–970.

Hassin R, Haviv M (1995) Equilibrium strategies for queues with impatient customers. *Operations Research Letters* 17(1):41–45.

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).

Ibrahim R (2018) Sharing delay information in service systems: a literature survey. *Queueing Systems* 89(1-2):49–79.

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

Jouini O, Akşin Z, Dallery Y (2011) Call centers with delay information: Models and insights. *Manufacturing & Service Operations Management* 13(4):534–548.

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

Krippendorff K (2018) *Content analysis: An introduction to its methodology* (Sage publications).

Kroll EB, Vogt B (2008) Loss aversion for time: an experimental investigation of time preferences. *Working Paper Series* .

Kuhlen AK, Brennan SE (2013) Language in dialogue: When confederates might be hazardous to your data. *Psychonomic bulletin & review* 20:54–72.

Kumar P, Dada M (2021) Investigating the impact of service line formats on satisfaction with waiting. *International Journal of Research in Marketing* 38(4):974–993.

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* .

Luo X, Tong S, Fang Z, Qu Z (2019) Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases. *Marketing Science* 38(6):937–947.

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* .

Porter MF (1980) An algorithm for suffix stripping. *Program* 14(3):130–137.

Prahl A, Van Swol L (2017) Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting* 36(6):691–702.

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.

Shimkin N, Mandelbaum A (2004) Rational abandonment from tele-queues: Nonlinear waiting costs with heterogeneous preferences. *Queueing Systems* 47:117–146.

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.

Soman D, Shi M (2003) Virtual progress: The effect of path characteristics on perceptions of progress and choice. *Management Science* 49(9):1229–1250.

Sun J, Zhang DJ, Hu H, Van Mieghem JA (2022) Predicting human discretion to adjust algorithmic prescription: A large-scale field experiment in warehouse operations. *Management Science* 68(2):846–865.

Tan TF, Netessine S (2020) At your service on the table: Impact of tabletop technology on restaurant performance. *Management Science* 66(10):4496–4515.

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.

Van Mieghem JA (2000) Price and service discrimination in queuing systems: Incentive compatibility of gc μ scheduling. *Management Science* 46(9):1249–1267.

Yu Q, Allon G, Bassamboo A (2017) How do delay announcements shape customer behavior? an empirical study. *Management Science* 63(1):1–20.

Electronic Companion

EC.1. Supporting Analysis for Survey 1

Below we provide details and data for Survey 1 described in §2 of the manuscript.

EC.1.1. Survey Setup and Questions

The data were collected using the Prolific platform.[9] 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.]

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*

[9] 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.

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%).

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

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

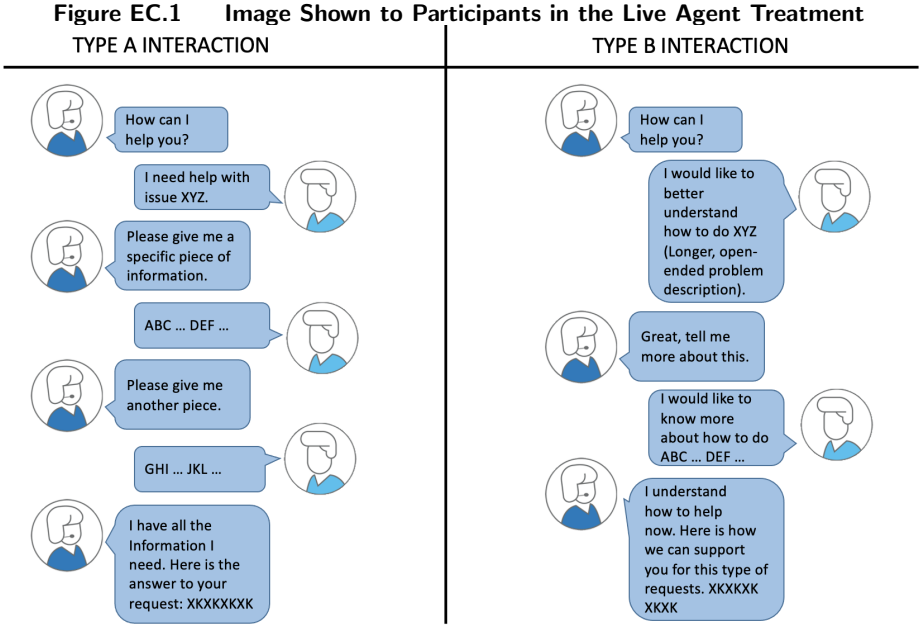
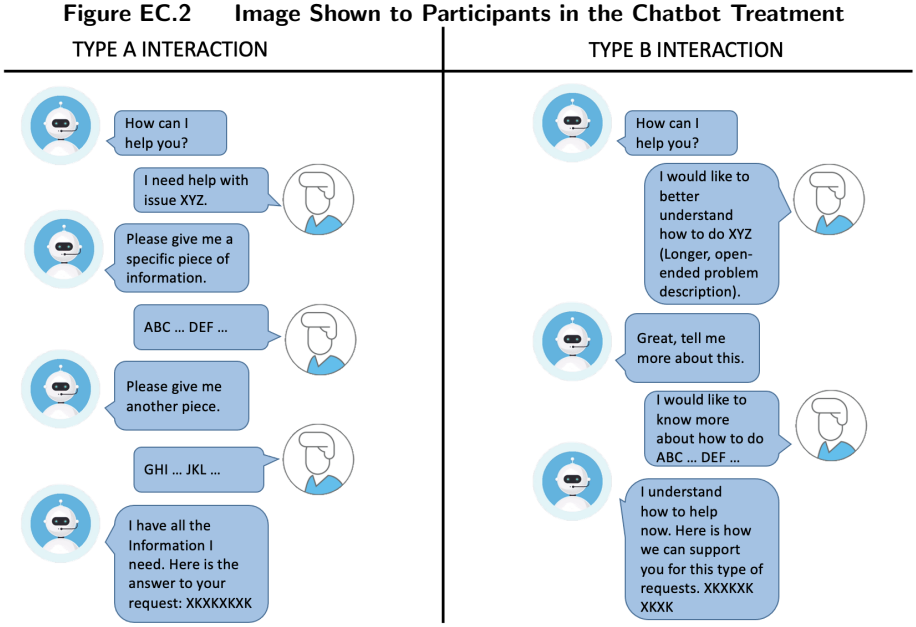


Figure EC.2 Image Shown to Participants in the Chatbot Treatment



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.2. Supporting Analysis for Survey 2 (§2 and §4.1.2)

Below we provide details and data for Survey 2 described in §2 and §4.1.2 of the manuscript.

EC.2.1. Survey Setup and Questions

The survey was conducted in July 2023 on the Prolific platform. A total of 202 respondents were recruited from the US-based population (49% female, average age: 38). All respondents received a show up payment of $3.00 and an additional payment of $2.00 at the end of the study.[10] The survey included a replication of most of the questions asked in Survey 1, with the addition of several new questions, including a manipulation check for the experiments in §4 and §6.

EC.2.2. Replication of Questions from Survey 1

Below we reproduce the portion of Survey 2 which involved a replication of the Survey 1 questions. 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 (7.84%) Yes (90.20%) Not Sure (1.96%).
Chatbot: No (5.00%) Yes (93.00%) Not Sure (2.00%).

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

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]. Aiming for 1-2 sentences, please answer the following questions.*

- *What caused you to contact customer support? 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, using a chatbot [live customer support], or 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 [a chatbot]?*

[10] Prolific workers who had participated in the previous experiments or surveys were excluded from participation.

[Note: a valid response required a minimum of 50 characters for each qeustion.]

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 (49.91%) Yes (51.09%).
Chatbot: No (61.29%) Yes (38.71%).

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)? [Question dropped in Survey 2]

Q5. *Of the two interaction types below* [Note: Figures EC.1/EC.2], *which one more closely resembles the customer support experience you described?* [Question dropped in Survey 2]

Q6. *How long did you have to wait until the agent [chatbot] became available?*
Live Agent: Less than 1 minute (33.33%) 1-2 minutes (30.39%) At least 3 minutes (36.28%).
Chatbot: Less than 1 minute (79.00%) 1-2 minutes (16.00 %) At least 3 minutes (5.00%).

Q7. *How long did you have to wait for the agent [chatbot] relative to your initial expectation?*[Question dropped in Survey 2]

Q8. *How long did the interaction with the agent [chatbot] last?*
Live Agent: Less than 1 minute (13.73%) 1-2 minutes (8.82%) At least 3 minutes (77.45%).
Chatbot: Less than 1 minute (10.00%) 1-2 minutes (36.00%) At least 3 minutes (54.00%).

Q9. *How long did the interaction last relative to your initial expectation?* [Question dropped in Survey 2]

Q10. *Did you have to share any information (e.g., order number, address, name, date of birth) with the agent [chatbot]?* [Question dropped in Survey 2]

Q11. *Approximately how many questions did the agent [chatbot] ask you during the interaction?* [Question dropped in Survey 2]

Q12. *Was the agent [chatbot] able to resolve your request?*

Live Agent: No, I was transferred to another agent (4.35%) No, I had to call a different number to resolve the issue (1.09%) No, the issue remained unresolved (7.61%) Yes (86.96%).

Chatbot: No, I was transferred to another agent (26.88%) No, I had to call a different number to resolve the issue (13.98%) No, the issue remained unresolved (17.20%) Yes (41.94%).

Q13. *Overall, how satisfied were you with the customer support interaction? (1: very dissatisfied, 5: very satisfied)*

Live Agent: 3.27 on average.

Chatbot: 2.27 on average.

EC.2.3. Additional Questions

Q14. *How did you interact with the live agent [chatbot]?*

Live Agent: I called a phone number (68.48%) I used live chat (31.52%)

Chatbot: By talking (2.15%) By typing (chat) (97.85%)

Q15. *In day-to-day interactions with customer service, which of the following most accurately describes you?*

I prefer live agents because I like interacting with a human. (15.84%) I prefer chatbots because I do not like interacting with a human. (12.38%) I prefer live agents because they can handle more complicated requests. (54.46%) I prefer chatbots because they are faster to access. (17.33%)

EC.2.4. Manipulation Check

Survey 2 included an attitudinal test that validates the key experimental manipulations in §4 and §6. In particular, §4 introduces two experimental manipulations: (1) an informational manipulation (which we refer to as *CONTEXT*) which explicitly present the channel choice as the choice between a chatbot and a live agent, and (2), a qualitative manipulation (which we refer to as *FEEL*), which varies the nature of the waiting experience in each channel. In the survey, we presented respondents with each manipulation and then asked them to rate each type of manipulation as more human-like or more chatbot-like. The methodology and results are presented next.

EC.2.4.1. Methodology There were two versions of the manipulation check. Each participant experienced only one version, with the version being assigned at random at the beginning of the survey. In the first version (N=106), respondents were asked to rate two different 20-second waiting experiences. The first experience required holding down a button for the progress bar to fill (Figure 4(a)). The second experience involved pressing certain keys at intervals for the progress bar to fill, similar to the waiting experience in Experiment 1 (Figure 4(b)). After completing both experiences (in a randomly assigned order), participants were asked to rate whether each experience resembled a chatbot or a live agent interaction.

In the second version of the survey (N=96), participants completed the same tasks but with added context: one experience included a gif-animated interaction between a call center agent and a customer (Figure 4(c)), and the other featured a chatbot producing messages (Figure 4(d)). The key variable of interest is the percentage of respondents rating each type of interaction as being more human-like or more bot-like.

EC.2.4.2. Results Table EC.1 reports the key result of the survey. We use proportion tests for statistical comparisons, however, Chi-square tests yield similar results. Consider first the *FEEL* manipulation. 31.13% of respondents rated holding down the button (Channel A) as being more human-like vs. 27.36% in the keystrokes manipulation (Channel B), with the difference not being statistically significant (Proportion test, $p = 0.545$). Further, 42.45% of respondents rated Channel A as being more bot-like vs. 60.38% for Channel B (proportion test, $p = 0.009$). Next consider the *FEEL+CONTEXT* manipulation. Here, 50.00% of respondents rated Channel A as being more human-like vs. 15.62% for Channel B ($p \ll 0.001$). Conversely, 25.00% perceived Channel A to be more bot-like, vs. 79.17% for Channel B ($p \ll 0.001$). The remaining respondents chose the *"Not sure"* category.

Table EC.1 Manipulation Check

	"Felt more like a live agent interaction"			*"Felt more like a chatbot interaction"*			*"Not sure"*		
	Channel A	Channel B	$p-$value	Channel A	Channel B	$p-$value	Channel A	Channel B	$p-$value
FEEL manipulation	31.13%	27.36%	0.545	42.45%	60.38%	0.009	26.42%	12.26%	0.009
FEEL+CONTEXT manipulation	50.00 %	15.62 %	$\ll 0.001$	25.00 %	79.17 %	$\ll 0.001$	25.00 %	5.21 %	$\ll 0.001$

Notes: Table reports percentage of respondents in each category, as well as $p-$values from non-parametric tests of equality of proportions of each response category between channel A and channel B.

Taken together, these comparisons validate our manipulations. There are significant differences in channel perceptions both with and without context, but the differences are stronger when context is added.

EC.3. Supporting Analysis for Experiments

In this section we present supporting analyses for §3-§6. Table EC.2 presents extended analysis for §3, in particular the increase of Channel B uptake with the improvement of its parameters, discussed in §4.2.1 and shown in Figure 3. This table includes additional specifications without demographic controls (columns 1 and 3), as well as the estimates for the channel parameters ($p^{succ}, t^{service_B}$ and t^{line_B}). Analogously, Table EC.3 presents extended analysis for §4, in particular, Table 7 and Table EC.4 presents extended analysis for §4, in particular, Table 9. This table also includes specifications with an interaction term between *LIVE* treatment and age, discussed in the main text. Finally, Table EC.5 presents extended analysis for §4, in particular, Table 11.

Table EC.2 Channel Preferences in the *BASELINE* Treatment

	(1)	(2)	(3)	(4)
Dependent Variable:	Channel B	Channel B	Channel B	Channel B
p^{succ}	13.453***	13.459***	33.093***	33.101***
	(0.812)	(0.812)	(2.129)	(2.127)
$t^{service_B}$	-0.369***	-0.369***	-0.612***	-0.612***
	(0.022)	(0.022)	(0.040)	(0.040)
t^{line_B}	-0.200***	-0.200***	-0.372***	-0.373***
	(0.012)	(0.012)	(0.024)	(0.024)
Female		-0.475		-0.890
		(0.491)		(0.956)
Age		0.00503		0.0411
		(0.0211)		(0.0446)
Quiz errors		-0.292		-0.607
		(0.252)		(0.484)
Risk aversion		0.207*		0.427*
		(0.114)		(0.230)
Intercept	2.765***	2.336**	-0.692	-2.827
	(0.674)	(1.184)	(1.274)	(2.438)
Sample	All subjects	All subjects	Consistent subjects	Consistent subjects
Observations	3234	3234	2673	2673
Subjects	98	98	81	81

Notes: Random effect logit regression coefficients are reported. Dependent variable is the channel choice (Channel B = 1). 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). *** $p < 0.01$ ** $p < 0.05$ * $p < 0.1$.

Table EC.3 Channel Preferences in Experiment 2

Dependent Variable:	(1) Channel B	(2) Channel B	(3) Channel B	(4) Channel B
BASELINE	-	-	-	-
FEEL	-0.101	-0.073	-0.401	-0.510
	(0.328)	(0.313)	(0.612)	(0.580)
CONTEXT	0.039	-0.108	-0.009	-0.340
	(0.325)	(0.309)	(0.598)	(0.563)
FEEL+CONTEXT	-0.849**	-0.806**	-1.512**	-1.602***
	(0.333)	(0.317)	(0.614)	(0.579)
p^{succ}	12.598***	12.601***	27.947***	27.937***
	(0.396)	(0.396)	(0.912)	(0.911)
$t^{service_B}$	-0.336***	-0.336***	-0.556***	-0.557***
	(0.010)	(0.010)	(0.018)	(0.018)
t^{line_B}	-0.175***	-0.175***	-0.316***	-0.316***
	(0.005)	(0.005)	(0.011)	(0.011)
Female		0.213		0.302
		(0.231)		(0.418)
Age		-0.0208**		-0.0206
		(0.00887)		(0.0161)
Quiz errors		-0.472***		-1.193***
		(0.135)		(0.266)
Risk aversion		0.263***		0.450***
		(0.0548)		(0.101)
Intercept	2.255***	2.238***	0.263	-0.021
	(0.385)	(0.573)	(0.698)	(1.053)
Sample	All subjects	All subjects	Consistent subjects	Consistent subjects
Observations	13002	13002	10725	10725
Subjects	394	394	325	325
Pairwise Comparisons of Treatment Effects ($p-$values)				
BASELINE = FEEL	0.759	0.512	0.815	0.379
BASELINE = CONTEXT	0.905	0.988	0.727	0.545
BASELINE = FEEL+CONTEXT	0.011	0.014	0.011	0.006
FEEL = CONTEXT	0.668	0.516	0.912	0.766
FEEL = FEEL+CONTEXT	0.025	0.073	0.021	0.060
CONTEXT = FEEL+CONTEXT	0.007	0.013	0.027	0.027

Notes: Random effect logit regression coefficients are reported. Dependent variable is the channel choice (Channel B = 1). 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). *** $p < 0.01$ ** $p < 0.05$ * $p < 0.1$.

Table EC.4 Channel Preferences in Experiment 3

	(1) Channel B	(2) Channel B	(3) Channel B	(4) Channel B	(5) Channel B	(6) Channel B
BASELINE	-	-	-	-	-	-
FEEL+CONTEXT	-0.906**	-0.806**	0.784	-1.599**	-1.493**	2.093
	(0.387)	(0.368)	(1.186)	(0.692)	(0.652)	(2.170)
LIVE	-0.563	-0.613*	2.119*	-1.055	-1.094*	4.606**
	(0.378)	(0.360)	(1.123)	(0.676)	(0.633)	(2.080)
p^{succ}	12.62***	12.62***	12.62***	26.94***	26.96***	26.96***
	(0.469)	(0.469)	(0.469)	(1.029)	(1.028)	(1.028)
$t^{service_B}$	-0.366***	-0.366***	-0.366***	-0.562***	-0.563***	-0.563***
	(0.0129)	(0.0129)	(0.0129)	(0.0216)	(0.0216)	(0.0216)
t^{line_B}	-0.181***	-0.181***	-0.181***	-0.308***	-0.308***	-0.308***
	(0.00668)	(0.00668)	(0.00668)	(0.0121)	(0.0121)	(0.0121)
Female		-0.173	-0.113		-0.282	-0.190
		(0.304)	(0.302)		(0.533)	(0.527)
Age		-0.037***	0.006		-0.054**	0.0412
		(0.0126)	(0.0225)		(0.0221)	(0.0428)
Quiz errors		-0.336**	-0.299*		-0.654**	-0.552*
		(0.171)	(0.170)		(0.302)	(0.300)
Risk aversion		0.289***	0.289***		0.496***	0.502***
		(0.0730)	(0.0724)		(0.129)	(0.127)
Age × *FEEL+CONTEXT*			-0.045			-0.098*
			(0.0310)			(0.0556)
Age × *LIVE*			-0.078**			-0.158***
			(0.030)			(0.055)
Intercept	2.886***	3.521***	1.928*	0.681	1.484	-2.085
	(0.459)	(0.748)	(1.005)	(0.791)	(1.328)	(1.904)
Sample	All subjects	All subjects	All subjects	Consistent subjects	Consistent subjects	Consistent subjects
Observations	9900	9000	9900	8283	8283	8283
Subjects	300	300	300	251	251	251
Marginal effect of *LIVE* at...			***p*−value:**			***p*−value:**
Age= 20			0.332			0.178
Age= 25			0.717			0.454
Age= 30			0.571			0.847
Age= 35			0.087			0.139
Age= 40			0.010			0.009
Age= 45			0.003			0.002
Age= 50			0.002			0.001
Age= 55			0.002			0.001
Age= 60			0.003			0.001

Notes: Random effect logit regression coefficients are reported. Dependent variable is the channel choice (Channel B = 1). Specifications (1)-(3) include all subjects that passed the screening questions. Specifications (4)-(6) include only the subjects with consistent choices throughout the task. All specifications control for the decision set number (which serves as the time period variable in the panel data set). *** $p < 0.01$ ** $p < 0.05$ * $p < 0.1$. The bottom panel reports significance levels for the *LIVE* treatment effects at different levels of the *Age* variable.

Table EC.5 Channel Preferences in Experiment 4

Dependent Variable:	(1) Channel B	(2) Channel B	(3) Channel B	(4) Channel B
BASELINE	-	-	-	-
BASELINE+NUDGE	0.547	0.524	0.979	0.809
	(0.390)	(0.380)	(0.762)	(0.736)
FEEL+CONTEXT	-0.932**	-0.915**	-1.815**	-1.890**
	(0.391)	(0.383)	(0.772)	(0.744)
FEEL+CONTEXT+NUDGE	0.358	0.317	0.478	0.351
	(0.390)	(0.381)	(0.775)	(0.745)
p^{succ}	13.75***	13.75***	31.37***	31.39***
	(0.430)	(0.430)	(1.039)	(1.037)
$t^{service_B}$	-0.385***	-0.385***	-0.682***	-0.682***
	(0.0116)	(0.0116)	(0.0227)	(0.0227)
t^{line_B}	-0.191***	-0.191***	-0.370***	-0.370***
	(0.00592)	(0.00592)	(0.0124)	(0.0124)
Female		-0.0512		0.0937
		(0.283)		(0.548)
Age		-0.0107		-0.0136
		(0.0112)		(0.0219)
Quiz errors		-0.382**		-1.031***
		(0.156)		(0.318)
Risk aversion		0.232***		0.457***
		(0.0699)		(0.137)
Intercept	2.631***	2.464***	1.417*	0.863
	(0.431)	(0.703)	(0.810)	(1.377)
Sample	All subjects	All subjects	Consistent subjects	Consistent subjects
Observations	12672	12672	10824	10824
Subjects	384	384	328	328

Notes: Random effect logit regression coefficients are reported. Dependent variable is the channel choice (Channel B = 1). 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. All specifications control for the decision set number. Specifications (2) and (4) control for demographic variables: age, gender, number of quiz errors, and risk aversion. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.1$.

EC.4. Experimental Stimuli

EC.4.1. Experiments 1 and 2: Details and Instructions

The data were collected on the Prolific platform. 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. The instructions for Experiment 2 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 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.4.2. Experiment 3: Script and Training

As in Experiments 1-2, 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 August 2023, on weekdays between 9am and 6pm Eastern Time.

Channel B was unchanged relative to Experiment 2 (*FEEL+CONTEXT* and *CONTEXT* treatments). In Channel A we recruited two research assistants and trained them using a chat script. We reproduce the chat script below. Depending on the choices, participants may interact with the research assistant up to three times (because three of the choices are chosen to be experienced in real time at the end of the experiment). In addition, participants interact with the research assistant prior to making any choices to test the interface.

Participant: *[starts conversation]*

Experimenter: *"Hello. Looks like we are good to go. Do you have any questions?"*

Participant: *[responds]*

Experimenter: "What is your issue type?"

Participant: *[enters issue type]*

Experimenter: "Got it. Advancing you."

[*Experimenter starts service process. Participant sees the progress bar fill which takes* $t^{service_A}$ *seconds. Participant is able to move on to next page after* $t^{service_A}$ *seconds*].

EC.4.3. Experiment 4: Details and Instructions

As in Experiments 1-3, 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. Instructions for *BASELINE+NUDGE* treatment were unchanged relative to Experiment 1 and instructions for *FEEL+CONTEXT+NUDGE* treatment were unchanged relative to Experiment 2. The sole difference were the decision screens shown in Figure 8.

EC.5. Random Utility Models

In this section we provide the estimation approach for our various utility models to explain the observed gatekeeper aversion in Experiment 1. Since the reward parameter cannot be separately identified by channel, we normalize it to zero, which reduces the choice problem to a waiting-cost minimization problem. Under a given specification, we denote the cost that participant i incurs in Channel A for decision j by $C_{ij}^A(\boldsymbol{\theta}, t_{ij}^{line_A}, t_{ij}^{service_A})$, which depends on the time in line ($t_{ij}^{line_A}$), the time in service ($t_{ij}^{service_A}$) and the population vector of parameters that describe the sensitivity to different types of waiting ($\boldsymbol{\theta}$). The Channel B waiting cost is denoted by $C_{ij}^B(\boldsymbol{\theta}, p_{ij}^{succ}, t_{ij}^{service_B}, t_{ij}^{line_B}, t_{ij}^{service_A})$, and depends on the time in service with the gatekeeper ($t_{ij}^{service_B}$), the success probability (p_{ij}^{succ}), and contingent on failure, the time spent waiting for and being served by the second server ($t_{ij}^{line_B}, t_{ij}^{service_A}$). Finally, the term ϵ_{ij}^A (ϵ_{ij}^B) is the idiosyncratic shock of choosing Channel A (B) in decision j. Then, the respective utilities that customer i receives by choosing each channel in decision j are then given by:

$$U_{ij}^A(\boldsymbol{\theta}) = -C_{ij}^A(\boldsymbol{\theta}, t_{ij}^{line_A}, t_{ij}^{service_A}) + \epsilon_{ij}^A,$$
$$U_{ij}^B(\boldsymbol{\theta}) = -C_{ij}^B(\boldsymbol{\theta}, p_{ij}^{succ}, t_{ij}^{service_B}, t_{ij}^{line_B}, t_{ij}^{service_A}) + \epsilon_{ij}^B.$$

By assuming that the shock terms are i.i.d type-1 extreme value distributed, the choice probabilities are in the familiar closed-form Logit probabilities and $\boldsymbol{\theta}$ can be estimated via the maximum likelihood method.

In §3.3 we proposed three explanations that could rationalize the observed gatekeeper aversion: the content of wait, risk with respect to time, and an aversion to delay after failure. We next present the cost specifications ($C_{ij}^A(\boldsymbol{\theta}, t_{ij}^{line_A}, t_{ij}^{service_A}), C_{ij}^B(\boldsymbol{\theta}, p_{ij}^{succ}, t_{ij}^{service_B}, t_{ij}^{line_B}, t_{ij}^{service_A})$) that capture each of these explanations as well as the specifications for every combination of two of the three explanations:

1. **Linear:**
$$C_{ij}^A = \alpha(t_{ij}^{line_A} + t_{ij}^{service_A})$$
$$C_{ij}^B = \alpha[t_{ij}^{service_B} + (1 - p_{ij}^{succ})(t_{ij}^{line_B} + t_{ij}^{service_A})]$$
2. **Content of Wait:**
$$C_{ij}^A = \alpha^{line} t_{ij}^{line_A} + \alpha^{service} t_{ij}^{service_A}$$
$$C_{ij}^B = \alpha^{service} t_{ij}^{service_B} + (1 - p_{ij}^{succ})(\alpha^{line} t_{ij}^{line_B} + \alpha^{service} t_{ij}^{service_A})$$
3. **Risk:**
$$C_{ij}^A = \alpha(t_{ij}^{line_A} + t_{ij}^{service_A}))^\beta$$
$$C_{ij}^B = p_{ij}^{succ}(\alpha t_{ij}^{service_B})^\beta + (1 - p_{ij}^{succ})(\alpha(t_{ij}^{service_B} + t_{ij}^{line_B} + t_{ij}^{service_A}))^\beta$$
4. **Delay:**

$$C_{ij}^A = \alpha(t_{ij}^{line_A} + t_{ij}^{service_A})$$
$$C_{ij}^B = \alpha t_{ij}^{service_B} + (1 - p_{ij}^{succ})\gamma\alpha(t_{ij}^{line_B} + t_{ij}^{service_A})$$

5. **Content of Wait + Risk:**

$$C_{ij}^A = (\alpha^{line} t_{ij}^{line_A} + \alpha^{service} t_{ij}^{service_A})^\beta$$
$$C_{ij}^B = p_{ij}^{succ}(\alpha^{service} t_{ij}^{service_B})^\beta + (1 - p_{ij}^{succ})(\alpha^{service} t_{ij}^{service_B} + \alpha^{line} t_{ij}^{line_B} + \alpha^{service} t_{ij}^{service_A})^\beta$$

6. **Content of Wait + Delay:**

$$C_{ij}^A = \alpha^{line} t_{ij}^{line_A} + \alpha^{service} t_{ij}^{service_A}$$
$$C_{ij}^B = \alpha^{service} t_{ij}^{service_B} + (1 - p_{ij}^{succ})\gamma(\alpha^{line} t_{ij}^{line_B} + \alpha^{service} t_{ij}^{service_A})$$

7. **Risk + Delay:**

$$C_{ij}^A = \alpha(t_{ij}^{line_A} + t_{ij}^{service_A}))^\beta$$
$$C_{ij}^B = p_{ij}^{succ}(\alpha t_{ij}^{service_B})^\beta + (1 - p_{ij}^{succ})\gamma(\alpha(t_{ij}^{service_B} + t_{ij}^{line_B} + t_{ij}^{service_A}))^\beta$$

In Table EC.6 we present the estimates and standard errors. We remark that Specification 4 (delay) achieves the best fit (in AIC) of the single-explanation specifications. Moreover, of the two-explanation specifications, (7) achieves better fit than that of (4) but only nominally. Finally, we estimated the specification that allows for all three explanations but it did not achieve better fit (in AIC) than that of (4).

Table EC.6 Waiting Cost Estimates: BASELINE

	(1) **Linear**	(2) **Content**	(3) **Risk**	(4) **Delay**	(5) **Content + Risk**	(6) **Content + Delay**	(7) **Risk + Delay**
α	0.212***		0.158***	0.249***			0.254***
	(0.020)		(0.014)	(0.022)			(0.041)
α^{line}		0.224***			0.176***	0.247***	
		(0.021)			(0.019)	(0.021)	
$\alpha^{service}$		0.359***			0.216***	0.277***	
		(0.027)			(0.028)	(0.025)	
β			1.223***		1.134***		0.996***
			(0.025)		(0.028)		(0.086)
γ				1.290***		1.225***	1.198***
				(0.041)		(0.045)	(0.042)
LL	-1422.98	-1101.64	-1097.36	-1075.91	-1092.68	-1073.65	-1073.47
AIC	2847.95	2207.28	2198.72	2155.81	2191.36	2153.29	2152.93

Bootstrapped standard errors. $^{***}p < 0.01$

EC.6. Nudge Implications

In this section we estimate cost savings resulting from implementing the expected waiting time nudge documented in §6. 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 first estimate a random utility model of the customer choice between the live agent and chatbot channels under the *FEEL+CONTEXT* and *FEEL+CONTEXT+NUDGE* treatments. We then use the estimated parameters to predict aggregate customer demand (traffic) into each channel under a range of system parameters with and without the nudge. Finally we perform counterfactual analyses by comparing the optimized staffing costs in the M/D/1 queuing regimes under the "no nudge" and the "nudge" service designs, and evaluate the cost savings.

EC.6.1. Random Utility Estimation

We used the same Maximum Likelihood approach for estimating the cost parameters as described in Appendix EC.5. In this case, we used Specification 6, which allows for differences in linear waiting costs based on content and for uplifting the second stage waiting cost when the chatbot fails. Since interacting with Channel A is different in content than with Channel B, we allow for two separate linear costs of service ($\alpha^{service_A}, \alpha^{service_B}$). We also separately estimate the delay uplift with and without the nudge (γ, γ^{nudge}). We present the estimates in Table EC.7 below.

Table EC.7 Waiting Cost Estimates: FEEL + CONTEXT and FEEL + CONTEXT + NUDGE

Symbol	Estimate	Standard Error
α^{line}	0.219***	(0.017)
$\alpha^{service_A}$	0.246***	(0.025)
$\alpha^{service_B}$	0.269***	(0.020)
γ	1.260***	(0.058)
γ^{nudge}	1.085***	(0.048)

Bootstrapped standard errors. $^{***}p < 0.01$.

We note that, as was the case under the *BASELINE* estimates (see (6) in Table EC.6), the waiting costs while in service exceed the in-line waiting cost. Moreover, the cost with the bot is higher than that of the agent ($\alpha^{service_A} = 0.246$, $\alpha^{service_B} = 0.269$). Indeed, this raises the disutility of choosing the bot, which rationalizes the algorithm aversion we observed under the *FEEL+CONTEXT* treatment. Finally, we note that the second-stage uplift is significantly lower under the nudge treatment ($\gamma = 1.260, \gamma^{nudge} = 1.085$). This decreases the expected disutility of the chatbot option, which rationalizes the increased uptake of the chatbot under the nudge.

EC.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. The relative demand for each channel is formed according the logit choice probabilities using the estimates in Table EC.7. 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, which lead to choice probabilities, which in turn affect the waiting times. The procedure to compute equilibrium staffing level (and costs) is as follows:

- **Fix System Design** As we showed in our experiments, chatbot uptake depends in part on the offered waiting times of customers who go directly to the live agent (t^{line_A}) and those who go to the agent after chatbot failure (t^{line_B}). In our analysis we hold these two equal ($t^{line_A} = t^{line_B} = t^{line}$) but vary t^{line} over a variety of offered waiting times. Moreover, we determine the presence (or absence) of the nudge (indicated with the binary variable N), as well as the remaining system parameters that affect the choice probabilities $\boldsymbol{S} = \{t^{line}, t^{service_A}, t^{service_B}, p^{succ}, N\}$.
- **Calculate Customer Flows** Given the system design ($\boldsymbol{S}$) and the estimated utility parameters ($\boldsymbol{\theta}$), we can calculate the arrival rate to the live server as follows. Let the customer traffic to the system be λ per unit time. Then a fraction of the customers, $\rho^A(\boldsymbol{\theta}, \boldsymbol{S})$, go directly to the live channel, where on average the focal customer i spends t^{line} in queue and $t^{service_A}$ with the agent. In contrast, the remaining fraction, 1 – $\rho^A(\boldsymbol{\theta}, \boldsymbol{S})$, of customers first go the chatbot channel where customer i spends $t^{service_B}$ in service. Further, in the event that the bot fails to resolve the issue, which occurs with probability (1-p^{succ}), the customer is transferred to the live channel and on the average spends t^{line} in queue and $t^{service_A}$ with the agent. Thus, of the total arrival rate λ, the arrival rate to the live channel is given by

$$\lambda^A(\boldsymbol{\theta}, \boldsymbol{S}) = \lambda \cdot (\rho^A(\boldsymbol{\theta}, \boldsymbol{S}) + (1 - \rho^A(\boldsymbol{\theta}, \boldsymbol{S}) \cdot (1 - p^{succ})).$$

- **Calculate Staffing Cost:** 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. 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 time can be calculated as follows (derived from Tijms 2003, p. 59):

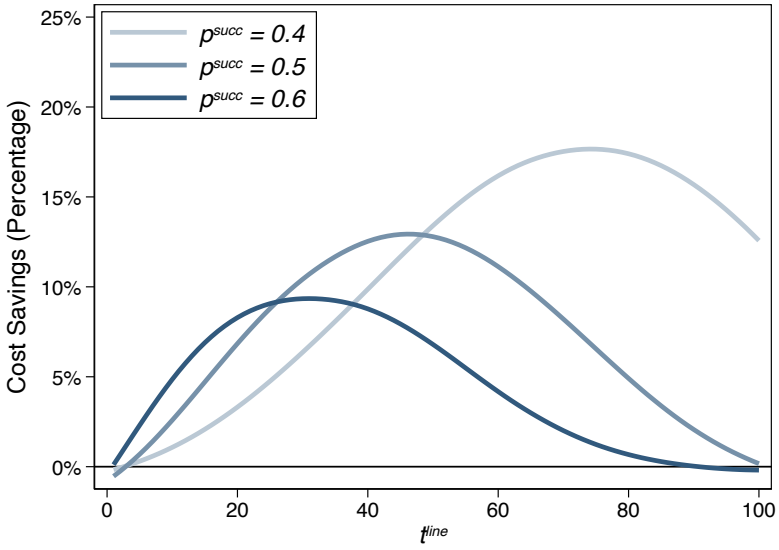
$$\mu(\lambda^A(\boldsymbol{\theta}, \boldsymbol{S})) = \frac{\lambda^A(\boldsymbol{\theta}, \boldsymbol{S}) + \sqrt{\lambda^A(\boldsymbol{\theta}, \boldsymbol{S})^2 + \frac{2 \cdot \lambda^A(\boldsymbol{\theta}, \boldsymbol{S})}{t^{line} + t^{service_A}}}}{2} \tag{EC.EC.6.1}$$

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

EC.6.3. Cost Comparisons

We estimate staffing costs with and without the nudge. 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 $t^{service_A}$ to 20 and $t^{service_B}$

Figure EC.3 Cost Savings Achieved by Expected Time Nudge



to 20. We vary t^{line} by increments of 1 from 1 to 100. 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 $\mu(\lambda^A(\boldsymbol{\theta}, \boldsymbol{S}))$). Figure EC.3 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 t^{line}. When 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 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 t^{line} that the nudge has a significant enough effect on demand to substantially decrease staffing costs, with a maximum savings of 20.2% when $p^{succ} = 0.4$.[11]

[11] For robustness, we also repeated our analysis under the $M/M/1$ regime and found qualitatively similar results, with a maximum savings of 19.1% when $p^{succ} = 0.4$. Detailed results are available from the authors upon request.