Deploying 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

Abstract. Problem definition: Despite recent advances in Artificial Intelligence, the use of chatbot technology in customer service continues to face adoption hurdles. This paper explores reasons for these adoption hurdles and tests several service design levers to increase chatbot uptake. **Methodology/results:** We use incentivized online experiments to study chatbot uptake in a variety of scenarios. The results of these experiments are threefold. First, people respond positively to improvements in chatbot performance; however, the chatbot channel is utilized less frequently than expected-time minimization would predict. A key driver of this underutilization is the reluctance to engage with a gatekeeper process, i.e., a process with an imperfect initial service stage and possible transfer to a second, expert service stage – a behavior we term *gatekeeper aversion*. We show that gatekeeper aversion can be further amplified by a secondary hurdle, algorithm aversion. Second, chatbot uptake can be increased by providing customers with average waiting times in the chatbot channel, as well as by being more transparent about chatbot capabilities and limitations. Third, methodologically, we show that chatbot adoption can depend on experimental implementation. In particular, chatbot adoption decreases further as (i) stakes are increased, (ii) the human/algorithmic nature of the server is manipulated with more realism. **Managerial Implications:** Our results suggest that firms should continue to prioritize investments in chatbot technology. However, less expensive, process-related interventions can also be effective. These may include being more transparent about the types of queries that are (or are not) suitable for chatbots, emphasizing chatbot reliability and quick resolution times, as well as providing faster live agent access to customers who experienced chatbot failure.

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

1. Introduction

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 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.

However, academic research into the drivers of chatbot technology adoption remains scarce. 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, on whether anthropomorphism (human-likeness) helps or hurts adoption. 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 deploying chatbots, their integration into the broader service delivery strategy, and their effects on the cost and performance of a service system. In this paper we seek to address this gap and study chatbot technology from a service operations perspective. We focus on chatbot adoption as a choice among several service channels offered within a service system, each with its own unique processes and customer experiences.

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 entry 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 many healthcare or hospitality settings, which *require* the patient or customer to go through the gatekeeper to begin service, chatbot operators often allow customers to *choose* between a live agent and a chatbot. In this study we examine the determinants of this channel choice and test several levers to increase chatbot uptake.

1.1. Study design

The starting point of our investigation is a retrospective survey, in which we ask 400 respondents to describe a recent customer service episode, either with a chatbot or with a live agent. Quantitative and qualitative analyses of their testimonies suggest a key trade-off in channel choice: chatbots are faster to access but have a lower request resolution rate. In contrast, live agents typically require some wait to access but are much more reliable in resolving customer requests. This insight helps motivate a simple model of channel choice (§2) which is then tested in our experiments (§3-5).

The first experiment (§3) focuses on identifying adoption hurdles. In this experiment we present participants with a series of choices between two alternatives. The first alternative ("Channel A") represents the live (A)gent and involves some waiting in line to access the server. The server then resolves the request with probability 1. The second alternative ("Channel B") represents the (B)ot 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. In both channels, upon successful resolution of the service request the customer receives a fixed monetary reward, which represents service completion. Depending on the parameters in a decision, the expected-time minimizing choice may be either Channel A or Channel B. There are

three experimental conditions: a *Context* treatment, in which the type of the server (human or bot) is explicitly revealed, a *No Context* treatment, in which all contextual cues are removed and participants choose between two visually identical (but process-differentiated) channels, and a *No Context, Deterministic* treatment which removes the uncertainty from Channel B. These treatments enable us to separately identify process-related preferences that exist independently of contextual framing from preferences related to the algorithmic nature of the service provider.

In the second experiment (§4) we focus on potential remedies for chatbot underutilization. Drawing on the literature in behavioral operations and decision theory, we test two alternative designs. In particular, we present participants with choices that are mathematically identical to those in Experiment 1 but change how information is presented. First, drawing on research on operational transparency (Buell and Norton 2011, Buell et al. 2017, Balakrishnan et al. 2022), the *Context + No Transparency* treatment deliberately reduces operational transparency. In this treatment, the chatbot always suggests a solution for each customer request, with the offered solution being either correct or incorrect. This is different from our *Context* treatment, where the chatbot is transparent and truthfully reports whenever it is able to handle a request or not. Second, in the *Context + Nudge* treatment we nudge participants to focus on the potential time savings offered by the chatbot by explicitly presenting the total average waiting times for both channels.

In the third experiment (§5) we turn to the methodological challenge of measuring algorithmic aversion in the customer service setting by introducing two treatments that add realism to our experimental setup. In the first treatment (*Context + Live*) we replicate the *Context* treatment but use actual humans (research assistants, blind to the experimental hypotheses) who play the role of live agents and who interact with participants using a live chat tool. In the second treatment (*Context + Hold*) we make salient differences between live agents and chatbots by requiring continuous, physical engagement in Channel A (representing the live agent) while retaining click-based interaction in Channel B (representing the chatbot).

1.2. Key Results and Contributions

The results of our experiments show that Channel B uptake declines as the chatbot service time grows longer, chatbot failure rate increases, or the wait for a live agent following chatbot failure increases. In other words, better operational performance leads to higher uptake. Nonetheless, across all three experiments, Channel B uptake remains considerably below what one would predict from a purely expected-time minimization perspective. Our results are summarized in Table 1.

In Experiment 1 we first show that Channel B underutilization is tied primarily to process-related hurdles. That is, participants are willing to spend more time in the system (in expectation) in order to avoid interacting with a gatekeeper channel, whether or not the decision is contextualized (as a choice between a live agent and a chatbot) or not. We term this behavior *gatekeeper aversion*. We further decompose gatekeeper aversion into two distinct components: risk aversion (preference for a less uncertain service time

duration) and transfer aversion (preference for continuous, rather than fragmented, multi-stage service processes). While risk aversion is well-documented in financial contexts (Holt and Laury 2002, Harrison and Cox 2008), customer behaviors in the presence of uncertain waiting times and fragmented service processes have received little attention in the behavioral literature (Allon and Kremer 2018). Thus, our first theoretical contribution is to document and characterize this important customer preference.

Continuing with Experiment 1, we show that algorithmic context may further affect chatbot uptake. The standard result in the literature is that AI assistance is often underutilized, even when AI performs as well as or better than a human alternative (Dietvorst et al. 2015, Logg et al. 2019, Castelo et al. 2019). We first follow the standard approach in the literature and conduct a series of pre-tests that hold the processes and performance constant across the two channels and vary only their labels and visual cues. These pre-tests do not detect any algorithm aversion, suggesting that the classic result that algorithmic errors loom larger than human errors does not hold in our setting. Nonetheless, we show that algorithm aversion still matters. Specifically, we show that gatekeeper aversion and algorithm aversion may interact to produce significantly lower chatbot uptake than can be explained by gatekeeper aversion alone, particularly when the stakes (waiting times in both channels) are high. Thus, our second theoretical contribution is to show that algorithm aversion can serve as an *amplifier*, reinforcing the reluctance to use a service channel with a gatekeeper structure.

In Experiment 2 we show that the aversions identified in Experiment 1 can be mitigated by varying how information is presented to customers. In particular, both transparency about chatbot capabilities and the average waiting time nudge can significantly increase chatbot adoption, although their effectiveness varies with time scale. Specifically, operational transparency matters when durations are short (suggesting that its effect is washed out when stakes are higher), whereas the effect of the nudge is more robust.

Table 1 Results Summary

	Effect	**Effect detected?**
	Experiment 1: Adoption Hurdles (§3)	
H1.1:	Transfer aversion } Gatekeeper Aversion	***
H1.2:	Risk aversion } Gatekeeper Aversion	***
Pre-tests of *Context* manipulation:	Algorithm aversion (Standalone effect)	n.s.
H1.3:	Algorithm aversion (Amplifying effect)	n.s. for short dur., ** for long dur.
	Experiment 2: Remedies (§4)	
H2.1:	Average Waiting Time Nudge	***
H2.2:	Transparency	** for short dur., n.s. for long dur.
	Experiment 3: Alternative Measurements of Algorithm Aversion (§5)	
H3.1:	Algorithm aversion (*Context* + *Live*)	**
H3.2:	Algorithm aversion (*Context* + *Hold*)	***

Note: Asterisks denote $p-$values after Bonferroni-Holm multiple hypothesis adjustment, following Holm (1979) and List et al. (2019). Effect sizes and standard errors are based on specifications in Tables 3, 4, and 5. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$, n.s. = not significant.

These results of Experiments 1 and 2 suggest practical ways to increase chatbot adoption: through fast-tracking chatbot customers by shortening their wait for a live agent after chatbot failure, through communicating operational advantages via explicit, time-based metrics and through providing a candid account of chatbot capabilities, rather than attempting to handle all inquiries without disclosure. In §4.4 we use a structural estimation of utility parameters to build a counterfactual model of channel-joining behavior and show that our interventions achieve substantial staffing cost savings (of up to 19.7%) in moderately congested systems. Thus, a practical contribution of our study is to identify inexpensive and easily implementable service design interventions that can increase chatbot adoption and generate substantial cost savings.

In the third experiment we show that increasing the realism of the interactions produces behaviors that are consistent with our original design (with similar aversion magnitudes). However, greater realism introduces a small but significant increase in algorithm aversion compared to the *Context* treatment. Thus, we contribute to the methodological discourse on measuring algorithmic attitudes by showing that experimental designs that rely on contextual framing alone may underestimate algorithm aversion, compared to designs that involve longer interactions or vary the human versus algorithmic nature of the interaction in a more realistic manner.

2. Retrospective Survey, Literature and Experiment Design

To motivate our model and experimental approach, we first report the results of a retrospective survey about real-life chatbot usage and experiences. We then introduce the model, review how the existing literature addresses the problem, and close by describing our experimental approach.

2.1. Retrospective Survey

To better understand key decision trade-offs faced by customers when choosing between chatbots and live agents, we conducted two waves of a retrospective survey on Prolific ($N = 400$, see Appendix EC.1 for details). Participants were asked to recall recent customer service interactions involving either a chatbot or a live agent, including details about wait times, issue resolution success, and overall satisfaction. The survey revealed significant differences in wait times to access customer service: about 77%–79% of chatbot users reported waits under one minute, compared to only 24%–33% for live agent interactions. However, chatbots resolved customer requests far less reliably, with success rates ranging between 34%–42%, compared to approximately 79%–87% for live agents. Overall satisfaction ratings were consistently higher for live agents (3.1 out of 5) relative to chatbots (2.2 out of 5). These results suggest the following. First, despite technological advancements, chatbot customer service experiences continue to be rated more poorly relative to live agent experiences. Understanding and mitigating adoption hurdles of chatbot technology thus continues to be an important practical concern for service managers. Second, the survey data highlight a key trade-off in customer channel choice: chatbots offer minimal wait times but fail frequently, while live agents require longer waits yet resolve requests more reliably. These insights guide both our model of channel choice and our experimental design.

2.2. Stylized Model of Channel Choice

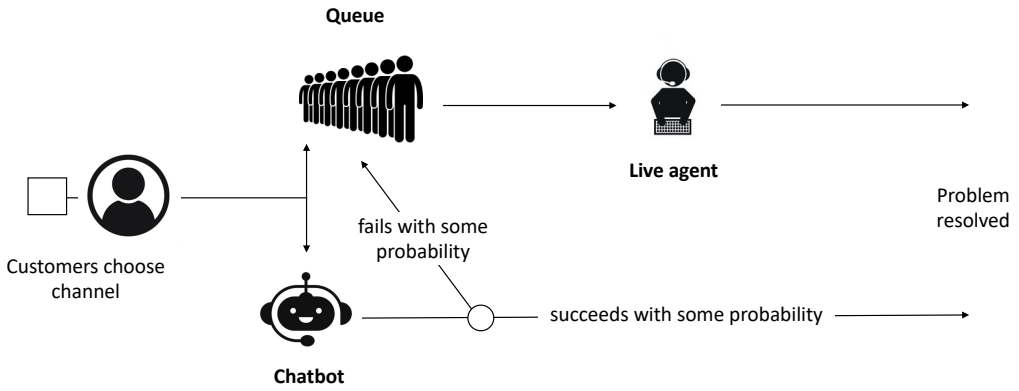
Building on our survey findings (§2.1), we model chatbot request resolution as a gatekeeper service process (Shumsky and Pinker 2003, Freeman et al. 2017, Hathaway et al. 2022) – a multi-stage process with an imperfect initial stage and a potential second stage with an expert service provider. Figure 1 illustrates the choice between a single-stage live agent channel and a two-stage chatbot channel, where the chatbot acts as a gatekeeper. Consider first the live agent channel. The customer must wait in line, after which the server resolves the request with certainty, and the customer exits.[1] Next, consider the chatbot channel. There is no queue, so the customer proceeds immediately to the chatbot interaction. Because the chatbot's problem-solving skills are limited, it only succeeds probabilistically. If the chatbot succeeds, the customer exits immediately. If it fails, then the customer waits in line before being served by a live agent, after which the request is resolved and the customer exits. Notably, the duration of the wait in line can be channel-dependent, and in our experiments and structural model we will consider priority queue designs that give chatbot customers a priority bump.

2.3. Related Literature

Standard economic reasoning used in traditional queue joining models (see Naor 1969, Allon and Kremer 2018) suggests that, when presented with a choice like the one in Figure 1, customers will select the channel that minimizes the expected time they would spend in the system. However, prior work in marketing, behavioral operations, and decision theory has identified several behaviors that suggest potential deviations from expected-time minimization in this setting. We discuss this work below.

[1] In practice, a small portion of requests handled by the live agent – between 13% and 21% based on our survey results – may be transferred to a second, expert agent. We do not examine such scenarios here to focus on the key trade-offs and to simplify choices for our experimental participants. However, such scenarios as well as other extensions of the decision problem in Figure 1 are studied analytically in Dada et al. (2025).

Figure 1 Channel Choice



Risk Aversion First, the channels in Figure 1 differ in the amount of risk they entail. While risk preferences for money have been extensively studied (Holt and Laury 2002, Eckel and Grossman 2008, Harrison and Cox 2008), relatively little is known about how individuals manage risk in the time domain. Some studies invoking Prospect Theory (Kahneman and Tversky 1979, Tversky and Kahneman 1992) suggest that because time expenditures are viewed as losses, individuals may be risk-neutral (Kroll and Vogt 2008) or even risk-seeking with respect to time (Abdellaoui and Kemel 2014). However, more frequently, research finds risk-averse behavior in the time domain (Leclerc et al. 1995, Festjens et al. 2015, Flicker and Hannigan 2022). Importantly, much of this work relies on hypothetical decisions and examines longer time intervals than those typical in customer service settings. In contrast, our experiments are incentivized, so that participants' choices have consequences for how they spend time in the experiment. Moreover, because higher stakes have been shown to amplify risk aversion in financial decisions (Holt and Laury 2002, Harrison and Cox 2008), we use two treatment arms in our design: one with shorter and one with longer waiting time durations. This will allow us to explore how choices evolve as the stakes increase.

Transfer Aversion Multi-stage waiting experiences were studied by Carmon and Kahneman (1996), Kumar et al. (1997) and Kumar and Dada (2021). These papers show that customer satisfaction depends not only on the total time spent in the system but may fluctuate within and across waiting stages. Different from us, these studies focus on the affective response (self-reported in-process and ex-post satisfaction), while we follow the revealed-preference approach and study queue-joining behaviors. Soman and Shi (2003) show that the progression path to the goal (in our case, having one's service request resolved) can matter as much as the total time spent in the system. Buell (2021) shows that customers often exhibit last-place aversion in queues, suggesting that a focus on expected waiting times alone may oversimplify the waiting experience. Althenayyan et al. (2022) look at fairness perceptions and show that customers experience in-queue delays differently depending on the source of the delay. Importantly, none of these studies look at multi-stage processes where the number of service stages is uncertain (i.e., gatekeeper processes), or where the nature of the server (human or algorithmic) is varied, leaving open the question of whether and how these factors interact.

Algorithm Aversion The algorithm aversion literature focuses on settings such as 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), order picking (Sun et al. 2022), and recommendation settings (Yeomans et al. 2019). A common finding across many studies is that people exhibit aversion toward algorithms, particularly after seeing them err. However, some research also documents algorithm appreciation, where users prefer algorithmic advice, particularly for tasks perceived as objective and data-driven (Castelo et al. 2019, Longoni et al. 2019). Our approach differs from this literature in two ways. First,

while the literature focuses on worker behavior, we focus on customer decisions; it is therefore not obvious whether the existing findings will hold in our setting. Second, our survey evidence (§2.1) suggests a structural difference between channels: chatbot systems typically operate as gatekeeper systems, whereas live agent systems typically involve longer waits but only a single service stage. This is different from the standard experimental paradigm wherein algorithmic assistance is assumed to make the user better off (Prahl and Van Swol 2017, Yeomans et al. 2019). Given these contextual differences, we develop a novel experimental approach to measure algorithm aversion in settings such as ours, where not only the nature of the server (human or algorithmic) but also the service design may differ across decision alternatives. In particular, we measure algorithm aversion by keeping the service design constant and by varying whether contextual information is presented to (or withheld from) participants.

Chatbot Adoption in Services Research specifically examining chatbot adoption in services remains relatively limited and often emphasizes anthropomorphism – rendering the bot to be more human-like – as a strategy to increase adoption (Sheehan et al. 2020, Adam et al. 2021, Schanke et al. 2021). Anthropomorphism can boost engagement and satisfaction by creating more natural conversational interactions, but its effectiveness varies across individual preferences and cultural contexts (Benke et al. 2022, Luo et al. 2019). More relevant for us is Castelo et al. (2023), who experimentally show that customers dislike service bots because they perceive them as a way for the service provider to cut costs at the customers' expense. Their study focuses on examining trust and fairness perceptions towards chatbots. In contrast, we explore operational dimensions of chatbot adoption, such as waiting times and chatbot effectiveness within a gatekeeper system, a common framework used in the service design literature (Shumsky and Pinker 2003, Freeman et al. 2017, Hathaway et al. 2022). Our approach thus adds to the chatbot adoption literature by integrating the chatbot experience into an operational system, by studying factors that drive adoption decisions as well as by testing operational levers to increase adoption.

2.4. Experiment Design

Our experiments serve three goals. First, they examine the extent to which the constructs identified in prior literature (risk aversion, transfer aversion, algorithm aversion, see §2.3) affect choices in the customer support setting, and explore potential ways in which these constructs interact (Experiment 1). Second, they test several managerial interventions to increase chatbot uptake (Experiment 2). Third, they allow us to evaluate different methodological approaches to eliciting and measuring algorithm aversion in the customer service setting (Experiment 3).

All experiments: In all experiments, participants make a series of binary choices between the live agent channel (Channel A) and the bot channel (Channel B), as shown in Figure 1. The choice is repeated for a range of problem parameters. Depending on the parametrization, Channel A or Channel B is the channel that minimizes total expected time spent in the system. Participants are incentivized to report their preferences

truthfully by having to experience a subset of their choices (in real-time) before receiving their payments. Table 2 summarizes all three experiments, the hypotheses and the participant numbers.

Experiment 1: This experiment examines channel choices in three settings: (1) a contextualized setting in which the algorithmic nature of the chatbot is explicitly disclosed, (2) a neutral, context-free setting in which the live-agent/chatbot nature of the channel is not revealed, and (3) a context-free setting in which both channels present the decision-maker with a deterministic sequence of events, thus allowing us to isolate the role of risk preferences. More specifically, the experiment is organized as a between-subjects 3 (treatment) $\times 2$ (time scale) design with the following treatments:

- *Context* Treatment. This treatment presents participants with a contextualized choice between two formats: the "Live Agent" and the "Chatbot" format. The sequence of stages in each channel is shown in Figure 1.
- *No Context* Treatment. This treatment is analogous to the *Context* treatment but presents participants with a setting where the choice is between two unnamed waiting formats that are visually identical (but continue to differ in the sequence of waiting stages, as described in Figure 1).
- *No Context, Deterministic* Treatment. This treatment is analogous to the *No Context* treatment in that it does not present participants with any contextual information. However, different from the *No Context* treatment, Channel B is now deterministic. This will allow us to separately identify the presence of risk aversion and transfer aversion.

As noted earlier, we use two between-subject treatment arms: a treatment arm with short time durations and a treatment arm with longer time durations. The size of the stakes (in our case, waiting time durations) is a commonly examined dimension in a variety of experiments studying both financial and time-relted choices (See, for example, Holt and Laury 2002, Abdellaoui and Kemel 2014); we therefore included two time scales in our design. In the short duration conditions, the average duration of a stage is 20 seconds across decisions, while in the long duration conditions it is 40 seconds. This results in average total waits of 40 seconds (resp.: 80 seconds) in the short (resp.: long) treatment arm.

Experiment 2: In Experiment 2 we focus on the practical implementation challenges faced by firms deploying chatbots. We continue to use the *Context* treatment as our baseline for comparisons. Against that baseline, we examine the effects of two manipulations: one related to how the chatbot capabilities are presented to the user, and a second one, related to how the waiting times are displayed in each channel:

- *Context + No Transparency* treatment. In this treatment the chatbot always suggests a resolution to the issue, regardless of whether it is viable. This is in contrast to the *Context* treatment in which the chatbot is transparent about being able (or not) to resolve a given issue.
- *Context + Nudge* treatment. This treatment is analogous to the *Context* treatment but adds the expected waiting times (in line + in service) for each channel. The expected waiting times serve as a nudge for the decision-maker to choose the channel that minimizes expected waiting duration.

Table 2 Experiment Overview

Objectives	Treatments (Between-subject)		Treatment Description	No. of Subjects (Recruited/ Passed comprehension screening/ Passed consistency checks)
	Short time durations:	Long time durations:		
Experiment 1 (§3): What are key drivers of chatbot uptake in customer service?	*Context*	*Context*	Contextualized channel choice	270/ 252/ 207
	No Context	*No Context*	Context removed	238/ 227/ 183
	No Context, Deterministic	*No Context, Deterministic*	Context and uncertainty removed	263/ 253/ 207
	Short time durations:	Long time durations:		
Experiment 2 (§4): What can firms do to increase chatbot uptake?	*Context + No Transparency*	*Context + No Transparency*	Chatbot attempts all requests instead of admitting to not having a solution	271/ 254/ 213
	Context + Nudge	*Context + Nudge*	Added average waiting time information	268/ 252/ 214
	Short time durations:			
Experiment 3 (§5): How does the nature of the service process affect algorithm aversion?	*Context + Live*		Real-time chat with human agent	116/ 106/ 91
	Context + Hold		Channel-specific interaction mode	106/ 102/ 86

As before, we examine a treatment arm with short time durations and a treatment arm with longer durations. Most importantly, all decisions in the *Context + No Transparency* and *Context + Nudge* treatments are mathematically identical to the *Context* treatment, with the only difference being how the choices and the interactions are presented.

Experiment 3 The purpose of the third experiment is to take a broader, more methodological view of characterizing the service process by examining alternative ways to measure algorithm aversion in a controlled experimental setting. To do so, we run two treatments that make the qualitative differences between the chatbot and live agent channels more salient:

- *Context + Live* treatment. In this treatment the chatbot channel continues to use click-based prompts, while the live agent channel is staffed by a human research assistant.
- *Context + Hold* treatment. In this treatment the chatbot channel continues to use click-based prompts, while participants in the live agent channel must hold down a button to complete the service process.

Comparing chatbot uptake in these treatments to our baseline (*Context* treatment) allows us to test whether our initial experimental findings are robust to more realistic implementations of service interactions and helps broaden our methodological contributions.

3. Experiment 1: Adoption Hurdles

In Experiment 1 we focus on unpacking the drivers of the choice between a live agent service channel and an algorithmic chatbot.

3.1. Methodology

3.1.1. Participants, Pre-tests and Treatments A total of 771 participants were recruited on Prolific to participate in the experiment; 732 of them passed the screening questions, of which 597 passed the consistency checks. Additionally, 221 participants were recruited to participate in the pre-tests to the experiment, of which 213 passed all the checks. The pre-tests were designed to validate our *Context* manipulation and to position our findings within the broader behavioral literature on algorithm aversion. All participants were US-based with an approval rating of at least 98%, and were restricted to participating in one session only. All experiments were programmed in oTree (Chen et al. 2016).

As noted earlier, the experiment consisted of three treatment conditions: *Context*, *No Context*, and *No Context, Deterministic*. In addition, there were two treatment arms: one with short time durations and a second one with longer time durations (doubled times). Each participant was randomly assigned to one of the treatment arms (short/long) and to one of the treatment conditions within that arm. Figure 2 shows the sequence of screens in each treatment and introduces notation for the relevant parameters. In all conditions, waiting in line (t^A_{line}, t^B_{line}) was programmed to look the same. However, interactions between the participant and the server ($t^A_{serve}, t^B_{serve_1}, t^B_{serve_2}$) depended on the treatment. In the *Context* treatment, choices were contextualized. In the instructions and on choice screens, participants were explicitly told that they were choosing between a live agent and a chatbot. Further, the interaction with each type of server consisted of channel-specific prompts (Figure EC.1a). In this condition, we expect to see all three aversions: transfer aversion, risk aversion and algorithm aversion. In the *No Context* treatment conditions, choices were context-free and the interaction between participants and servers was programmed to look identical across channels (see Figure EC.1b). In this treatment, we expect to see transfer aversion and risk aversion, but no algorithm aversion. Finally, in the *No Context, Deterministic* treatment, the participant always experienced both service stages in Channel B whereas in the other two, the participant experiences one service stage with probability p^B and two with probability $1 - p^B$. However, the durations of the service stages were adjusted such that the expected times in each decision were identical to the corresponding decision in the non-deterministic treatments. Given that both context and uncertainty is removed, in this treatment we expect to see transfer aversion only.

3.1.2. Instructions and Demo Figure 3 describes the experiment protocol. After being randomly assigned to a treatment, reading the instructions and answering comprehension questions, participants experience a demo of both channels. The Channel A demo is parametrized with $t^A_{line} = t^A_{serve} = 20$ seconds for short scale treatments and 40 seconds for long scale treatments. The Channel B demo follows and includes both the successful and failed chatbot resolution scenarios, with $t^B_{serve_1} = t^B_{line} = t^B_{serve_2} = 20$ seconds for short (40 seconds for long). The demo is visually representative of the actual experience in each channel and thus differs by treatment. Screenshots of the interface are in Appendix EC.3.

Figure 2 Flow of Waiting Stages by Treatment

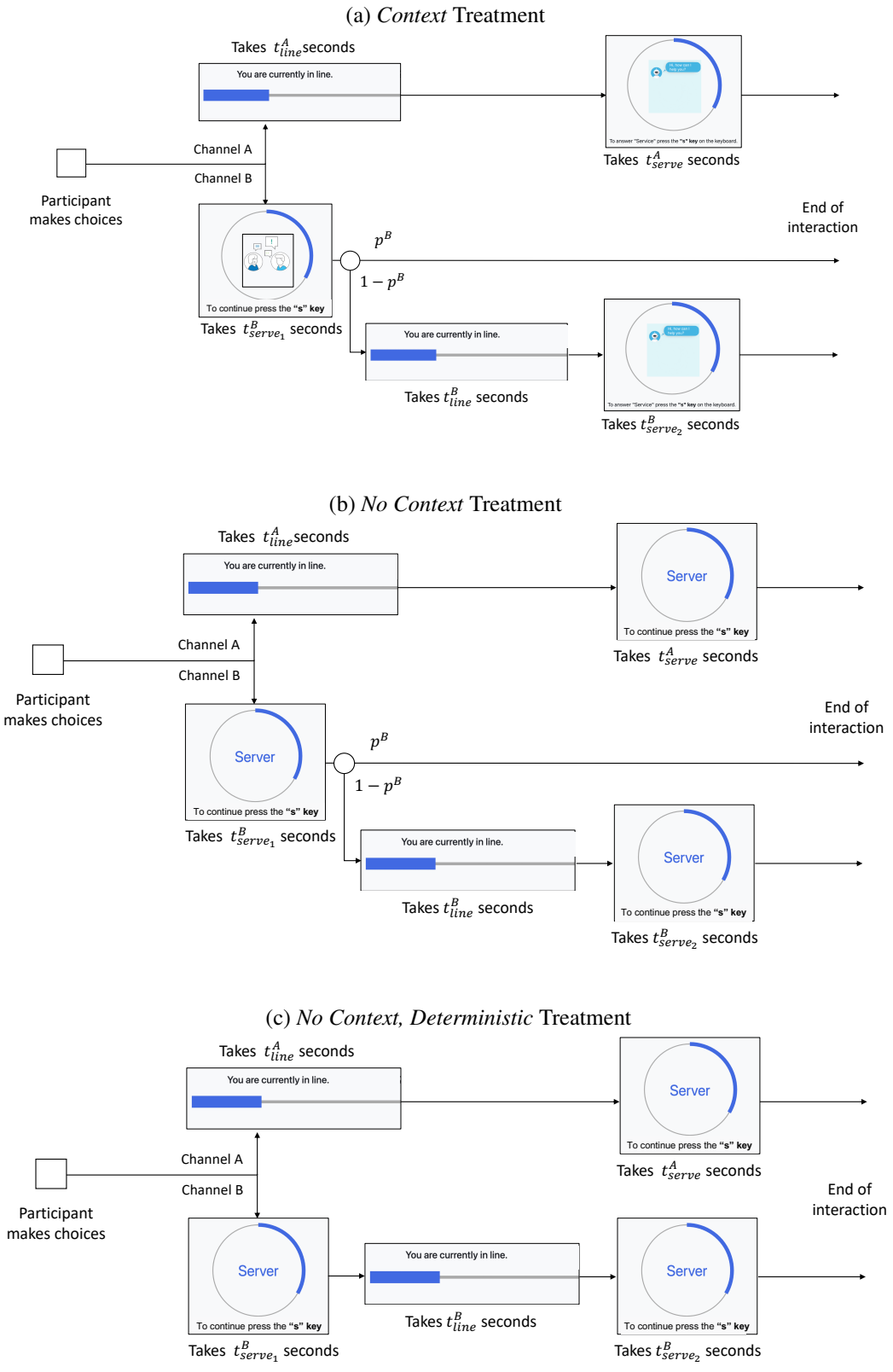
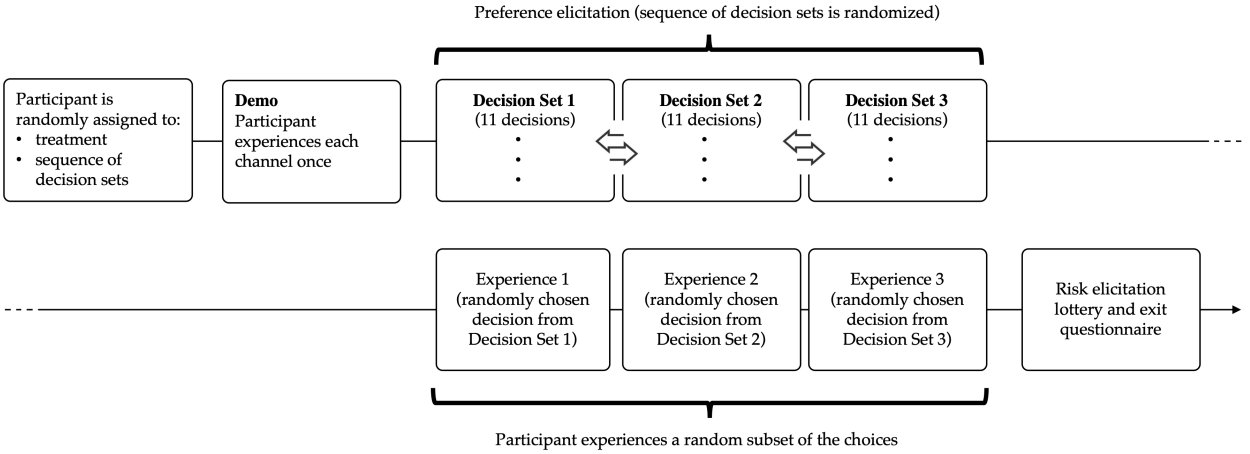


Figure 3 Experiment Protocol



3.1.3. Decisions and Parameters After the demo, participants make a total of 33 decisions, subdivided into three decision sets of 11 decisions. The sequence of decision sets was randomized to control for any order effects. Each decision is a binary choice between Channel A and Channel B. The chatbot capability p^B, the chatbot service time $t^B_{serve_1}$, and the waiting time after potential chatbot failure t^B_{line} differ across the 33 decisions. In particular, p^B ranges from 0.25 to 0.75 in increments of 0.05. In short duration conditions, $t^B_{serve_1}$ ranges from 10 to 30 seconds in 2-second increments, and t^B_{line} ranges from 0 to 40 seconds in 4-second increments. Long duration conditions double these time parameters: $t^B_{serve_1}$ ranges from 20 to 60 seconds, and t^B_{line} from 0 to 80 seconds. Within each decision set, Channel A minimizes expected waiting times in the first five decisions, Channel B minimizes expected waiting times in the last five decisions, and both channels yield identical expected waiting times in the sixth decision. Complete parameter listings for all 33 decisions are in Table EC.2 for short durations, Table EC.3 for long durations, and Table EC.4 for the *No Context, Deterministic* treatment.

3.1.4. Elicitation Within each decision set, we used the Multiple Price Lottery mechanism (Holt and Laury 2002) to elicit preferences. The basic idea of this mechanism is to present participants with a list of binary decisions, where one of the alternatives becomes more desirable as one goes down the list. For example, in Decision Set 1 we varied the success rate of the chatbot (p^B) in steps of 0.05 from 0.25 to 0.75 percent, while all other parameters were held constant. 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 between channels as Decision 1 in Decision Set 2 and Decision 1 in Decision Set 3, and similarly for the remaining 10 decisions).

3.1.5. 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. The average time spent in the experiment was 18 (resp.: 23) minutes in the short (resp.: long) time conditions. The average payment was \$6.65 (resp.: \$8.14).[2]

3.2. Hypotheses

Our hypothesis testing approach is summarized in Table 3. Before testing our hypotheses we will perform a series of pre-tests in which Channels A and B will have identical performance and sequence of stages and will only differ in the visual cues. The pre-tests will help identify algorithm aversion in the absence of process or performance-related differences. Subsequently, we will test for the presence of each type of aversion by examining the differences in mean Channel B uptake across treatments.

Table 3 Experiment 1: Hypothesis Testing Approach

	Treatment		
Pre-tests	*No Context, Deterministic*	*No Context*	*Context*
Algorithm aversion	Transfer aversion (H1.1)	Transfer aversion + Risk aversion (H1.2)	Transfer aversion + Risk aversion + Algorithm aversion (H1.3)

Consider first the *No Context, Deterministic* treatment. In this treatment channels A and B are visually identical, and there is no risk in either channel. Therefore, transfer aversion is the sole mechanism that may drive potential deviation from expected-time minimization. While the literature on multi-stage waiting experiences is limited (see §2.3), some work suggests that departures from a single-stage service process can lower satisfaction (e.g., Soman and Shi 2003, Kumar and Dada 2021). Therefore, we hypothesize:

H1.1 (Transfer Aversion): Average Channel B uptake in the *No Context, Deterministic* treatment is below 5.5 (risk-neutral theory prediction).[3]

Recall that in the *No Context* treatment Channel B entails risk, while in the *No Context, Deterministic* treatment, neither channel is risky. Further, in both of these treatments, channels A and B are visually identical and include no contextual cues. Therefore, a treatment comparison between the *No Context* and the *No Context, Deterministic* treatments will isolate risk aversion.

[2] In addition to the main task, for which participants received a fixed participation payment, 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.

[3] Recall that there are 11 decisions in each decision set. Given the parameterizations in the experiment, expected time minimization predicts that a decision-maker switches from Channel A to Channel B in decision 6 where both channels offer the same. If we assume random tie breaks, 5.5 is the theory prediction. However, our results are robust to alternative tie-breaking procedures for the sixth decision, i.e., to using 5 out of 11 or 6 out of 11 as our theory benchmark. See Tables EC.2-EC.4 for full listings of parameters in all treatments.

H1.2 (Risk Aversion): Conditional on expected waiting times in each channel, participants in the *No Context* treatment choose Channel B at a lower rate than in the *No Context, Deterministic* treatment.

Our *No Context* and *Context* treatments differ only in how the channels are presented, i.e., whether the algorithmic nature of Channel B is made salient. Therefore, a comparison of these two treatments will allow us to isolate algorithmic aversion.

H1.3 (Algorithm Aversion): Conditional on expected waiting times in each channel, participants in the *Context* treatment choose Channel B at a lower rate than in the *No Context* treatment.

To test H1.1 we perform one sample $t-$tests, comparing empirically observed chatbot uptake against 5.5, the theoretically optimal uptake under expected-time minimization. To test H1.2-H1.3, we perform random effects logit regressions and examine the significance of treatment coefficients. Because our hypotheses make a directional prediction, we will report one-sided $p-$values adjusted for multiple hypothesis testing using the Bonferroni-Holm procedure for each family of hypotheses (Holm 1979, List et al. 2019).

3.3. Results

We begin by reporting the results of two pre-tests designed to validate our *Context* manipulation and to better position our findings within the broader behavioral literature on algorithm aversion. After reporting the results of these pre-tests, we present descriptive statistics and formal hypothesis tests.

3.3.1. Pre-tests The first pre-test ($N = 113$) examined whether participants had any intrinsic preference for (or bias against) the visual stimuli used to represent the human or the algorithmic channel. Participants experienced two interactions: a 20-second interaction with a chatbot and a 20-second interaction with a human agent, presented in randomized order. (See Figures EC.1 b and c for theses visual stimuli, which are also used in the *Context* treatments.) Participants then selected one provider (chatbot or live agent) for an additional 40-second interaction; this choice served as our outcome variable. A total of 56 of 113 participants chose the live agent (49.56%), which is not significantly different than 50% (Proportion test, $p = 0.904$). Thus, neither set of visual stimuli produced a bias towards one channel over the other.

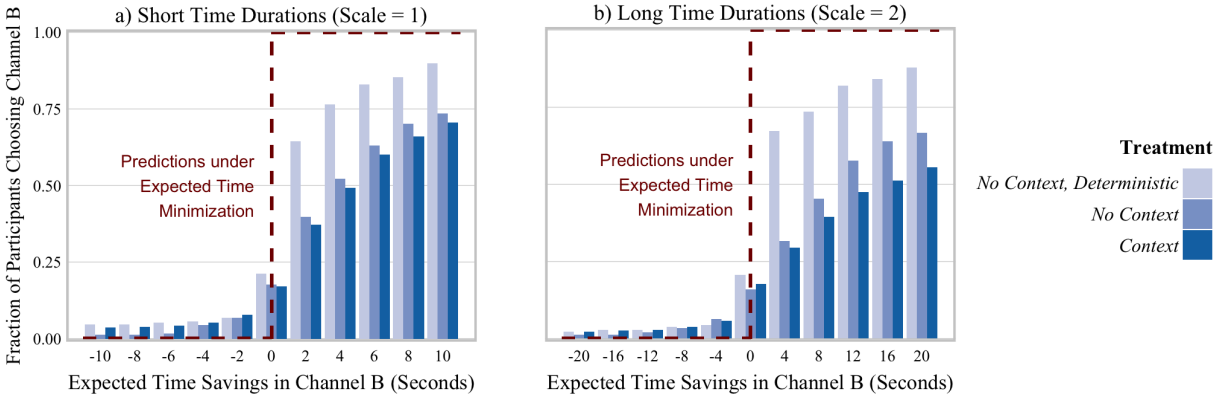
The second pre-test ($N = 100$) was similar in that each participant made a single decision after experiencing each channel once. However, now participants chose between two identical *gatekeeper* processes, each offering a 50% chance of immediate resolution following an initial 40-second interaction, and a 50% chance of requiring an additional 40-second wait in line plus a 40-second interaction with a second-stage human agent. The only difference between the two channels was the labeling and visuals of the initial server as either live agent or chatbot. Results again revealed no significant preference for either channel: 49 out of 100 participants chose the live agent (49%), vs. 51 participants (51%) chose the chatbot (Proportion test, $p = 0.920$). Thus, our experimental stimuli do not produce a detectable bias towards humans or algorithms

when performance (error rates) and processes are held constant between human and algorithmic alternatives. However, as will become clear later in this section, algorithm aversion can still play an important role as an *amplifier*, reinforcing gatekeeper aversion and further reducing chatbot adoption when the chatbot is associated with existing process and performance differences.

3.3.2. Descriptive Statistics Figure 4 shows the share of participants choosing Channel B (representing the chatbot) in each of the six conditions, conditional on the expected time difference between the two channels. Within each decision set, Channel B becomes increasingly more attractive for higher-numbered decisions, with Decision 6 marking the indifference point. Several observations are in order. First, in all six conditions, chatbot uptake is substantially below the expected-time minimization benchmark. This is particularly visible in the right half of each graph (where time savings in Channel B are positive), with a gap ranging between 10 and 70 percentage points. Further, much of this behavior appears to be tied to the process-related features of the choice, with a substantial Channel B underutilization even in the absence of context. Indeed, adding context further reduces chatbot uptake, but it does so only by a small margin (0 to 10 percentage points), with a more pronounced change in panel b. These observations offer some preliminary support for hypotheses H1.1-H1.2 but suggest that H1.3, i.e., algorithm aversion, may only be observed for the longer duration conditions, i.e., when the stakes are higher. Finally, comparing panel a and panel b, longer time durations decrease chatbot uptake in all three treatments (with the difference being particularly large for the *No Context* and *Context* conditions). While not part of our formal hypothesis development, we will nonetheless test the significance of this difference in post-hoc analysis (§3.3.4).

3.3.3. Hypothesis Tests We next test H1.1 - H1.3, i.e., examine whether chatbot uptake is below what expected-time minimization would predict, and whether it is affected by the presence of context. Because we are simultaneously testing three hypotheses, it is appropriate to adjust the reported significance levels for multiple hypothesis testing. The description of the Holm-Bonferroni adjustment procedure as well as both the raw (unadjusted) and the adjusted $p-$values are included in Table EC.5 in the Appendix.

Figure 4 Channel B (Chatbot Channel) Uptake in Experiment 1



We begin by testing H1.1, i.e., the presence of transfer aversion. To test H1.1 we use $t-$tests and compare observed channel uptake with 5.5 (out of 11), the expected-time minimization benchmark. Our data strongly reject H1.1: average Channel A uptake is 4.47 (resp.: 4.32) in the short (resp.: long) time durations condition; both values are significantly smaller than 5.5, $p \ll 0.001$. To test H1.2 we use random effects logit regressions. The regression coefficients are reported in Table 4. Column 1 presents the full data set. We use the *No Context* treatment as the omitted variable, because this treatment is used for comparisons in both H1.2 and H1.3. Note first that the difference between the *No Context* and *No Context, Deterministic* treatments is statistically significant ($p \ll 0.001$). The absence of uncertainty significantly increases Channel B uptake. However, the difference between the *Context* and *No Context* treatments is not statistically significant ($p = 0.179$). Columns 2-3 replicate the analysis, but focus on either the short durations conditions (col. 2) or the long durations conditions (col. 3). The same pattern of results emerges, with the difference being that the effect sizes are somewhat higher for the long time durations. Indeed, the *Context* treatment dummy is significant at $p = 0.046$ in col. 3, suggesting that algorithm aversion is present for longer time durations.

Result 1: *H1.1-H1.2 are supported. We find evidence for both transfer aversion and risk aversion. H1.3 is partially supported. Adding customer service context significantly reduces chatbot uptake for long, but not for short time durations.*

3.3.4. Additional Analysis Most lab experiments, and ours, involve only a small number of relatively short interactions, which can limit external validity. To address this concern, we performed three sets

Table 4 Channel Preferences in Experiment 1

Dependent Variable:	(1) Channel B (Chatbot Channel)	(2) Channel B (Chatbot Channel)	(3) Channel B (Chatbot Channel)
No context Treatment	Omitted category	Omitted category	Omitted category
No context, deterministic Treatment	2.134***	1.800***	2.585***
	(0.401)	(0.522)	(0.668)
Context Treatment	-0.560	-0.054	-1.116**
	(0.416)	(0.545)	(0.661)
Time scale $= 2$	-0.821**		
	(0.323)		
Channel Performance Controls	Yes	Yes	Yes
Demographic Controls	Yes	Yes	Yes
Sample	Full Sample	Time Scale $= 1$	Time Scale $= 2$
Observations	19701	9504	10197
Subjects	597	288	309

Notes: Random effects logit regression coefficients are reported. Dependent variable is channel choice (Channel B $= 1$). Standard errors are clustered at subject level. Decision set number and decision number within the decision set are controlled for. The following demographic variables are controlled for: age, gender, number of quiz errors and the Eckel-Grossman risk aversion measure (administered after the main task). H1 tests are one-sided, with $p-$values adjusted for multiple hypothesis testing using Bonferroni-Holm procedure. The remaining tests are two-sided. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.1$.

of supplementary analyses. First, we check for potential learning effects. To do so, we verify that the decision set variable is not statistically significant. Indeed, based on regression specifications in Table 4, we find no evidence of learning, with all decision set coefficients being well above the significance threshold ($p > 0.256$). Second, we examine potential time scale effects. In particular, Figure 4 suggested that all three aversions increase with longer time durations. Column (1) in Table 4 confirms that this effect is significant: the scale effect is negative and statistically significant at $p = 0.011$, with chatbot adoption decreasing by an average 4.4 percentage points when durations are long. Thus, the duration of the interactions is an additional contributor to the willingness (or reluctance) to engage with chatbot technology. Finally, we compare participant behavior in the experiment with their responses to the post-experimental questions regarding everyday chatbot use. We find that participants who reported prior use of chatbots outside of the experiment showed significantly higher Channel B uptake in the experiment, with an increase between 13% and 26% relative to those participants who have had no prior experience with chatbots (rank sum tests $p < 0.01$).

3.4. Discussion

The results of Experiment 1 offer several interesting insights. As expected, participants respond positively to improved chatbot performance and choose Channel B more frequently as its time (in line and in service) becomes shorter and as the chatbot's probability of success increases. At the same time, participants' choices deviate significantly from expected-time minimization: the majority of participants frequently chose the channel with longer expected waiting times, suggesting a significant aversion to gatekeeper systems both with and without uncertainty and both in the presence and in the absence of contextual cues.

In our pre-test (§3.3.1) we saw that context alone did not significantly drive channel preferences. Similarly, context had only minimal effects in our main experiment, when durations were short. However, when we increased the stakes (doubling the time durations in both channels), we observed significant algorithm aversion in addition to gatekeeper aversion. Thus, algorithm aversion can amplify peoples' reluctance to use a gatekeeper channel. From a theory standpoint, this suggests that algorithmic attitudes are malleable and can interact with structural biases around the gatekeeper process itself. This is different from the classic result that people are more sensitive to algorithmic than to human errors even when humans and algorithms perform equally well (Dietvorst et al. 2015). If chatbots performed at the same level as live agents, our results suggest that their uptake will be closer to rational theory predictions. However, as long as chatbots remain limited in their capabilities (as is currently the case in practice; see §2.1), customers will continue to underutilize them – both due to their aversion to gatekeeper processes and because chatbots have become emblematic of that very experience.

From a more practical standpoint, the result that algorithmic and process-related hurdles can be mutually reinforcing suggests that firms need to consider chatbots as part of a larger, multi-channel service design problem instead of treating chatbots as a standalone AI issue. Building on this idea, in the next section we test two practical remedies aimed at increasing chatbot uptake.

4. Experiment 2: Remedies

We have so far seen that the bulk of Channel B (chatbot) underutilization is tied to the gatekeeper structure of the chatbot channel. We have also seen that chatbot uptake may be further reduced for request types that involve longer durations. In this section, we propose and test two managerial levers to counteract these adoption hurdles.

4.1. Treatments

Experiment 2 introduces two new treatments. As in Experiment 1, each treatment consists of two between-subject conditions: one with short and one with long (doubled) time durations. We continue to use the *Context* treatment from Experiment 1 as our baseline for comparisons. The new treatments are the *Context + No Transparency* and the *Context + Nudge* treatment. In the *Context + No Transparency* treatment the chatbot always suggests a solution to the request, regardless of whether it is able to resolve it successfully. This is in contrast to the *Context* treatment, in which the chatbot is transparent about its capabilities and simply reports not being able to resolve an issue if this happens to be the case. See Figure EC.2 for the relevant screens displayed in each treatment. In the *Context + Nudge* treatment we provide participants with expected waiting times for each channel and decision (see Figure EC.3). The parameters as well as the remaining prompts and instructions are unchanged relative to the *Context* treatment. We recruited 539 participants for the new treatments via Prolific, of whom 427 participants passed all comprehension checks and consistency checks.

4.2. Theory and Hypotheses

To develop hypotheses we leverage the rich behavioral literatures in economics and operations. The *Context + No Transparency* treatment is inspired by prior work on operational transparency, which has consistently shown that revealing the processes underlying service delivery increases trust and perceived value (Buell and Norton 2011, Buell et al. 2017). Balakrishnan et al. (2022) show that feature transparency reduces algorithm aversion in forecasting tasks. Somewhat different from this literature, we focus on outcome (as opposed to process) transparency. We expect that being transparent about the ability of the chatbot to resolve a given request may increase trust and thus increase chatbot uptake. Specifically, we compare a transparent chatbot that explicitly communicates its capabilities and limitations versus one that attempts to handle all requests without such disclosure. Formally, we hypothesize the following:

H2.1 (Transparency): Conditional on expected times in both channels, participants in the *Context + No Transparency* treatment choose Channel B at a lower rate than in the *Context* treatment.

Our second intervention (*Context + Nudge*) leverages insights from behavioral economics about how people process complex choices under uncertainty. In particular, decision-makers often fail to optimally integrate outcomes and probabilities, instead focusing on particularly salient aspects of the choice (Arieli

et al. 2011, Aimone et al. 2016a,b). The *Context + Nudge* intervention is aimed at directing decision-makers' attention towards objective performance metrics (expected times) and away from format preferences or channel biases. This intervention thus offers a lightweight "nudge" that could increase the share of people choosing the chatbot option, when this option helps them save time (in expectation). Formally, we test the following hypothesis:

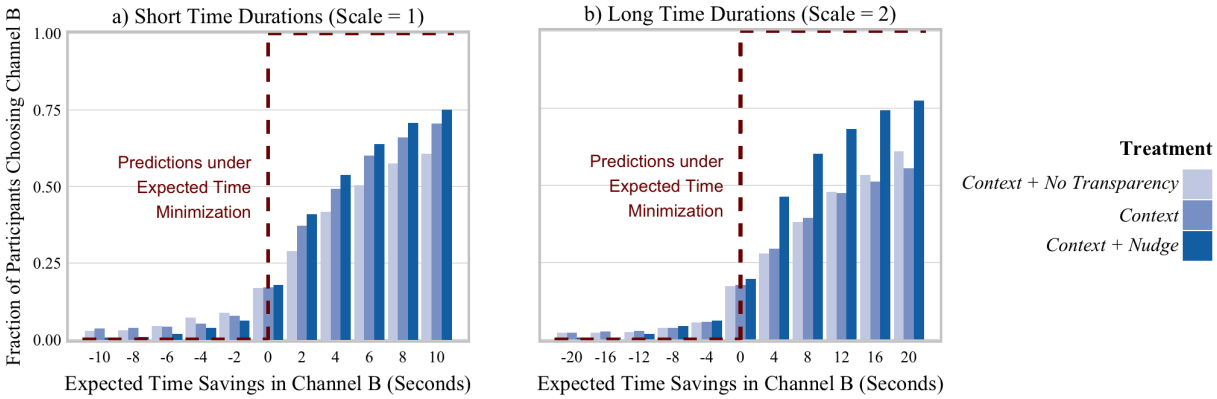
H2.2 (Nudge): Conditional on expected times in both channels, participants in the *Context* treatment choose Channel B at a lower rate than in the *Context + Nudge* treatment.

4.3. Results

As before, we present average chatbot uptake by treatment, and then test our hypotheses using random effects logit regressions.

4.3.1. Descriptive Statistics Figure 5 shows chatbot uptake in each of the six conditions and in each of the 11 decisions. First, relative to the *Context + No Transparency* treatment, transparency (*Context* treatment) appears to increase chatbot uptake for short time durations (panel a), with increases between five and ten percentage points in the right half of the graph. This is consistent with H2.1. However, transparency appears to have only a minimal effect in the long time durations conditions. Second, consistent with H2.2, the nudge appears to increase chatbot uptake in both treatments, though the increase is quite small under short time durations (at most five percentage points) and quite strong under longer time durations (up to 20 percentage points). Further, note that the scale effect (reduced chatbot uptake for longer time durations) observed in Experiment 1, while still present on average, does not hold across all treatments. In particular, comparing the dark blue bars in panel a) with those in panel b), we observe that there is no discernible scale effect in the *Context + Nudge* condition. This suggests that the effect of the nudge is quite dominant in focusing participant attention on the decision-theoretic fundamentals of the choice, bringing their decisions closer to the theoretic predictions.

Figure 5 Channel B (Chatbot Channel) Uptake in Experiment 2



4.3.2. Hypothesis Tests To test H2, we regressed channel choice on the treatment dummies. Table 5 presents the estimates. As before, in Table 5 we report one-sided, Bonferroni-Holm adjusted $p-$values. Table EC.6 presents details of the adjustment procedure. In column (1) we present test results for the pooled sample. *Context* treatment is used as the baseline (omitted category). Several observations are in order. First, transparency appears to have little effect on chatbot uptake in the aggregate: as predicted, the *Context + No Transparency* treatment coefficient is negative but the effect is not statistically significant ($p = 0.138$). Second, the effect of the nudge is positive and statistically significant ($p = 0.008$). In column (2) we focus on short time durations. Here, the effects of transparency are statistically significant: being transparent about chatbot capabilities significantly increases chatbot uptake relative to the condition where the chatbot always produces a solution ($p = 0.019$). Further, the effect of the nudge is not statistically significant ($p = 0.669$). Finally, column (3) focuses on the long durations conditions and shows that the effect of transparency is not statistically significant ($p = 0.664$), while the nudge significantly increases uptake ($p \ll 0.001$).

Result 2: *H2.1 is partially supported. Chatbot channel uptake is increased under operational transparency, but the effect is only observed when time durations are short. H2.2 is supported. The nudge helps increase chatbot uptake.*

Table 5 Channel Preferences in Experiment 2

Dependent Variable:	(1) Channel B (Chatbot Channel)	(2) Channel B (Chatbot Channel)	(3) Channel B (Chatbot Channel)
Context Treatment	Omitted category	Omitted category	Omitted category
Context + *No Transparency* Treatment	-0.412	-1.078**	0.225
	(0.378)	(0.521)	(0.529)
Context + *Nudge* Treatment	0.883***	-0.221	2.049***
	(0.364)	(0.506)	(0.523)
Time scale $= 2$	-0.463		
	(0.302)		
Channel Performance Controls	Yes	Yes	Yes
Demographic Controls	Yes	Yes	Yes
Sample	Full Sample	Time Scale $= 1$	Time Scale $= 2$
Observations	20922	10428	10494
Subjects	634	316	318

Notes: Random effects logit regression coefficients are reported. Dependent variable is the channel choice (Channel B $= 1$). Standard errors are clustered at subject level. Decision set number and decision number within the decision set are controlled for. The following demographic variables are controlled for: age, gender, number of quiz errors and the Eckel-Grossman risk aversion measure (administered after the main task). H2 tests are one-sided, with $p-$values adjusted for multiple hypothesis testing using Bonferroni-Holm procedure. The remaining tests are two-sided. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.1$.

4.3.3. Discussion In Experiment 2, we tested the effectiveness of two levers that managers can use to increase chatbot uptake: providing transparency about the chatbot's capabilities and highlighting potential

time savings offered by the chatbot channel. Both remedies yield positive results. However, the effectiveness of each intervention varies with the duration of the interactions. When waiting times are short, participants are already choosing the chatbot relatively frequently, reducing the incremental benefit of emphasizing time savings. Under these shorter time horizons, participants appear more attuned to how the chatbot operates. Thus, being transparent about what the chatbot can and cannot do helps increase adoption. In contrast, when waiting times are long, participants are more reluctant to select the chatbot. In this scenario, highlighting the time-saving advantage is more compelling than additional transparency about capabilities. More broadly, the observed behaviors suggest that when the consequences of decisions are small, decision-makers focus on channel presentation and appearance of the channels. However, as waiting times increase and stakes are higher, decision-makers shift attention away from interactive features and toward efficiency considerations. Therefore, the relative effectiveness of different channel designs may depend on the time horizon.

4.4. Structural Estimation of Staffing Cost Savings

To quantify the potential operational benefits of our experimental manipulations (queue time durations, transparency, nudge) we focus on staffing – a key driver of controllable costs that motivated the deployment of customer service chatbots in the first place. We do this in three steps. We first formulate and estimate a random utility model of the customer choice between the live agent and chatbot channels. We then use the model estimates to predict customer arrival rates to each channel under various conditions. Finally, we calculate staffing costs based on the staffing levels necessary to maintain promised waiting times in each channel. This approach allows us to evaluate a range of counterfactual scenarios and estimate potential cost savings by explicitly accounting for endogenous customer channel selection.

4.4.1. Structural Estimation of Utility Parameters To accurately predict channel arrival rates, we first need to estimate a customer utility function. In EC.4, we consider several plausible candidate utility functions. As is typical in structural estimation, increasing the number of parameters generally improves model fit but can reduce interpretability and intuition. Balancing this trade-off, we ultimately select a model that captures the aversion to using Channel B through a simple linear specification:

$$U^A_{ij}(\boldsymbol{\theta}) = r - c_{line} \cdot t^A_{line_{ij}} - c_{agent} \cdot t^A_{serve_{1ij}} + \epsilon^A_{ij}, \tag{4.1}$$

$$U^B_{ij}(\boldsymbol{\theta}) = r - c_{bot} \cdot t^B_{serve_{1ij}} - (1 - p^B_{ij}) \cdot (c_{nt} + \beta \cdot (c_{line} \cdot t^B_{line_{ij}} + c_{agent} \cdot t^B_{serve_{2ij}})) + \epsilon^B_{ij}, \tag{4.2}$$

where $U^A_{ij}(\boldsymbol{\theta})$ and $U^B_{ij}(\boldsymbol{\theta})$ represent the utilities from receiving service through Channel A and Channel B, respectively, for participant i in decision j. The parameter vector $\boldsymbol{\theta}$ includes the reward for service (r), waiting costs per second in line, with the agent, and with the chatbot ($c_{line}, c_{agent}, c_{bot}$), a lump-sum disutility if the chatbot fails and lacks transparency about its capabilities (c_{nt}), and a multiplier applied to the disutility from delays when the chatbot fails (β). This structure allows us to explicitly model how customers weigh different service channels based on expected waiting times and information availability. We normalize r

to zero and estimate $\boldsymbol{\theta}$ using maximum likelihood across the three treatments (*Context, Context + Nudge, Context + No Transparency*), setting c_{nt} to 0 in the *Context* and *Context + Nudge* treatments and estimating distinct β parameters for treatments with and without the nudge. All parameters are statistically significant, with estimates and bootstrapped standard errors reported in EC.4.

4.4.2. Demand Estimation To calculate staffing levels and costs for our counterfactual scenarios we need to know the arrival rates to each channel (demand intensities λ^A and λ^B). We make the standard assumption that customers arrive according to a Poisson arrival process, with demand intensity λ, which then splits into λ^A and λ^B based on the offered waiting times and success probabilities. In particular, the relative demand for each channel is formed according to the logit choice probabilities $\rho^A(\boldsymbol{\theta})$ and $\rho^B(\boldsymbol{\theta})$, which can be derived from equations (4.1) - (4.2) and the estimates in Table EC.9. 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 in which the waiting times initially promised (t^A_{line}, t^B_{line}) match the actual average waiting times $(\bar{t}_{line})$ arising from the endogenously determined arrival rates resulting from customer channel choices $(\rho^A(\boldsymbol{\theta}), \rho^B(\boldsymbol{\theta}))$. We provide an intuitive overview of this procedure. We first compute the choice probabilities $(\rho^A(\boldsymbol{\theta}), \rho^B(\boldsymbol{\theta}))$ for a range of counterfactual parameters, i.e., t^A_{line}, t^B_{line}, the presence (or absence) of the nudge and/or transparency, as well as the remaining system parameters. We then use these probabilities to calculate channel demand, where bot demand (λ^B) is simply the portion of total demand intensity λ that is directed to the chatbot, and live agent demand (λ^A) is made up of two components: the customers who choose the live agent channel and the customers who choose the chatbot channel but experience chatbot failure and are redirected to the live agent. Finally, by modeling the system as an $M/D/1$ queuing regime, we are able to derive the live agent service rate μ required to deliver the announced waiting times.[4] If we assume that staffing costs increase linearly in the service rate μ (proxy for staffing level), we can use our derivation to estimate staffing costs.

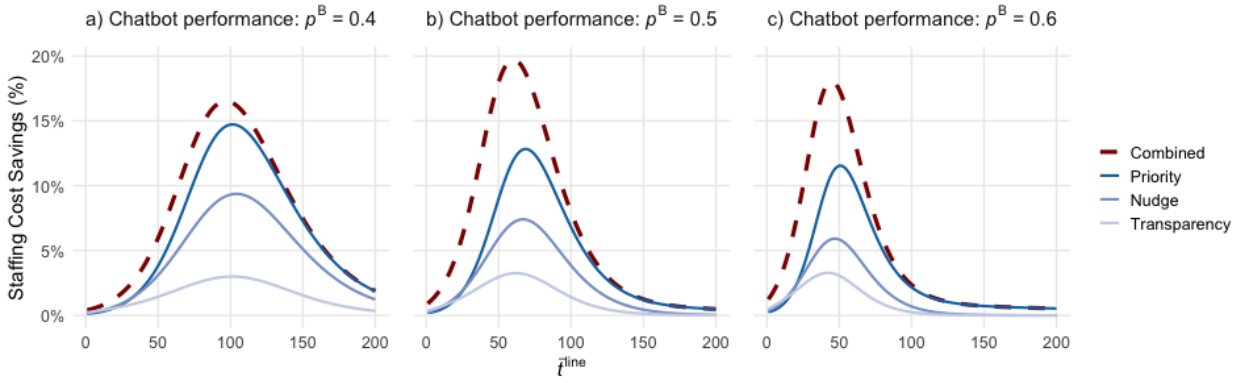
4.4.3. Results We estimate staffing costs in five scenarios. In the baseline service design the chatbot is not transparent, no nudge is applied and the queue for the live agent is pooled (where $t^A_{line} = t^B_{line}$). This is compared against four scenarios that apply either transparency, the nudge, a priority queue for chatbot

[4] We first compute the average sojourn time in the live agent channel, $T^A(\boldsymbol{\theta}) = \bar{t}_{line}(\boldsymbol{\theta}) + t^A_{serve}$, where $\bar{t}_{line}(\boldsymbol{\theta})$ is the average time that a customer spends in line waiting for the live agent, weighted by the proportion of the live agent demand coming from each channel. Then the live agent service rate μ required to deliver the announced waiting times can be calculated as follows (derived from Tijms 2003, p. 59): $\mu(\lambda^A(\boldsymbol{\theta}), T^A(\boldsymbol{\theta})) = \frac{\lambda^A(\boldsymbol{\theta}) + \sqrt{\lambda^A(\boldsymbol{\theta})^2 + \frac{2\cdot\lambda^A(\boldsymbol{\theta})}{T^A(\boldsymbol{\theta})}}}{2}$. The results are virtually identical if we use a $M/M/s$ system to model staffing costs.

users (where t^A_{line} is set lower than t^B_{line})[5], or all three interventions combined. Figure 6 shows the relative staffing cost savings for counterfactual scenarios defined by $\bar{t}^{line} \in [1, 200]$, and by $p^B = \{0.4, 0.5, 0.6\}$.[6]

Figure 6 offers several managerial insights. First, in all scenarios the cost savings peak at an interior value of $\bar{t}_{line}$. When $\bar{t}_{line}$ is sufficiently low, as would be the case in a low-congestion system, joining the live agent queue for quick, guaranteed resolution is so attractive that any intervention has little effect on demand, explaining the low cost savings. Conversely, when $\bar{t}_{line}$ is high, as would be the case in a highly congested system, any intervention has little effect as the bot is already perceived as an attractive option to avoid the long wait for the live server. It is only for intermediate values of $\bar{t}_{line}$ that the interventions have a strong enough effect on demand to substantially decrease staffing costs. Second, among the three interventions, prioritizing chatbot customers for access to live agents has the highest potential to reduce staffing needs, yielding cost savings of between 11.5% (in panel c) and 14.7% (in panel a). This suggests that operational efficiency gains from priority queues exceed those from information-based interventions. Third, the highest cost savings occur for chatbots of intermediate capability ($p^B = 0.5$), with up to 19.7% cost savings, suggesting that design interventions have their greatest impact when chatbots are neither too ineffective nor too advanced, representing an important transition zone where customer channel choice is most malleable.

Figure 6 Counterfactuals: Staffing Cost Savings



[5] To compare staffing costs with and without a priority queue for chatbot customers, we set $\bar{t}_{line}$ to be the same under each comparison between baseline and prioritization, but set t^A_{line} and t^B_{line} under prioritization such that the weighted average was equal to $\bar{t}_{line}$ and that $t^B_{line} = 0.9 \cdot t^A_{line}$. In other words, if the chatbot fails, the customer receives a 10 percentage point bump in the queue relative to a customer who immediately chose the agent channel. Giving strict priority to chatbot users would result in factors even lower than our chosen factor of 0.9. We chose a fairly conservative priority factor and held it constant across all scenarios to make like comparisons.

[6] Further assumptions on system parameters are as follows. We set λ to 0.1, resulting in a system utilization between 75% and 80% – a utilization level commonly used in queuing analysis of moderate-to-heavy traffic. We set all service times to 20. Finally, we set the unit staffing cost to 1 (i.e., the staffing cost is simply given by the equation in the previous footnote).

5. Experiment 3: Alternative Measurements of Algorithm Aversion

In Experiments 1 and 2 we examined key drivers of chatbot adoption and identified ways to overcome barriers to use. In both of those experiments, participants interacted with servers through identical click-based prompts across channels. This enabled precise control and measurement of interaction times, ensuring that the actual time spent matched the time promised when participants made their channel selections. Although this design provided strong experimental control, it necessarily reduced interaction realism. Recognizing the trade-off between experimental control and realism, in Experiment 3 we replicate our original design, but add more authentic representations of how users may experience interactions across different service channels.

5.1. Methodology

5.1.1. Treatments Experiment 3 consisted of two treatments, which we will refer to as the *Context + Live* treatment and the *Context + Hold* treatment. A total of 222 new subjects were recruited to participate in these treatments of whom 177 passed all comprehension and consistency checks. The training materials for the research assistants representing the live agents are in EC.3.3. As in Experiments 1 and 2, participants made 33 decisions, with three of these decisions being implemented after the decisions were submitted. As in Experiments 1 and 2, participants were shown a demo of each channel prior to making their decisions. Both treatments were conducted using the short duration parametrization (See Table EC.2 for parameters).

The *Context + Live* treatment was identical to the *Context* treatment in Experiment 1, with the only 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 (See EC.3.3 for the script). At the beginning of the service process, the participant indicated their assigned "issue type", after which the research assistant started the service process, which then lasted for the required duration ($t^B_{serve_2}$ seconds).

The *Context + Hold* treatment was identical to the *Context* treatment in Experiment 1, with the only difference being that the interaction mode differed between the channels. In particular, we made the human/algorithmic nature of the server more salient. In interacting with the server representing the live agent, participants were required to hold down a button for the duration of service. In contrast, in interacting with the server representing the chatbot, participants continued using keystrokes to interact, exactly like in Experiment 1. See EC.3.3, particularly, Fig. EC.4-EC.5 for the description of the experimental stimuli.[7]

[7] The hold vs. click-based interaction modes were chosen to represent a more continuous interaction with an agent vs. a more fragmented interaction with a bot. To ensure that these interaction modalities were valid representations of the channels, we performed a manipulation check (with a separate group of respondents), which confirmed that the manipulations were associated with the channel as intended. See Appendix EC.3.3 for details.

5.1.2. Theory and Hypotheses

In Experiments 1 and 2, we followed an approach common in the experimental literature and examined how contextual information, i.e., variations in instructions and visuals, affects chatbot uptake. However, context is only one part of understanding algorithm aversion, particularly in customer service, where *experiential* factors also play a role. The interventions in Experiment 3 are designed to amplify these experiential differences between channels. In particular, the visible presence of a human may increase algorithm aversion by creating a stronger contrast and making the alternative (Chatbot channel) appear more algorithmic. Similarly, introducing physical engagement, such as holding a button rather than clicking prompts in Channel A, may increase the perception of interacting with a human by simulating agency and control over the interaction. Thus, we hypothesize:

H3.1 (Live): Conditional on expected waiting times, participants in the *Context + Live* treatment choose the Chatbot channel less frequently than in the *Context* treatment.

H3.2 (Hold): Conditional on expected waiting times, participants in the *Context + Hold* treatment choose the Chatbot channel less frequently than in the *Context* treatment.

5.2. Results

5.2.1. Descriptive Statistics

Figure 7 shows average chatbot uptake by decision, with the *Context* treatment added as a comparison baseline. The figure suggests that chatbot uptake goes down in the *Context + Live* treatment relative to the *Context* treatment, with decreases between two and eight percentage points, depending on the decision. Further, chatbot uptake also goes down in the *Context + Hold* treatment, with decreases between two and seventeen percentage points. Thus, both manipulations aimed at making the perceptual differences between channels more salient appear to increase algorithm aversion, providing some initial support for H3.1 and H3.2.

5.2.2. Hypothesis Tests

To test H3 we regressed channel choice on the treatment dummies. Table 6 presents the estimates. As before, in Table 6 we report one-sided, Bonferroni-Holm adjusted $p-$values

Figure 7 Channel B (Chatbot Channel) Uptake in Experiment 3

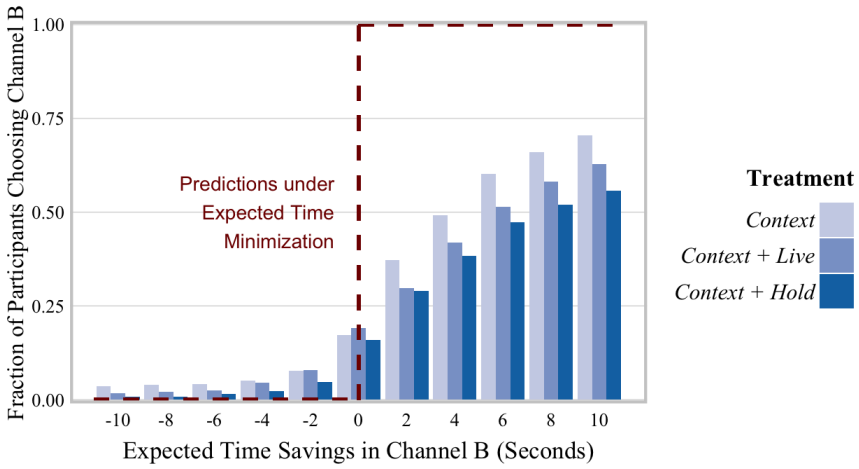


Table 6 Channel Preferences in Experiment 3

Dependent Variable:	(1) Channel B (Chatbot Channel)
Context Treatment	Omitted category
Context + *Live* Treatment	-1.064**
	(0.536)
Context + *Hold* Treatment	-1.737***
	(0.540)
Channel Performance Controls	Yes
Demographic Controls	Yes
Sample	Full Sample
Observations	9240
Subjects	280

Notes: Random effects logit regression coefficients are reported. Dependent variable is the channel choice (Channel B $= 1$). Standard errors are clustered at subject level. Decision set number and decision number within the decision set are controlled for. The following demographic variables are controlled for: age, gender, number of quiz errors and the Eckel-Grossman risk aversion measure (administered after the main task). H3 tests are one-sided, with $p-$values adjusted for multiple hypothesis testing using Bonferroni-Holm procedure. The remaining tests are two-sided. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.1$.

(Table EC.7 presents the details of the multiple hypothesis adjustment procedure). Table 6 shows that both the effects of introducing perceptual differences between channels (through manipulating how the user interacts with the server), and introducing a live human into the interaction, reduces chatbot uptake. In particular, adding a live human to the service process decreased chatbot uptake by 7.02 percentage points, on average ($p = 0.013$). Similarly, introducing server-specific interaction modes (holding down a button for live agent vs. prompt clicking for chatbot) decreased chatbot uptake by 9.92 percentage points, on average ($p < 0.01$). Thus, the regression results confirm that the observed increase in algorithm aversion is indeed statistically significant.[8]

Result 3 (Algorithm Aversion): *H3.1 and H3.2 are supported. Channel B uptake is reduced under the Context + Live and the Context + Hold manipulations relative to the Context treatment.*

5.3. Discussion

In Experiment 3 we explored the trade-off between experimental control and external validity when studying chatbot adoption in service settings. Adding realistic elements such as live interactions and physical engagement creates a more authentic experience but can potentially reduce experimental control, making it more difficult to measure exact times spent or rule out alternative explanations. Despite these more realistic implementations, the overall magnitude of algorithm aversion in Experiment 3 was comparable to that in

[8] In post-hoc comparisons we find that the difference between *Context* + *Live* and *Context* + *Hold* treatment dummies is not statistically insignificant ($p = 0.225$), suggesting that the two manipulations led to similar levels of algorithmic aversion.

Experiment 1, suggesting that our key results are robust to alternative implementations. At the same time, we saw that perceptual differences contribute modest (though statistically significant) effects to the overall aversion. This has both practical and methodological implications. From the practical standpoint, it suggests that the effects of the remedies in Experiment 2 (priority queues, transparency and reducing perceived uncertainty through nudges) likely present a lower bound on potential cost savings. Indeed, the stronger aversion levels observed in Experiment 3 provide more room for improvement, which may lead to even greater savings from improvements in service design. Lastly, from a methodological perspective, the results presented in this section suggest that future research on algorithmic technology adoption should exercise caution when using purely contextual, framing-based manipulations, as these can significantly underestimate algorithm aversion.

6. 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 and their operational implications. In this paper we studied chatbot adoption by first soliciting and analyzing testimonies from chatbot users, by using these user stories to formulate a key trade-off in channel choice, by examining how users navigate this trade-off in incentivized experiments, and by evaluating the cost savings achievable through simple process-design interventions.

Summary of Results The standard approach in the service operations literature is to model queue-joining behavior as a simple expected-time minimization problem (Naor 1969, Hassin and Haviv 2003). Our experimental results suggest that this approach may oversimplify behavior when one service channel has a gatekeeper structure. Specifically, we found that most decision-makers avoided the gatekeeper channel even when this channel offered shorter expected waiting times. We termed this behavior *gatekeeper aversion* and characterized is as a combination of risk and transfer aversion. Separately, we identified another hurdle to chatbot uptake – *algorithm aversion* – and showed that it can amplify gatekeeper aversion. In particular, while algorithm aversion was not detectable in isolation, i.e., in our pre-tests where the service processes and performance were identical across channels, associating the gatekeeper channel with the chatbot further decreased participants' willingness to use it. Lastly, we examined potential remedies for chatbot under-utilization and showed that priority queues, operational transparency and an average waiting time nudge increase adoption. Using counterfactual analysis of channel joining behavior in an $M/D/1$ system, we showed that, when combined, these remedies can yield staffing cost savings of up to 19.7%.

Contributions We make several contributions to theory, practice, and experimental methodology. On the theory front, we extend the conversation on human-AI interfaces, which has traditionally focused on AI adoption among workers (Dietvorst et al. 2015, Yeomans et al. 2019, Jussupow et al. 2020), to customer

decisions in service channel selection. By connecting this research stream to the service operations literature on queue-joining behavior (Allon and Kremer 2018) and gatekeeper service systems (Shumsky and Pinker 2003, Freeman et al. 2017), we show that the classic finding that errors loom larger when made by algorithms does not seamlessly generalize to our customer service context. Instead, algorithm aversion plays a more subtle role: it amplifies existing reluctance to engage with a channel that has a gatekeeper structure. The implication of this result is that the reluctance to using chatbot technology is likely to drop substantially once this technology reaches performance levels similar to humans. Second, we make a more practical contribution by showing that low-cost, easily implementable service design interventions – such as introducing priority queues, providing average wait information, and truthfully revealing chatbot capabilities – can significantly increase chatbot adoption and generate substantial cost savings for the service provider. Third, we develop a novel experimental approach for eliciting algorithm aversion in service systems and show that experimental designs relying solely on contextual framing may underestimate algorithm aversion, compared to designs that vary the human vs. algorithmic nature of the interaction in a more realistic manner.

Limitations and Extension To keep the experiments focused, we did not model certain aspects of channel choice, such as the language and style of service interactions, the seamlessness of transitions between gatekeepers and experts, the uncertainty in interaction times, or the residual probability of expert failure. Examining these features may add realism to our setup and may improve the generalizability of our findings. Avenues for future research also include the role of algorithmic preferences when interacting with firm-specific vs. general-purpose chatbots (e.g., ChatGPT), the use of chatbots to perform purely diagnostic work vs. more task-oriented functions, and the role of privacy concerns and the use of customer data in these service interactions. Finally, our experiments in §5, which tested basic variations in interaction mode, suggest that richer server-customer interaction environments such as voice, video, or virtual reality deserve further study as they could produce markedly different levels of algorithm aversion.

Outlook The customer service context provides an ideal setting for our study because it allows precise control and communication of waiting times, involves outcomes that are binary (a request is either resolved or not), and is largely conducted online rather than in physical retail stores. Therefore, our online experiments mirror service interactions with high fidelity to the real-world setting. However, given that similar technology adoption challenges exist beyond customer service, we believe that some of our results may apply more broadly. Examples include self-checkout stations in grocery stores, automated check-in kiosks at airports, or digital ordering systems in restaurants. Future research could examine whether key behaviors identified in chatbot interactions – such as gatekeeper aversion and algorithm aversion – are also observed in these alternative automated settings, and how service design interventions like operational transparency and pooled/dedicated queues might mitigate technology underutilization more broadly. As self-service technologies continue to evolve and mature, controlled experiments offer a powerful tool that can add to our understanding of customer behavior and service design.

References

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

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.

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

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

Buell RW, Norton MI (2011) The labor illusion: How operational transparency increases perceived value. *Management Science* 57(9):1564–1579.

Carmon Z, Kahneman D (1996) The experienced utility of queuing: experience profiles and retrospective evaluations of simulated queues. *Durham, NC: Fuqua School, Duke University* .

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

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

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.

Dada M, Hathaway B, Kagan E (2025) Customer service operations: A gatekeeper framework. *Production and Operations Management, forthcoming* .

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).

Harrison GW, Cox JC (2008) *Risk aversion in experiments* (Emerald Group Publishing).

Hassin R, Haviv M (2003) *To queue or not to queue: Equilibrium behavior in queueing systems*, volume 59 (Springer Science & Business Media).

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

Holm S (1979) A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics* 65–70.

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

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

Jussupow E, Benbasat I, Heinzl A (2020) Why are we averse towards algorithms? a comprehensive literature review on algorithm aversion .

Kahneman D, Tversky A (1979) Prospect theory: An analysis of decision under risk. *Econometrica* 47(2):363–391.

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.

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

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

List JA, Shaikh AM, Xu Y (2019) Multiple hypothesis testing in experimental economics. *Experimental Economics* 22:773–793.

Logg JM, Minson JA, Moore DA (2019) Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151:90–103.

Longoni C, Bonezzi A, Morewedge CK (2019) Resistance to medical artificial intelligence. *Journal of Consumer Research* 46(4):629–650.

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.

Maynard N, Crabtree G (2020) Artificial intelligence 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.

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.

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.

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

Tversky A, Kahneman D (1992) Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty* 5:297–323.

Yeomans M, Shah A, Mullainathan S, Kleinberg J (2019) Making sense of recommendations. *Journal of Behavioral Decision Making* 32(4):403–414.

Electronic Companion

EC.1. Supporting Analysis for Survey (Wave 1 and Wave 2)

Below we provide questions, data and the results summary for the retrospective survey referred to in §2.1 of the manuscript. We conducted the survey in two waves - an exploratory survey in June 2022 ($N = 198$) and a confirmatory one in July 2023 ($N = 202$).

EC.1.1. Wave 1: Survey Setup and Questions

The data were collected in June 2022. A total of 202 respondents (55% female, average age: 33) were recruited on the Prolific platform. 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. 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-Q12). 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).

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%).

[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 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?*

[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: Participants see two figures with an interaction resembling a specific problem-solving inquiry and a more general, open-ended inquiry]*, 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 expected (2.60%).

Q8. *How long did the interaction with the agent [chatbot] last?*
Live Agent: Less than 1 minute (0.00%) 1-2 minutes (8.24%) At least 3 minutes (91.76%).
Chatbot: Less than 1 minute (2.60%) 1-2 minutes (27.27%) At least 3 minutes (70.13%).

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

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

Chatbot: No, my details were not required (35.06%) No, the agent was able to retrieve most of the details from the system (23.38%) Yes, I had to share those details (41.56%).

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

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

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

EC.1.2. Wave 2: Survey Setup and Questions

The second wave was also conducted on the Prolific platform and also consisted of two between-subject "treatments": a live agent and a chatbot treatment ($N = 106$ 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. This survey included the questions from Survey 1, with two additions. 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. Below we reproduce the questions asked in the survey, along with summary statistics of responses for multiple choice 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]?*

[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. *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%).

Q6. *How long did you have to wait for the agent [chatbot] relative to your initial expectation?*[Question dropped in Survey 2]

Q7. *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%).

Q8. *How long did the interaction last relative to your initial expectation?* [Question dropped in Survey 2]

Q9. *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]

Q10. *Approximately how many questions did the agent [chatbot] ask you during the interaction?* [Question dropped in Survey 2]

Q11. *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%).

Q12. *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.1.3. Additional Questions

Q13. *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%)

Q14. *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.1.4. Key Results and Discussion

EC.1.4.1. 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)*

EC.1.4.2. Survey 2 In the second survey 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 EC.1 summarizes the key findings from the survey.

First, 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%, $p = 0.034$). Second, 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 ($p \ll 0.01$). The majority of the chatbot users with unresolved requests were either transferred to a live agent (23.38%) or had to call a live agent (29.87%), thus spending additional time in the system until their request was resolved. Together, these comparisons suggest a speed-performance trade-off in channel choice: chatbots have a lower success rate but are faster to access, while live agents may require a longer wait but are usually the final step of the encounter.

Table EC.1 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.

EC.2. Technical Materials for Experiments

In this section, we present supporting technical materials for §3-§5. Tables EC.2 and EC.3 present the parameters used in the short and long time duration conditions in the *Context* and *No Context* treatments in Experiment 1 and in all treatments in Experiments 2 and 3. Table EC.4 presents the parameters used in the short *No Context, Deterministic* condition in Experiment 1. Tables EC.5 - EC.7 present multiple hypothesis adjustment results for the regression coefficients reported in Tables 4 - 6.

Table EC.2 Experimental Parameters (*No Context* and *Context* Treatments, Short Time Durations)

Decision Set 1: Varying Gatekeeper Success Rate (p^B)

	Channel A		Channel B				Expected time minimizing choice
	t^A_{serve}	$t^A_{serce_1}$	p^B	$t^B_{serve_1}$	t^B_{line}	$t^B_{serve_2}$	
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 ($t^B_{serve_1}$)

	Channel A		Channel B				Expected time minimizing choice
	t^A_{serve}	t^A_{serve}	p^B	$t^B_{serve_1}$	t^B_{line}	$t^B_{serve_2}$	
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^B_{line})

	Channel A		Channel B				Expected time minimizing choice
	t^A_{serve}	t^A_{serve}	p^B	$t^B_{serve_1}$	t^B_{line}	$t^B_{serve_2}$	
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 are in seconds.

Table EC.3 Experimental Parameters (*No Context* and *Context* Treatments, Long Time Durations)

Decision Set 1: Varying Gatekeeper Success Rate (p^B)

	Channel A		Channel B				Expected time minimizing choice
	t^A_{line}	t^A_{serve}	p^B	$t^B_{serve_1}$	t^B_{line}	$t^B_{serve_2}$	
Decision 1	40	40	0.25	40	40	40	Channel A
Decision 2	40	40	0.3	40	40	40	Channel A
Decision 3	40	40	0.35	40	40	40	Channel A
Decision 4	40	40	0.4	40	40	40	Channel A
Decision 5	40	40	0.45	40	40	40	Channel A
Decision 6	40	40	0.5	40	40	40	Indifferent
Decision 7	40	40	0.55	40	40	40	Channel B
Decision 8	40	40	0.6	40	40	40	Channel B
Decision 9	40	40	0.65	40	40	40	Channel B
Decision 10	40	40	0.7	40	40	40	Channel B
Decision 11	40	40	0.75	40	40	40	Channel B

Decision Set 2: Varying Gatekeeper Service Time ($t^B_{serve_1}$)

	Channel A		Channel B				Expected time minimizing choice
	t^A_{serve}	t^A_{serve}	p^B	$t^B_{serve_1}$	t^B_{line}	$t^B_{serve_2}$	
Decision 1	40	40	0.5	60	40	40	Channel A
Decision 2	40	40	0.5	56	40	40	Channel A
Decision 3	40	40	0.5	52	40	40	Channel A
Decision 4	40	40	0.5	48	40	40	Channel A
Decision 5	40	40	0.5	44	40	40	Channel A
Decision 6	40	40	0.5	40	40	40	Indifferent
Decision 7	40	40	0.5	36	40	40	Channel B
Decision 8	40	40	0.5	32	40	40	Channel B
Decision 9	40	40	0.5	28	40	40	Channel B
Decision 10	40	40	0.5	24	40	40	Channel B
Decision 11	40	40	0.5	20	40	40	Channel B

Decision Set 3: Varying Line Duration after Gatekeeper Failure (t^B_{line})

	Channel A		Channel B				Expected time minimizing choice
	t^A_{serve}	t^A_{serve}	p^B	$t^B_{serve_1}$	t^B_{line}	$t^B_{serve_2}$	
Decision 1	40	40	0.5	40	80	40	Channel A
Decision 2	40	40	0.5	40	72	40	Channel A
Decision 3	40	40	0.5	40	64	40	Channel A
Decision 4	40	40	0.5	40	56	40	Channel A
Decision 5	40	40	0.5	40	48	40	Channel A
Decision 6	40	40	0.5	40	40	40	Indifferent
Decision 7	40	40	0.5	40	32	40	Channel B
Decision 8	40	40	0.5	40	24	40	Channel B
Decision 9	40	40	0.5	40	16	40	Channel B
Decision 10	40	40	0.5	40	8	40	Channel B
Decision 11	40	40	0.5	40	0	40	Channel B

Notes: The sequence of decision sets was chosen at random for each participant. All time parameters are in seconds.

Table EC.4 Experimental Parameters (*No Context, Deterministic* Treatment, Short Time Durations)

Decision Set 1: Varying Second Service Stage Time ($t^B_{serve_2}$)

	Channel A		Channel B				Expected time minimizing choice
	t^A_{serve}	t^A_{serve}	p^B	$t^B_{serve_1}$	t^B_{line}	$t^B_{serve_2}$	
Decision 1	20	20	0	10	10	30	Channel A
Decision 2	20	20	0	10	10	28	Channel A
Decision 3	20	20	0	10	10	26	Channel A
Decision 4	20	20	0	10	10	24	Channel A
Decision 5	20	20	0	10	10	22	Channel A
Decision 6	20	20	0	10	10	20	Indifferent
Decision 7	20	20	0	10	10	18	Channel B
Decision 8	20	20	0	10	10	16	Channel B
Decision 9	20	20	0	10	10	14	Channel B
Decision 10	20	20	0	10	10	12	Channel B
Decision 11	20	20	0	10	10	10	Channel B

Decision Set 2: Varying First Service Stage Time ($t^B_{serve_1}$)

	Channel A		Channel B				Expected time minimizing choice
	t^A_{serve}	t^A_{serve}	p^B	$t^B_{serve_1}$	t^B_{line}	$t^B_{serve_2}$	
Decision 1	20	20	0	30	10	10	Channel A
Decision 2	20	20	0	28	10	10	Channel A
Decision 3	20	20	0	26	10	10	Channel A
Decision 4	20	20	0	24	10	10	Channel A
Decision 5	20	20	0	22	10	10	Channel A
Decision 6	20	20	0	20	10	10	Indifferent
Decision 7	20	20	0	18	10	10	Channel B
Decision 8	20	20	0	16	10	10	Channel B
Decision 9	20	20	0	14	10	10	Channel B
Decision 10	20	20	0	12	10	10	Channel B
Decision 11	20	20	0	10	10	10	Channel B

Decision Set 3: Varying Line Duration (t^B_{line})

	Channel A		Channel B				Expected time minimizing choice
	t^A_{serve}	t^A_{serve}	p^B	$t^B_{serve_1}$	t^B_{line}	$t^B_{serve_2}$	
Decision 1	20	20	0	10	30	10	Channel A
Decision 2	20	20	0	10	28	10	Channel A
Decision 3	20	20	0	10	26	10	Channel A
Decision 4	20	20	0	10	24	10	Channel A
Decision 5	20	20	0	10	22	10	Channel A
Decision 6	20	20	0	10	20	10	Indifferent
Decision 7	20	20	0	10	18	10	Channel B
Decision 8	20	20	0	10	16	10	Channel B
Decision 9	20	20	0	10	14	10	Channel B
Decision 10	20	20	0	10	12	10	Channel B
Decision 11	20	20	0	10	10	10	Channel B

Notes: The sequence of decision sets was chosen at random for each participant. All time parameters are in seconds.

Table EC.5 Experiment 1: Multiple Testing Adjustments Results

				Unadjusted		Holm-Bonferroni adjusted	
Hypothesis	Test Type	Test Statistic	p-value	Cutoff	Reject H_0?	Cutoff	Reject H_0?
Full Sample (Col. 1 of Table 3)							
H1.1 (Gatekeeper aversion)	t-test	-7.756	< 0.001	0.050	Yes	0.017	Yes
H1.2 (Risk aversion)	logistic	5.310	< 0.001	0.050	Yes	0.025	Yes
H1.3 (Algorithm aversion)	logistic	-1.340	0.090	0.050	No	0.050	No
Short Time Durations (Col. 2 of Table 3)							
H1.1 (Gatekeeper aversion)	t-test	-5.100	< 0.001	0.050	Yes	0.017	Yes
H1.2 (Risk aversion)	logistic	3.450	< 0.001	0.050	Yes	0.025	Yes
H1.3 (Algorithm aversion)	logistic	-0.100	0.460	0.050	No	0.050	No
Long Time Durations (Col. 3 of Table 3)							
H1.1 (Gatekeeper aversion)	t-test	-5.840	< 0.001	0.050	Yes	0.017	Yes
H1.2 (Risk aversion)	logistic	3.870	< 0.001	0.050	Yes	0.025	Yes
H1.3 (Algorithm aversion)	logistic	-1.690	0.046	0.050	Yes	0.050	Yes

All tests are one-sided. H1.1 uses t-test (negative effect). H1.2 and H1.3 use logistic regression coefficients (positive and negative effects respectively). Holm-Bonferroni sequential cutoffs are calculated as follows: Order the p-values from smallest to largest: $p_{(1)} \leq p_{(2)} \leq p_{(3)}$. For family-wise error rate $\alpha = 0.05$ and $m = 3$ hypotheses, compare each $p_{(k)}$ to its adjusted cutoff $\alpha_{(k)} = \alpha/(m-k+1)$. Thus, $p_{(1)}$ is compared to $\alpha/(m-1+1) = 0.05/3 = 0.017$, $p_{(2)}$ to $\alpha/(m-2+1) = 0.05/2 = 0.025$, and $p_{(3)}$ to $\alpha/(m-3+1) = 0.05/1 = 0.050$. Reject $H_{(k)}$ if $p_{(k)} \leq \alpha_{(k)}$ and all hypotheses $H_{(j)}$ with $j < k$ were rejected.

Table EC.6 Experiment 2: Multiple Testing Adjustments Results

				Unadjusted		Holm-Bonferroni adjusted	
Hypothesis	Test Type	Test Statistic	p-value	Cutoff	Reject H_0?	Cutoff	Reject H_0?
Full Sample (Col. 1)							
H2.1 (Transparency)	logistic	-1.090	0.139	0.050	No	0.050	No
H2.2 (Nudge)	logistic	2.430	0.008	0.050	Yes	0.025	Yes
Short Time Durations (Col. 2)							
H2.1 (Transparency)	logistic	-2.070	0.019	0.050	Yes	0.025	Yes
H2.2 (Nudge)	logistic	-0.440	0.669	0.050	No	0.050	No
Long Time Durations (Col. 3)							
H2.1 (Transparency)	logistic	0.430	0.664	0.050	No	0.050	No
H2.2 (Nudge)	logistic	3.920	< 0.001	0.050	Yes	0.025	Yes

All tests are one-sided. H2.1 and H2.2 use logistic regression coefficients. Holm-Bonferroni sequential cutoffs are calculated as follows: Order the p-values from smallest to largest: $p_{(1)} \leq p_{(2)}$. For family-wise error rate $\alpha = 0.05$ and $m = 2$ hypotheses, compare each $p_{(k)}$ to its adjusted cutoff $\alpha_{(k)} = \alpha/(m-k+1)$. Thus, $p_{(1)}$ is compared to $\alpha/(m-1+1) = 0.05/2 = 0.025$ and $p_{(2)}$ to $\alpha/(m-2+1) = 0.05/1 = 0.050$. Reject $H_{(k)}$ if $p_{(k)} \leq \alpha_{(k)}$ and all hypotheses $H_{(j)}$ with $j < k$ were rejected.

Table EC.7 Experiment 3: Multiple Testing Adjustments Results

				Unadjusted		Holm-Bonferroni adjusted	
Hypothesis	Test Type	Test Statistic	$p-$value	Cutoff	Reject H_0?	Cutoff	Reject H_0?
H3.1 (Live)	logistic	-1.987	0.024	0.050	Yes	0.050	Yes
H3.2 (Hold)	logistic	-3.238	0.001	0.050	Yes	0.025	Yes

All tests are one-sided and are based on logistic regression coefficients reported in Table 6. Holm-Bonferroni sequential cutoffs are calculated as follows: Order the p-values from smallest to largest: $p_{(1)} \leq p_{(2)}$. For family-wise error rate $\alpha = 0.05$ and $m = 2$ hypotheses, compare each $p_{(k)}$ to its adjusted cutoff $\alpha_{(k)} = \alpha/(m-k+1)$. Thus, $p_{(1)}$ is compared to $\alpha/(m-1+1) = 0.05/2 = 0.025$ and $p_{(2)}$ to $\alpha/(m-2+1) = 0.05/1 = 0.050$. Reject $H_{(k)}$ if $p_{(k)} \leq \alpha_{(k)}$ and all hypotheses $H_{(j)}$ with $j < k$ were rejected.

EC.3. Additional Experimental Details

EC.3.1. Experiment 1

The instructions for the *Context* treatment are reproduced on Researchbox (`https://researchbox.org/3917&PEER_REVIEW_passcode=BQXGDU`). Instructions for the remaining treatments are analogous, with the difference being that the live agent channel is referred to as "Channel A" and the chatbot channel as "Channel B". See Fig. EC.1 for an illustration of the visuals used in the experiment.

EC.3.2. Experiment 2

Screenshots of decision screens with and without transparency are shown in Figure EC.2. Screenshots of decision screens with and without the nudge are shown in Figure EC.3.

EC.3.3. Experiment 3

Context + Live Treatment: To perform service 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 in Channel B prior to making any choices to test the interface. The script is below. As in the remaining treatments, participants also test Channel A prior to making any choices.

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^A_{serve} seconds.].

The live chat interface is shown in Figure EC.4(a). The chat window popped up as soon as participants entered the first service stage in Channel A, or the second service stage in Channel B (if the first stage had failed). The experimenters were instructed to initiate service as soon as they received the correct letter from the participant. The interface for Channel B was kept identical to that of the *Context* treatment.

Context + Hold Treatment: The screenshots of the key experimental manipulation in the *Context + Hold* treatment are reproduced in Figure EC.5. Moreover, as part of Experiment 3, we performed an attitudinal test to validate the *Context + Hold* manipulation. In particular, we presented respondents with the click-based prompts and holding and then asked them to rate each type of manipulation as more human-like or more chatbot-like. There were two versions of the pre-test. 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. In the second version of the survey (N=96), participants completed the same tasks but with added context. In both versions, the first experience required holding down a button for the progress bar to fill (Figure EC.5(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 EC.5(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. The key variable of interest is the percentage of respondents rating each type of interaction as being more human-like or more bot-like.

Table EC.8 reports the results. Consider first the non-contextualized setting. 31.13% of respondents rated holding down the button as being more human-like vs. 27.36% in the click-based prompts manipulation, with the difference not being statistically significant (Proportion test, $p = 0.545$). Further, 42.45% of respondents rated holding as being more bot-like vs. 60.38% for click-based prompts (proportion test, $p = 0.009$). Next consider the contextualized version of the pre-test. Here, 50.00% of respondents rated holding as being more human-like vs. 15.62% for Channel B ($p \ll 0.001$). Conversely, 25.00% perceived clicking to be more bot-like, vs. 79.17% for Channel B ($p \ll 0.001$). The remaining respondents chose the *"Not sure"* category. These comparisons validate our interaction mode manipulation in the *Context + Hold* Treatment.

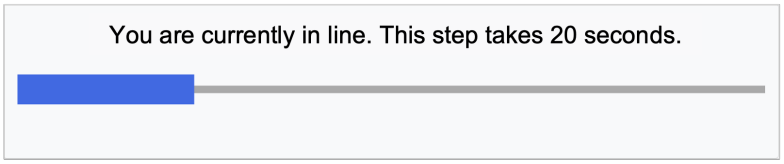
Table EC.8 Pre-Tests for Context + Hold

	"Felt more like a live agent interaction"			*"Felt more like a chatbot interaction"*			*"Not sure"*		
	Button Holding	Click-based Prompts	$p-$value	Button Holding	Click-based Prompts	$p-$value	Button Holding	Click-based Prompts	$p-$value
Non-contextualized	31.13 %	27.36 %	0.545	42.45 %	60.38 %	0.009	26.42%	12.26%	0.009
Contextualized	50.00 %	15.62 %	$\ll 0.001$	25.00 %	79.17 %	$\ll 0.001$	25.00 %	5.21 %	$\ll 0.001$

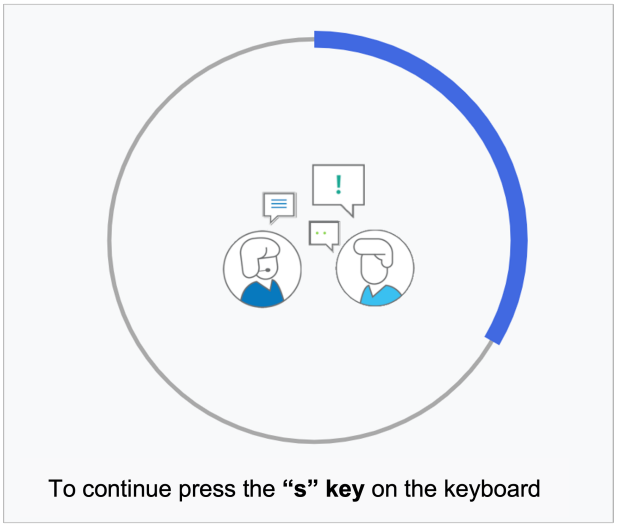
Notes: Table reports percentage of respondents in each category, as well as $p-$values from non-parametric tests of equality of proportions.

Figure EC.1 Waiting Experiences (Screenshots)

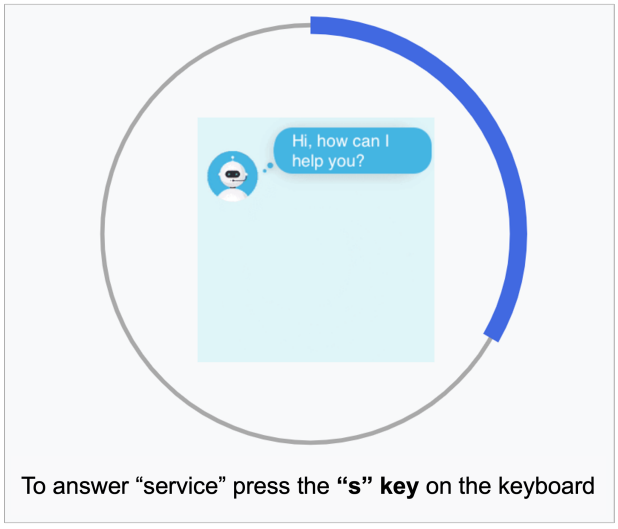
(a) All Treatments, In-line Screen



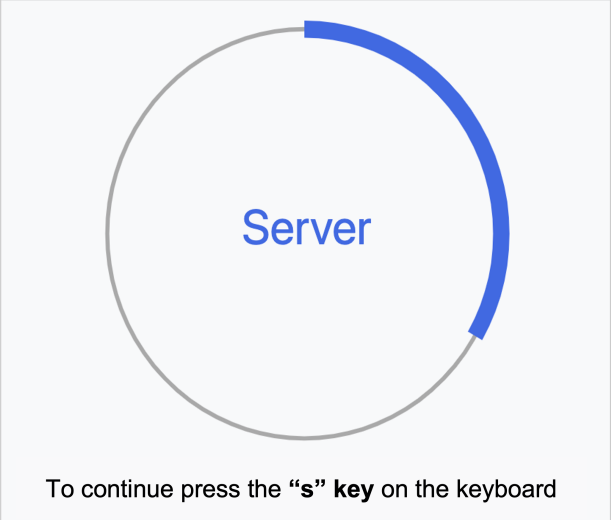
(b) *Context* Treatment, In-service Screen (Channel A)



(c) *Context* Treatment, In-service Screen (Channel B)



(d) *No Context* Treatment, In-service Screen (Channel A)



(e) *No Context* Treatment, In-service Screen (Channel B)

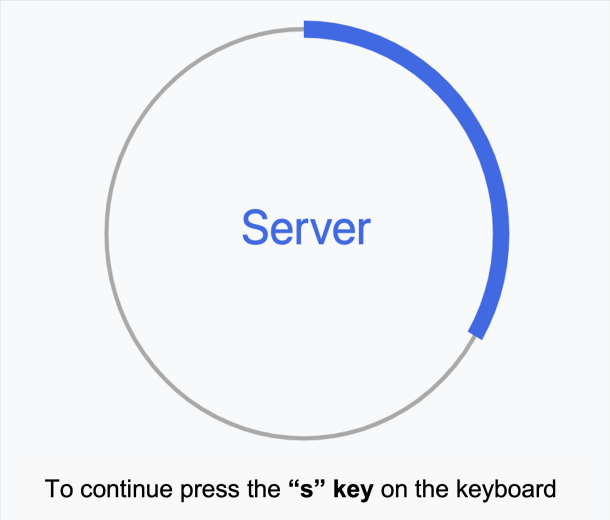
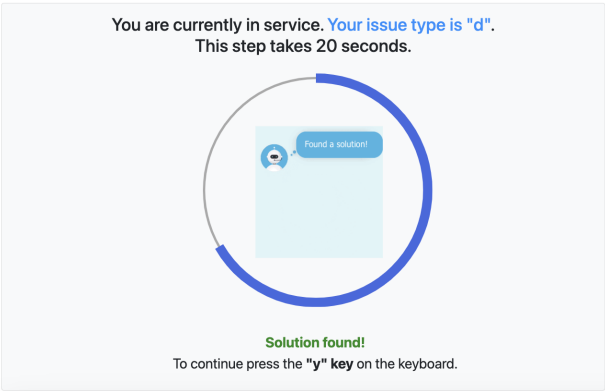
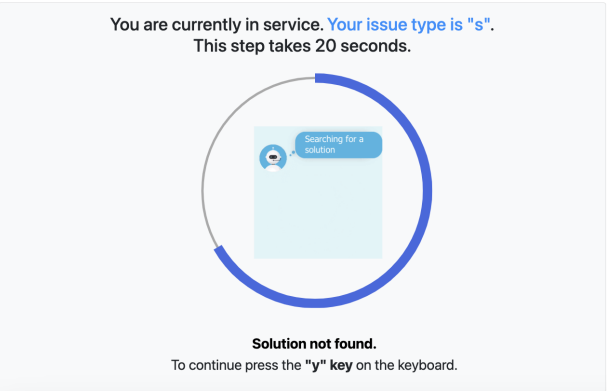


Figure EC.2 Waiting Experiences in *Context* Treatment vs. *Context + No transparency* Treatment

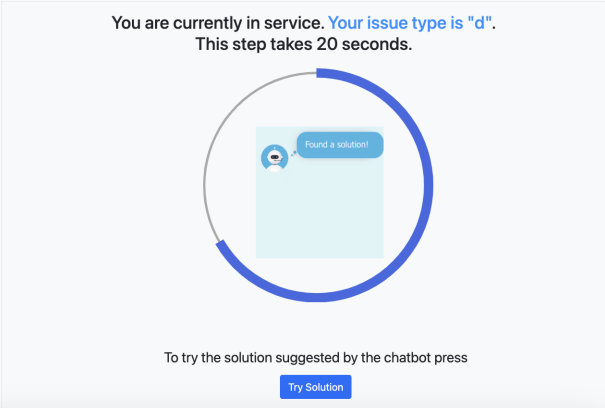
(a) *Context* Treat., Chatbot Succeeds



(b) *Context* Treat., Chatbot Fails



(c) *Context + No Transparency* Treat., Chatbot Succeeds



(d) *Context + No Transparency* Treat., Chatbot fails

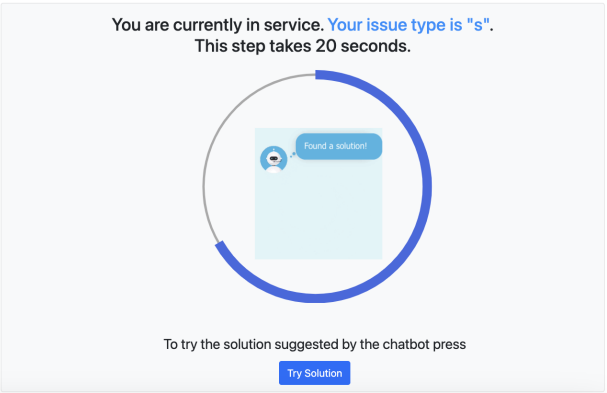


Figure EC.3 Choice Screens

(a) Example Choice Screen Without the Nudge

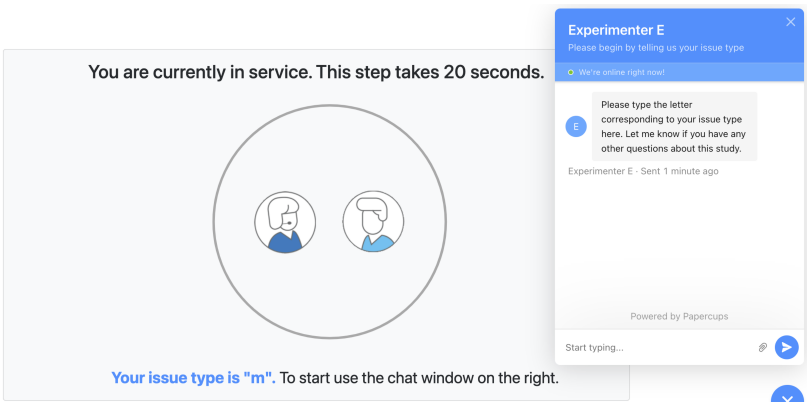
Scenario 1	20 + 20 = 40 sec. w. prob. 100%	○ ○	20 sec. w. prob. **25%**, 20 + 20 + 20 = 60 sec. w. prob. **75%**.
Scenario 2	20 + 20 = 40 sec. w. prob. 100%	○ ○	20 sec. w. prob. **30%**, 20 + 20 + 20 = 60 sec. w. prob. **70%**.
Scenario 3	20 + 20 = 40 sec. w. prob. 100%	○ ○	20 sec. w. prob. **35%**, 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) Example Choice Screen with the Nudge

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)**

Figure EC.4 *Context + Live* Treatment: Screenshots

(a) Interaction with a Live Agent



(b) Interaction with a Chatbot

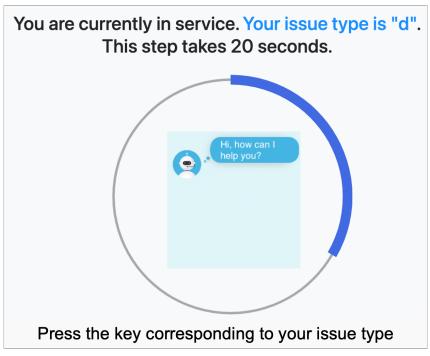
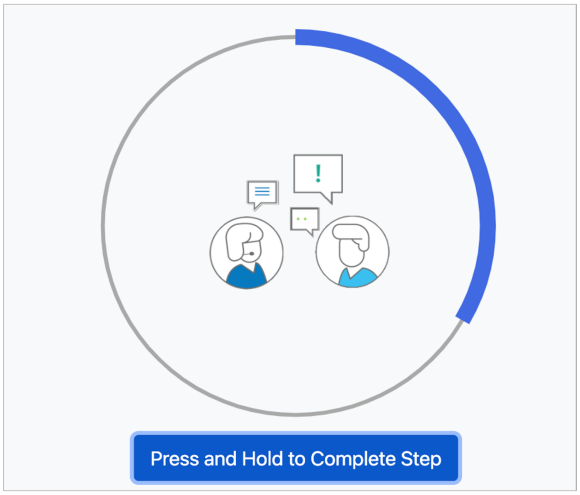
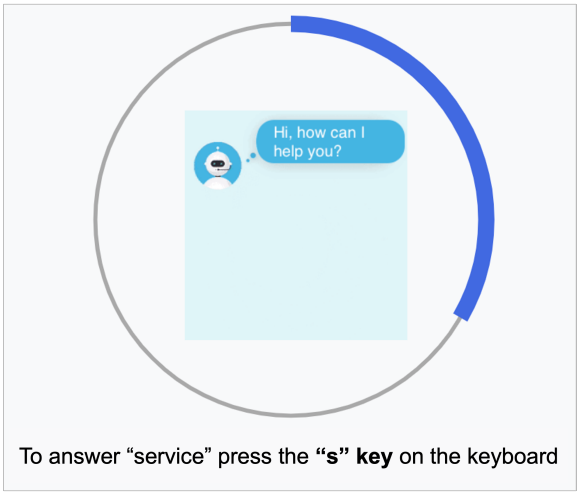


Figure EC.5 *Context + Hold* Treatment: Screenshots

(a) *Context + Hold* Treat., Interaction w. Live Agent



(b) *Context + Hold* Treat., Interaction w. Chatbot



EC.4. Channel choice utility model

In this section we estimate four specifications of the structural model and provide the estimates in Table EC.9. We first estimate a convex waiting cost model of exponential utility where c_{bot} is estimated separately for treatments with and without the nudge. We then estimate three linear specifications, two of which are restrictions of our model in §4.4. In Linear 1 we set β to 1, allowing differences in waiting costs across service stages to rationalize chatbot aversion, again picking up the effect of the nudge through separate estimates of c_{bot}. In Linear 2 we set all waiting costs equal, allowing β to fully rationalize chatbot aversion, with seperate estimates of β for the nudge and non-nudge treatments. Finally, Linear 3 is our selected specification (based on AIC), which allows for costs to vary by service stage and β to vary based on the nudge.

Table EC.9 Structural Estimates

Parameter	Exponential	Linear 1	Linear 2	Linear 3
c_{line}	0.0194***	0.1634***	0.1407***	0.1372***
	(0.0022)	(0.0078)	(0.0060)	(0.0062)
c_{agent}	0.0195***	0.1412***	0.1407***	0.1705***
	(0.0035)	(0.0084)	(0.0060)	(0.0087)
c_{bot}	0.0206***	0.2026***	0.1407***	0.1709***
	(0.0007)	(0.0090)	(0.0060)	(0.0073)
c^{bot} nudge	0.0192***	0.1911***		
	(0.0005)	(0.0077)		
c_{nt}	0.0993*	0.6002*	0.5209*	0.4218*
	(0.0514)	(0.2947)	(0.3599)	(0.2800)
β			1.3601***	1.2338***
			(0.0349)	(0.0286)
β nudge			1.3601***	1.1517***
			(0.0265)	(0.0301)
Log likelihood	-10502.82	-9078.70	-9072.70	-9014.47

*$p < .10$, ***$p < .001$.