Beyond Algorithm Aversion: The Role of Risk and Gatekeeper Aversion in AI Chatbot Adoption

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

Problem definition: Despite recent advances in large language models, chatbot technology continues to face adoption hurdles. We examine the drivers of choices between the chatbot channel and the live agent channel in customer service. **Methodology/results:** We report the results of four studies (a retrospective survey and three incentivized online experiments). We show that people respond positively to improvements in chatbot performance; however, the chatbot channel is utilized less frequently than expected time minimization would predict. This underutilization is caused by three separate mechanisms: *gatekeeper aversion* (aversion to any service format that may involve a transfer to a different server), *risk aversion* (aversion to uncertainty in the time spent in the service system) and *algorithm aversion* (aversion to an algorithmic service provider). Among the three, algorithm aversion is the weakest mechanism, and is only statistically significant for those aged 40 and above. In contrast, gatekeeper aversion and risk aversion are robustly present across age groups (and other demographics), and are comparable in size. **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 facilitating more seamless chatbot-to-agent transfers, 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

Chatbot technology has many uses, yet its commercial success is most visible in online customer support. Recent technological advances have significantly increased chatbot capabilities, improved their speed, enabled them to handle more complex, often unstructured customer queries, and reduced training and maintenance costs (Johannsen et al. 2018). These improvements have 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 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 the 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, Hasija et al. 2005, Freeman et al. 2017, Hathaway et al. 2022), where the chatbot plays the role of a gatekeeper that handles only a subset of the incoming requests, with the remaining requests being diverted to a live, human agent. This is because certain requests may be difficult to communicate or categorize, or because the chatbot may not be authorized to handle certain requests; for example, ones that involve large financial transactions. Thus, the chatbot serves as the entrance point to, but not necessarily the final step of, the service encounter, similar to a nurse in a hospital or a front desk receptionist in a hotel. Different from 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.[1] In this study we examine the determinants of this channel choice.

Our investigation consists of a user survey and three incentivized experiments. In our survey (§2) respondents retrospectively describe a recent customer service episode, either with a chatbot or with a live agent. Their testimonies suggest a key trade-off in channel choice: chatbots are faster to access but have a lower request resolution rate. In contrast, live agents typically require some wait, but are more reliable in resolving customer requests. This insight helps motivate our theoretical model of channel choice (§3). The core of the paper (§4-6) focuses on testing this model through a series of experiments.

In the first experiment (§4) we present participants with a series of choices between two alternatives. The first alternative ("Channel A") represents the live agent and involves some waiting in line to access service; the server resolves the request with probability 1, after which the participant

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

receives a monetary reward. The second alternative ("Channel B") represents the bot 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. We conduct two treatments: a *Baseline* treatment, in which the choice is described using neutral language, avoiding any contextual cues, and a *Context* treatment, in which the type of the server (human or bot) is explicitly revealed. In the *Context* treatment, the participant interacts either with an (animated) human server (in Channel A), or with a pre-programmed chatbot (in Channel B), depending on their choice. By using a naturalistic framing and presenting participants with contextual cues, the *Context* treatment allows us to separately identify the presence of algorithmic preferences ("algorithm aversion") in this setting, while keeping constant the process-related features and parameters of each channel.

The results of the first experiment show that Channel B uptake is reduced as the first service stage becomes longer in duration, has a higher failure rate, and when the customer needs to wait longer for the second service stage after the first one fails. That is, the better the performance of Channel B (representing the chatbot channel), the higher its uptake. However, in both treatments, Channel B uptake remains far below what expected time minimization would predict. Further, we find no significant treatment differences. In other words, the presence of contextual cues in the *Context* treatment does not affect chatbot uptake.

To perform an alternative test of the presence algorithm aversion, we conduct a second experiment (§5), in which we replicate the *Context* treatment, but employ real humans (research assistants, blind to the experimental hypotheses) who play the role of live agents and interact with participants using a live chat tool. We find that the presence of a live human significantly reduces chatbot (Channel B) uptake relative to the *Baseline* treatment. However, the drop in uptake is observed only for participants aged 40 and above, suggesting that age is an important moderator of algorithm aversion in this setting.

Lastly, in the third experiment (§6) we unpack the process-related (non-algorithmic) mechanisms driving reduced Channel B uptake. In particular, in the first two experiments, Channel B is both more uncertain, and has a different progression of stages (a gatekeeper stage, and possibly a second, expert stage), compared to Channel A, which involves no risk and only has a single service stage. To separately identify the relative importance of risk aversion (aversion to uncertainty in the time spent in the system), and gatekeeper aversion (aversion to a service format that may involve a transfer to a different server), we conduct a *Deterministic* treatment, in which neither of the two channels involve risk. We find that in this treatment Channel B uptake is increased relative to that of the *Baseline* treatment. However, there continues to be a significant gap between Channel B uptake and the expected time minimization benchmark, suggesting that risk and gatekeeper aversion are two separate drivers of channel choice.

Summarizing the results of our experiments, we identify three barriers to chatbot adoption: two process-related adoption hurdles that are not tied to the algorithmic nature of the chatbot (gatekeeper aversion and risk aversion) and an attitudinal bias against chatbot technology (algorithm aversion). In §7 we use the data from all three experiments to evaluate the relative importance of these behaviors and find that gatekeeper and risk aversion together account for approximately 86% of chatbot underutilization, while algorithm aversion accounts for the remainder. These results are confirmed across four additional robustness treatments in which we examine service process variations that test for alternative explanations and boundary conditions. We conclude §7 with a structural estimation of several variants of a customer utility model. The results of this estimation further support the robust presence of each of the three mechanisms in the choice data, and show that other factors play only a minimal role in explaining behavior.

Our contributions are threefold. First, we extend the conversation on human-AI interfaces, which has traditionally focused on AI adoption among workers (Dietvorst et al. 2015, Castelo et al. 2019), to customer decisions in service channel selection. Our findings support the presence of algorithm aversion in this setting but suggest that it is overshadowed by operational factors unrelated to the human or algorithmic nature of the channel (such as chatbot speed and capabilities). Second, we contribute to the methodological discussion of how to best measure algorithmic attitudes (Glikson and Woolley 2020). We show that the common experimental approach that relies on framing may underestimate the presence of algorithm aversion, compared to experiments that vary the human/robotic nature of the interaction in a more realistic manner. Third, we contribute to the service operations literature on queue-joining (Kremer and Debo 2016, Flicker and Hannigan 2022), and queue-switching and reneging behaviors (Akşin et al. 2020, Buell 2021), as well as to the studies examining the internal decision trade-offs between the value and the cost of waiting in lines (Naor 1969, Kremer and Debo 2016, Ülkü et al. 2020, Luo et al. 2022). Our contribution to this literature is that we study different waiting modalities, such as in-line and in-service waits, and present a novel way to systematically evaluate responses to the content of the wait, while carefully controlling for its duration.

2. Retrospective Survey

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

2.1. Methodology

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

2.1.1. Survey 1 The first wave of the survey was conducted on the Prolific platform and consisted of two versions or "treatments", assigned at random. In one version ($N = 100$) we asked respondents to describe a recent encounter with a live customer service agent. In the other version ($N = 98$) we asked a different set of respondents to describe an encounter with a chatbot. Respondents were US-based (55% female, average age: 33). All respondents received a show-up payment of $2.00 and an additional payment of $2.00 at the end of the study.

2.1.2. Survey 2 The second wave was also conducted on the Prolific platform and also consisted of two between-subject "treatments" ($N = 106$ and $N = 96$), analogous to Survey 1. Respondents were US-based (49% female, average age: 38). They received a show up payment of $3.00 and an additional payment of $2.00 at the end of the study. The questions were similar to Survey 1, with two key differences. First, we asked a more specific, free-form question about what had driven the respondent's decision to choose the chatbot (a live agent) channel. Second, we included a more structured question regarding the respondents' attitudes towards algorithms.

2.2. Results

2.2.1. Survey 1 The survey responses reveal a wide range of problem types and settings in which both live agents and chatbots are used, from e-commerce and technology-related issues, to travel and transportation, food and groceries, and finance and banking. The descriptions of these interactions further suggest two common themes. First, chatbots require minimal waiting to start the interaction. Users appreciated the instant availability of the chatbot and commented on the expediency with which their request was handled:

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

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

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

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

Table 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 ($p \ll 0.01$).[2] 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.

3. Utility Framework, Literature and Experiment Design

The survey data in §2 suggest a key trade-off in customer channel choice: chatbots require minimal waiting but fail frequently, whereas live agents typically require some time to access, but are more reliable in resolving customer requests. These survey insights guide our theoretical model, as well as our experimental design. To position our experiments within a broader conceptual framework,

[2] The majority of the chatbot users with unresolved requests were either transferred to a live agent (23.38%) or had to call a live agent (29.87%), thus spending additional time in the system until their request was resolved. See Appendix EC.2 for detailed data.

Table 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.

we first present a utility model that describes customer preferences in this setting and discuss how various behavioral constructs identified in prior literature would affect model parameters. We then use the model to develop testable hypotheses and present our experimental design.

3.1. A general utility model

Denote by u^j the utility received by a customer choosing service channel $j \in \{A, B\}$, where A denotes a live agent channel and B denotes a chatbot channel. The value of u^j depends on the reward r obtained by the customer from receiving service, minus the disutility of waiting, $c^j(\cdot)$. The reward r can represent having one's billing issue resolved, recovering a lost online password, scheduling an appointment, or some other inbound customer service request. The value of r is assumed to be known ex-ante and constant across channels.

We restrict our attention to processes with at most two service stages: a gatekeeper stage and an expert stage. The customer can incur waiting costs in four possible ways: waiting in line for a gatekeeper, waiting in service with a gatekeeper, waiting in line for an expert, and waiting in service with an expert. The total cost of waiting, $c^j(\cdot)$, can therefore depend on the durations of each of these four waits, which we denote by $t^j_{line_1}$, $t^j_{serve_1}$, $t^j_{line_2}$, $t^j_{serve_2}$, respectively. Denote by p^j the probability of successful request resolution in the first (gatekeeper) stage. Further, denote by $\boldsymbol{\theta}$ the vector of preference parameters, which describe customer preferences towards risk, interacting with gatekeepers, interacting with algorithms, and other factors. Then, the expected utility $\mathbb{E}[u^j]$ can be expressed as follows:

$$\mathbb{E}[u^j] = r - \mathbb{E}_{p^j}\left[c^j\left(t^j_{line_1}, t^j_{serve_1}, t^j_{line_2}, t^j_{serve_2} | \boldsymbol{\theta}\right)\right]. \tag{3.1}$$

3.2. Power Utility

Within the general utility framework in (3.1), we will focus on convex waiting cost functions $c^j(\cdot)$. Convex mappings between the time spent and the cost of waiting have been proposed and examined

in the queuing literature (Dewan and Mendelson 1990, Van Mieghem 2000, Veeraraghavan and Debo 2009, Ata and Olsen 2009).[3] The advantage of a convex cost function is that it can be used to model risk aversion – one of potential drivers of behavior in our setting, where one channel has a significantly higher failure rate (and thus risk) than the other. To fix ideas and develop hypotheses, it is useful to specify a functional form for $c^j(\cdot)$. Specifically, we will model channel waiting costs as a power transformation of a weighted sum of stage-specific durations. This functional form has the advantage of accommodating a variety of behavioral constructs that emerge from the literature, which we discuss in §3.3. We will consider several alternative specifications of $c^j(\cdot)$ in §7.

The waiting cost functions can be further simplified based on our survey results in §2. In particular, we have seen that the majority of respondents did not have to wait to access a chatbot; therefore, we will assume that $t^B_{line_1} = 0$. Further, we have seen that the majority of customers get their requests resolved on the first attempt when contacting a human service provider. In particular, our survey data showed request resolution rates of approximately 87% for the live channel vs. 42% for chatbots (§2.2.2). To focus on the key trade-offs, in our experiments we will assume that $p^A = 1$, while $p^B < 1$. After these simplifications, the channel-specific cost functions are as follows:

$$\mathbb{E}[c^A] = \left(\alpha^A_{line_1} t^A_{line_1} + \alpha^A_{serve_1} t^A_{serve_1}\right)^\gamma \tag{3.2}$$

$$\mathbb{E}[c^B] = p^B \left(\alpha^B_{serve_1} t^B_{serve_1}\right)^\gamma + (1-p^B)\left(\alpha^B_{serve_1} t^B_{serve_1} + \alpha^B_{line_2} t^B_{line_2} + \alpha^B_{serve_2} t^B_{serve_2}\right)^\gamma \tag{3.3}$$

where $\boldsymbol{\theta} = \{\alpha^A_{line_1}, \alpha^A_{serve_1}, \alpha^B_{serve_1}, \alpha^B_{line_2}, \alpha^B_{serve_2}, \gamma\}$ is the preference parameter vector. Note that if all α-parameters are equal and $\gamma = 1$, the problem simplifies to choosing the channel that minimizes total expected time spent in the system – a risk-neutral optimization benchmark that will serve as the null hypothesis for our experiments.

3.3. Behavioral factors

While our null is based on expected time minimization, prior work in marketing, behavioral operations, and decision theory has identified several behaviors that suggest potential deviations from this benchmark.

(i) **Risk Aversion** While risk preferences for money have been studied extensively (Holt and Laury 2002, Eckel and Grossman 2008, Harrison and Cox 2008, Kagel and Roth 2020), research on risk preferences in the time domain is limited. Some existing work invokes Prospect Theory (Kahneman and Tversky 1979, Tversky and Kahneman 1992) and suggests that since time expenditures are losses, people may be risk-neutral (Kroll and Vogt 2008) or even risk-seeking

[3] The use of convex cost functions is typically motivated by customer impatience, i.e., escalating displeasure with waiting (Shimkin and Mandelbaum 2004, Janakiraman et al. 2011), rather than by customer risk attitudes. A notable exception is Afèche et al. (2013), who model a joint time-money utility function that accommodates risk aversion.

for time (Abdellaoui and Kemel 2014). More frequently, however, research finds risk-averse behaviors in the time domain (Leclerc et al. 1995, Festjens et al. 2015, Flicker and Hannigan 2022). Note that much of this research is based on hypothetical decisions, and involves longer time intervals than those typical for the customer service setting. However, if risk aversion is present in our data, γ would be greater than 1, resulting in the cost function being convex in the sum of weighted stage costs.

(ii) **Gatekeeper Aversion** We use the term "gatekeeper aversion" to describe a preference for a more continuous, barrier-free service process with a single server, as opposed to a more fragmented process that may involve transfers and multiple service stages. Multi-stage waiting experiences were studied by Carmon and Kahneman (1996), Kumar et al. (1997), Soman and Shi (2003) 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. A notable exception is Buell (2021) who measures queue-switching behaviors as a response to one's position in a line.

Most relevant to our setting are Soman and Shi (2003) and Buell (2021). Similar to Soman and Shi (2003), customers in our setting may have a preference for moving continuously towards service completion and may experience a disutility from interruptions or slowdowns, even when the total duration of the wait is the same. Similar to Buell (2021), entering a new stage of a wait *after* a service stage has been completed (as in Channel B) may be perceived as psychologically equivalent to repositioning from the front and towards the back of the line. Therefore, for a fixed duration, customers may experience greater cost of waiting in line *after* chatbot failure than waiting in line *before* being served in the human channel. If this is the case, we would have $\alpha^A_{line_1} < \alpha^B_{line_2}$.

(iii) **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 other tasks in which worker performance can be augmented by an algorithm. Different from these studies, we focus on customer (rather than on worker) adoption of algorithmic technology. Most relevant to our study is Castelo et al. (2019), who find that people are especially averse to algorithms when the task at hand involves a subjective, interpersonal or taste-based component, as would be the case in many customer service settings. Also of relevance is the operations and marketing literature that finds that self-service technology adoption may have important effects on sales, customer loyalty, and satisfaction (Curran and Meuter 2005, Buell

et al. 2010, Tan and Netessine 2020, Castelo et al. 2023). Failures of self-service machines can trigger algorithm aversion (Chen et al. 2021), and disclosing the algorithmic nature of sales agents reduces sales (Luo et al. 2019). This literature suggests that the cost of waiting may be larger when the server is algorithmic than when it is human, i.e., $\alpha^B_{serve_1} > \alpha^A_{serve_1}$.

(iv) **Other Factors** There are other factors that may play a role in evaluating the cost of waiting. For example, waiting actively while being served may be preferred to waiting passively in line (Maister et al. 1984). Buell and Norton (2011) show that customers have a preference for seeing that service is being performed while waiting. In our (customer service) setting, waiting episodes in which customers wait in line may generate a stronger aversive response than episodes in which customers are being actively served, even when the time spent is the same. Such preferences may favor the chatbot channel, which involves less waiting in line. If this factor is sufficiently strong (relative to factors (ii) and (iii)), then we may observe that $\alpha^j_{line_k} > \alpha^j_{serve_k}$ for $j = \{A, B\}$ and $k = \{1, 2\}$.

3.4. Experiment Design and Hypotheses

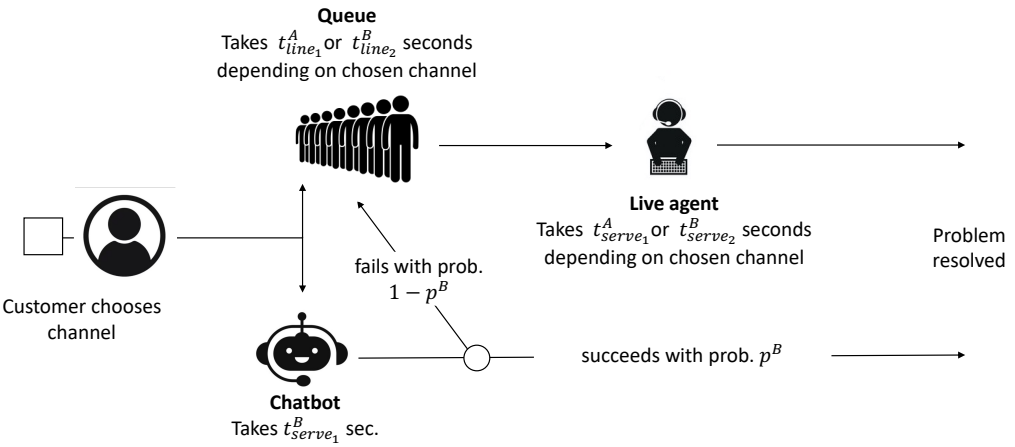
Our experiments examine the extent to which the constructs described in §3.3 affect channel choices.

All experiments: In all experiments, the decision-maker makes a series of binary choices between the live agent channel (Channel A) and the bot channel (Channel B). Channel A requires some waiting in line but resolves the request with certainty, while Channel B can be accessed immediately but is less reliable and transfers a large portion of requests to a live agent. The choice is shown in Figure 1. The choice is repeated for a range of problem parametrizations. In particular, we vary Channel B performance (specifically, $t^B_{serve_1}, t^B_{line_2}$ and p^B), as well as several qualitative features of the channels. Participants are incentivized to report their preferences truthfully by having to experience several of their choices (in real-time) before receiving their payments.

Experiment 1: This experiment examines channel choices in two settings: (1) a neutral, context-free setting in which the live-agent/chatbot nature of the channel is muted, and (2) a contextualized setting in which the algorithmic nature of the chatbot is explicitly disclosed. The experiment consists of two treatments:

Baseline Treatment. This treatment presents participants with a setting where the choice is between two unnamed waiting formats. The channel choice therefore comes down to a comparison of a channel with a deterministic duration (an in-line stage then an in-service stage) vs. a channel with an uncertain duration (either one in-service stage or an in-service stage, an in-line stage, then an in-service stage).

Figure 1 Channel Choice



Context Treatment. This treatment is analogous to *Baseline*, but presents participants with a choice between two named formats: the "Live Agent" or the "Chatbot" format. While the actions required to complete service in each channel are identical to those of *Baseline*, the experience is contextualized through pictures and animations to portray the back-and-forth between the server and participant.

Experiment 1 will allow us to test two hypotheses:

H1.1: *Participants in the Baseline treatment choose the channel that minimizes the total expected time spent in the system.*

H1.2: *Participants in the Context treatment choose each channel at the same rate as participants in the Baseline treatment.*

If H1.1 is rejected, then we can conclude that some combination of gatekeeper and risk aversion (items (i) and (ii) in §3.3) affects channel choices. If H1.2 is rejected, then we can conclude that algorithm aversion (item (iii) in §3.3), also plays a significant role.

Experiment 2: The second experiment consists of a single, *Live* Treatment. Similar to *Context*, this treatment also presents participants with a contextualized choice between the two channels. However, in this treatment, we employ real humans (research assistants, blind to the experimental hypotheses) who play the role of live agents and interact with participants using a live chat tool. The remainder of this treatment is identical to the *Context* treatment. This experiment will allow us to test the following hypothesis:

H2: *Participants in the Live treatment choose each channel at the same rate as participants in the Baseline treatment.*

If H2 is rejected, then we can conclude that introducing a live human into the service process affects channel choices. Further, if H1.2 is supported and H2 is rejected, we can conclude that

although adding contextual cues is insufficient to alter behavior, a more realistic implementation involving actual humans acting as live agents can be effective. This distinction will help differentiate between potential algorithm aversion from a mere association and algorithm aversion resulting from the experiential difference.

Experiment 3: The purpose of the third experiment is to separately identify risk aversion and gatekeeper aversion. To do so we conduct the *Deterministic* treatment. This treatment is analogous to the *Baseline* treatment in that it does not present participants with any contextual information. However, different from the *Baseline* treatment, the chatbot channel is deterministic. In particular, in the chatbot channel, the initial service stage always fails and the customer needs to wait in line for the second service stage. This treatment allows us to separate between risk aversion and gatekeeper aversion. If risk aversion is the dominant mechanism, then we would observe expected time minimizing choices in the *Deterministic* treatment data. Conversely, if gatekeeper aversion plays a role even in the absence of risk, then we will observe a significant deviation between expected time minimizing choices and the *Deterministic* treatment data. Formally, we test the following hypotheses:

H3.1: *Participants in the Deterministic treatment choose each channel at the same rate as participants in the Baseline treatment.*

H3.2: *Participants in the Deterministic treatment choose the channel that minimizes the total expected time spent in the system.*

The hypotheses are summarized in Table 2. In §4-6 we use reduced form statistical tests (tests of equality of means, random effects regressions) of H1-H3 to understand the relevance of each mechanism. Then, in §7, we pool our treatment data and estimate $\boldsymbol{\theta}$ (eq. (3.1)) to re-examine observed behaviors within a common utility framework and to assess their relative size and importance.

Table 2 Summary of Hypotheses and Tests

Behavioral construct	**Hypotheses**	**Treatment comparison (§4-6)**	**Parameter predictions based on expected time minimization**	**Parameter predictions based on behavioral literature (§3.3)**
Risk aversion	H1.1, H3.1	*Baseline* = *RN benchmark* (§4), *Baseline* = *Deterministic* (§6)	$\gamma = 1$	$\gamma > 1$
Gatekeeper aversion	H3.2	*Deterministic* = *RN benchmark* (§6)	$\alpha^A_{line_1} = \alpha^B_{line_2}$	$\alpha^A_{line_1} < \alpha^B_{line_2}$
Algorithm aversion	H1.2, H2	*Baseline* = *Context* (§4), *Baseline* = *Live* (§5)	$\alpha^A_{serve_1} = \alpha^B_{serve_1}$	$\alpha^A_{serve_1} < \alpha^B_{serve_1}$

Note: RN Benchmark is the theoretical chatbot uptake predicted by risk-neutral time minimization.

4. Experiment 1

In Experiment 1 we examine behaviors in two settings: one with a neutral framing where participants choose among two unnamed formats (*Baseline* treatment), and one where the same choices are made but the setting is contextualized (*Context* treatment).

4.1. Methodology

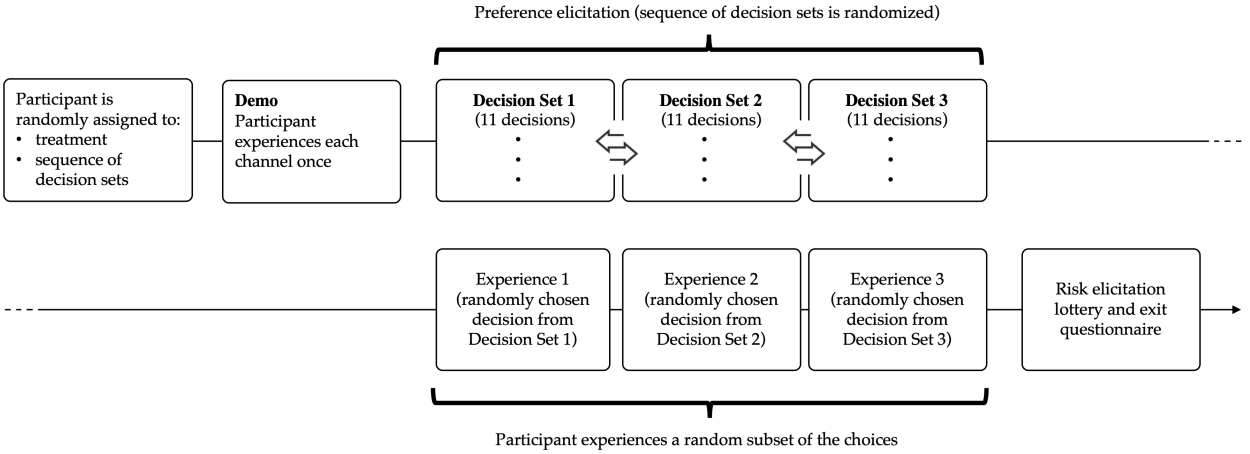
4.1.1. Channel choice We begin by describing the fundamentals of the channel choice, as shown in Figure 1. To access the server in Channel A, the participant first needs to wait in line. The wait in line takes $t^A_{line_1}$ seconds, and the wait in service takes $t^A_{serve_1}$ seconds. The server always succeeds in resolving the request and the participant exits the system. In Channel B, there is no line to start service, so that the participant immediately proceeds to interacting with the chatbot and spends $t^B_{serve_1}$ seconds to complete this interaction. The chatbot has limited problem-solving skills, resulting in some portion of chatbot interactions being redirected to a live agent, with p^B denoting the probability of chatbot success. If the chatbot succeeds in resolving the request, the participants exits the system. If the chatbot fails, the participant has to wait in line for $t^B_{line_2}$ seconds and then in service with a live agent agent for $t^B_{serve_2}$ seconds. After that, the participant exits the system.

4.1.2. Experiment protocol Figure 2 presents the experiment protocol. After being randomly assigned to a treatment, participants first experience a demo of both channels. The Channel A demo begins with a wait in line. While waiting in line, participants see a horizontal progress bar fill from left to right. After that, participants proceed to service, represented by a circular progress bar (to distinguish it from waiting in line). During service, participants are prompted to click on pre-specified keys for the progress bar to advance. They receive three such prompts, equally spaced throughout the service time. The Channel B demo is analogous and includes both the successful and failed chatbot resolution scenarios. During the demo, all stages last 20 seconds.

After the demo, participants make a total of 33 decisions, subdivided into three decision sets of 11 decisions, with varying parameters. The sequence of decision sets was randomized to control for any order effects. Each of the 33 decisions is a binary choice between Channel A and Channel B. The parameters $p^B, t^B_{serve_1}$ and $t^B_{line_2}$ vary across the 33 decisions in ways that will be discussed next.

4.1.3. Elicitation Table 3 presents the parameters used in the experiment. Within each decision set, we used the Multiple Price Lottery mechanism (Holt and Laury 2002) to elicit preferences. The basic idea of this mechanism is to present participants with a list of binary decisions, where one of the alternatives becomes more desirable as one goes down the list. In Decision Set 1 we varied the success rate of the chatbot (p^B) in steps of 0.05 from 0.25 to 0.75 percent, while all other

Figure 2 Experiment Protocol



parameters were held constant. In Decision Set 2 we varied the service time of the chatbot ($t^B_{serve_1}$) in steps of 2 from 30 to 10 seconds. In Decision Set 3 we varied the time spent in line if the chatbot fails ($t^B_{line_2}$) in steps of 4 from 40 to 0 seconds. Across all three decision sets, we kept constant the difference in expected times between the two alternatives for each decision within a decision set (i.e., Decision 1 in Decision Set 1 has the same expected time difference between channels as Decision 1 in Decision Set 2 and Decision 1 in Decision Set 3, and similarly for the remaining 10 decisions).

4.1.4. Incentives After a participant completed all their decisions, a subset of the participant's decisions was selected to be experienced in real time. Specifically, one decision from each decision set was selected at random for the real experience, resulting in each participant experiencing three of their 33 choices prior to receiving their (fixed) dollar payment and exiting the experiment. Thus, participants were incentivized to report their true preferences. Participants received \$1.50 as a show up fee, and an additional \$3 at the end of the experiment, once they had completed all their decisions and waiting experiences. The average time spent in the experiment was 18 minutes and the average payment was \$5.70.[4]

4.1.5. Treatments and Participants Experiment 1 consisted of two treatments: *Baseline* and *Context*. Both treatments presented participants with the same choices and parametrizations. In the *Baseline* treatment, choices were context-free and the interaction between participants and servers was programmed to look identical across channels (see Figure 3b-c). In the *Context* treatment, choices were contextualized. In the instructions and on choice screens, participants were

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

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 3d-e). The prompts appeared

Table 3 Experimental Parameters

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

	Channel A		Channel B				Expected time minimizing choice
	$t^A_{line_1}$	$t^A_{serce_1}$	p^B	$t^B_{serve_1}$	$t^B_{line_2}$	$t^B_{serve_2}$	
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
	$t^A_{line_1}$	$t^A_{serve_1}$	p^B	$t^B_{serve_1}$	$t^B_{line_2}$	$t^B_{serve_2}$	
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 ($t^B_{line_2}$)

	Channel A		Channel B				Expected time minimizing choice
	$t^A_{line_1}$	$t^A_{serve_1}$	p^B	$t^B_{serve_1}$	$t^B_{line_2}$	$t^B_{serve_2}$	
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.

in equally spaced intervals with a total of three prompts per interaction, and the progress bar resumed only after the participant responded to the prompt.

A total of 216 participants were recruited on Prolific; 199 of them passed the screening questions and attention checks and were admitted to the experiment. Each Prolific worker was restricted to participating in one session only. US-based workers with an approval rating of at least 98% were recruited. All experiments were programmed in oTree (Chen et al. 2016).

4.2. Results

Before presenting our results we comment briefly on the overall level of consistency in the collected data. We refer to a participant as "consistent" if, for each decision set in Table 3, their Channel B uptake is always weakly increasing in p^B and weakly decreasing in $t^B_{serve_1}$ and $t^B_{line_2}$. In our data, 166 out of 199 participants (83.41%) are consistent. We present our hypothesis tests both with the full sample and with the sample of consistent participants only.[5]

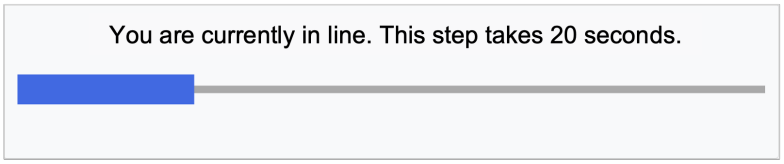
4.2.1. Descriptive Statistics Figure 4 shows the share of participants choosing Channel B in each of the 33 decisions. Only consistent participants are included. As one may expect, the proportion of participants choosing Channel B increases with p^B and decreases with $t^B_{serve_1}$ and $t^B_{line_2}$. Consistent with expected time minimization, very few participants (between 0% and 10%) choose Channel B when the parameters are such that the expected time in Channel B is higher than the expected time in Channel A ($p^B \leq 0.45$ in panel a, $t^B_{serve_2} \geq 22$ in panel b, $t^B_{line_2} \geq 24$ in panel c). However, for decisions where both channels have the same expected waiting time ($p^B = 0.5$ in panel a, $t^B_{serve_1} = 20$ in panel b, $t^B_{line_2} = 20$ in panel c), there is a preference for Channel A – over 75% of participants choose Channel A in those decisions. Further, a large group of participants (between 24% and 30%, depending on decision set) chooses Channel A for all 11 decisions. Finally, there are only minimal differences in channel uptake between the two treatments.

To better understand channel choices it is insightful to divide participants into discrete types. The type summary is shown in Table 4. First, consider the *Baseline* treatment. In this treatment, 17.28% of participants never choose the chatbot channel and 45.68% of participants choose the chatbot channel less frequently than predicted under expected time minimization in at least one decision set and are expected time minimizers in the remaining decisions. The third group minimizes expected wait across all three decision sets and the fourth group collects remaining participants, with each group containing fewer than 20% of participants. Thus, the majority of participants (62.96%) fall into categories (1) and (2), both of which avoid Channel B (gatekeeper channel)

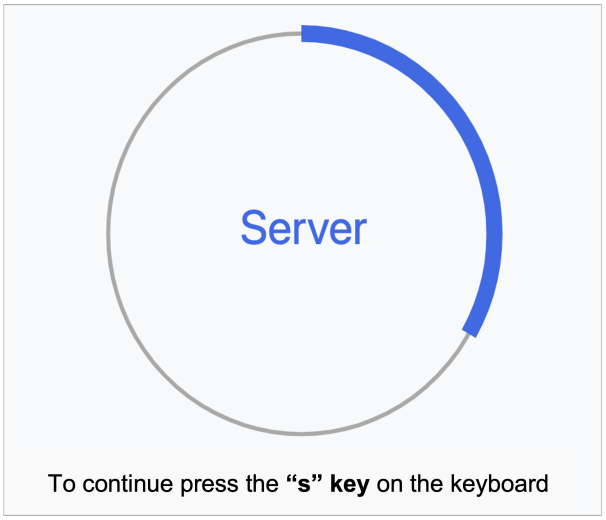
[5] Excluding inconsistent participants from the data is common practice in the experimental literature (see Charness et al. 2013, for comprehensive discussion). The share of consistent participants in our data compares favorably to the numbers reported in the prior literature using our elicitation method (Holt and Laury 2002, Charness et al. 2013), suggesting high levels of attention and engagement with the experimental stimuli.

Figure 3 Waiting Experiences (Screenshots)

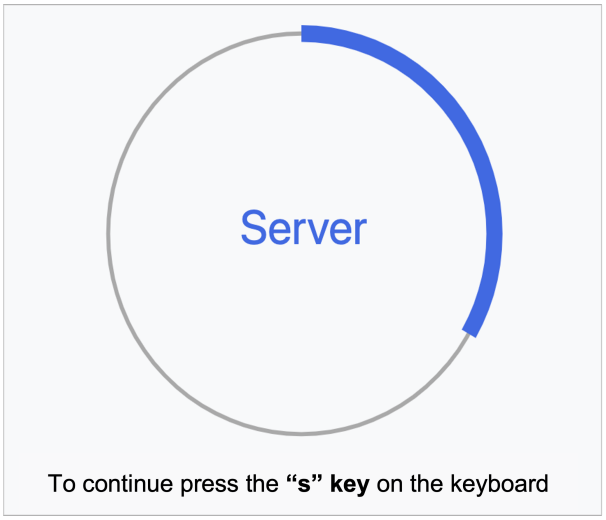
(a) All Treatments, In-line Waiting Stage



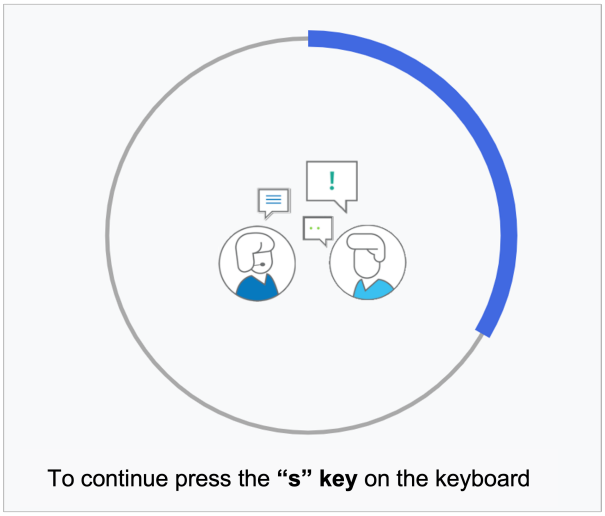
(b) *Baseline* Treatment, Interaction with a "Live Agent"



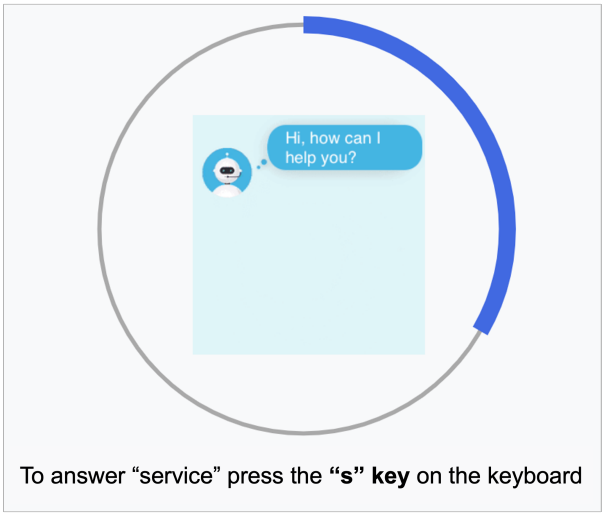
(c) *Baseline* Treatment, Interaction with a "Chatbot"



(d) *Context* Treatment, Interaction with a Live Agent



(e) *Context* Treatment, Interaction with a Chatbot



even in situations when it has a shorter expected time relative to Channel A. The numbers are quite similar in the *Context* treatment, with none of the treatment differences in proportions being statistically significant (at the 0.05 level).

4.2.2. Hypothesis Tests We next test H1.1 and H1.2, i.e., examine whether channel choices are consistent with expected time minimization, and whether they are affected by the presence

Table 4 Participant Types in Experiment 1

	No. of Channel A choices per decision set	No. of Channel B choices per decision set	Participant Type (1) Never chooses Channel B	(2) Underutilizes Channel B	(3) Expected time minimizer	(4) Other
Baseline treatment	7.72	3.28	17.28%	45.68%	18.52%	18.52%
Context treatment	7.76	3.24	14.23%	52.94%	8.24%	24.71%

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

of context. To test H1.1 we use $t-$tests and compare observed channel uptake in the *Baseline* treatment with 5.5, the expected time minimization benchmark. Our data strongly reject H1.1: average Channel A uptake is 7.59 in Decision Set 1, 7.71 in Decision Set 2 and 7.86 in Decision

Figure 4 Channel B uptake in Experiment 1

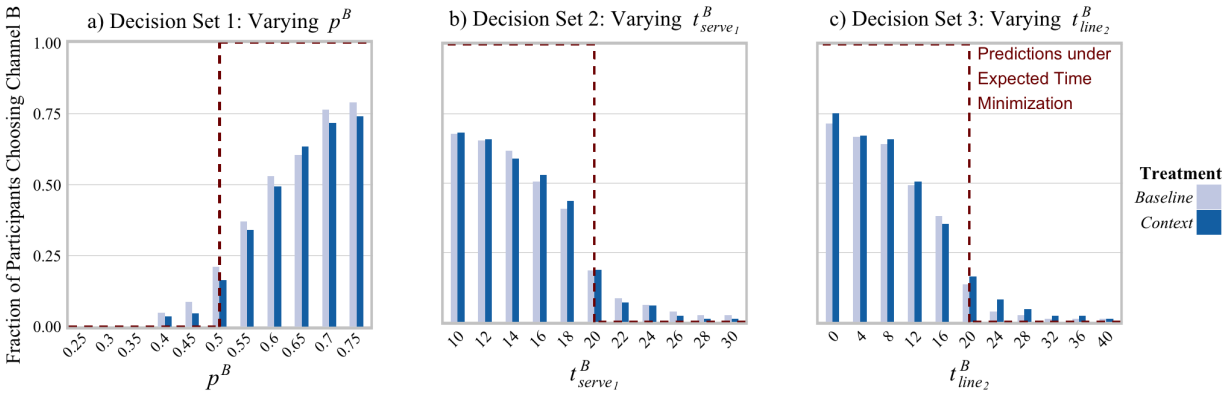


Table 5 Channel Preferences in Experiment 1

Dependent Variable:	(1) Channel B	(2) Channel B
Baseline	-	-
Context	-0.081	-0.267
	(0.306)	(0.561)
Channel B Performance Controls? ($p^B, t^B_{serve_1}, t^B_{line_2}$)	Yes	Yes
Demographic Controls?	Yes	Yes
Sample	All subjects	Consistent subjects
Observations	6567	5478
Subjects	199	166

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

Set 3 (each number is significantly different from 5.5, with each $p \ll 0.01$).[6] To test H1.2 we use random effects regressions, in which we examine the *Context* treatment coefficient.[7] The regression coefficients are reported in Table 5. Performing the tests for the full sample of participants (column (1)), or for the sample of consistent participants (column (2)) yields the same result: the *Context* treatment does not significantly affect chatbot uptake ($p = 0.790$ and $p = 0.634$). Thus, our first result is as follows:

Result 1: *H1.1 is rejected. Channel B uptake is lower than expected time minimization would predict. H1.2 is supported. Channel B uptake is not significantly different between the Baseline and Context conditions.*

4.3. Discussion

The results of Experiment 1 provide valuable insights into the factors influencing channel choice. Participants' choices deviate significantly from the expected time minimization benchmark, with a strong preference for Channel A (human agent) even when Channel B (chatbot) offers a shorter expected wait time. The majority of participants (63% to 67% depending on the treatment, see Table 4) consistently chose the channel with longer expected waiting times, suggesting that the aversion to channels with a gatekeeper structure affects a significant portion of the population.

Interestingly, the presence of contextual information in the *Context* treatment did not significantly alter behavior. This result is somewhat surprising, as one might expect that providing participants with a more realistic and engaging decision environment would influence their preferences. Indeed, prior literature on algorithm aversion shows that asking people to work together with algorithms may elicit an aversive response, even when the performance of such systems is the same as the performance of humans (Dietvorst et al. 2015, Castelo et al. 2019). Our results so far suggest that, in the customer service domain, such attitudes appear to be relatively weak. In the next two sections we will further examine both the role of algorithm aversion (§5), and the role of more process-related factors, such as risk aversion and gatekeeper aversion (§6).

5. Experiment 2

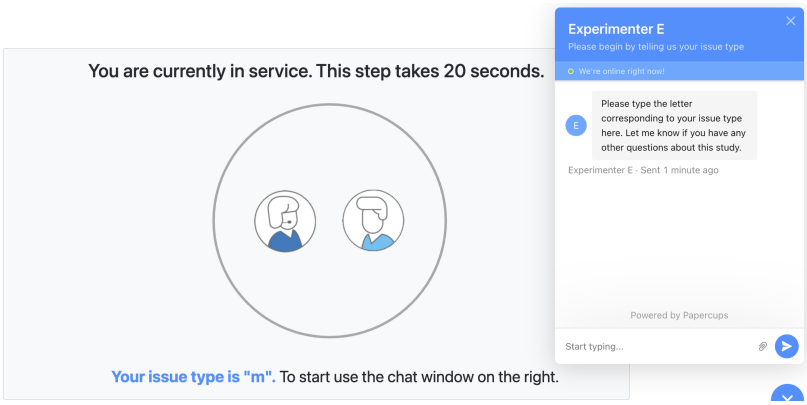
We have so far examined algorithmic attitudes using simple contextual cues (*Context* treatment in Experiment 1). While capturing one aspect of the channel choice, this manipulation may not fully mimic the differences of interacting with a human vs. with a machine. We next examine a setting where Channel A, representing the live agent, is staffed by actual live servers (research assistants).

[6] H1.1 continues to be rejected at $p \ll 0.01$ if we assume 6.0 instead of 5.5 as the benchmark for expected time minimization, which would be consistent with choosing the riskless Channel A in Decision 6, in which both channels yield the same expected time.

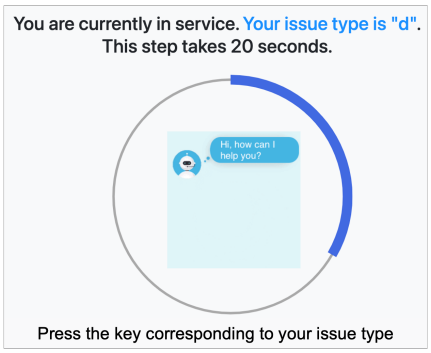
[7] Random effects regressions are more appropriate to test H1.2 due to the panel structure of the data. However, simple tests of equality of treatment means yield the same results as the regression-based tests presented in the text.

Figure 5 Screenshots of the Experiment

(a) Interaction with a Live Agent



(b) Interaction with a Chatbot



5.1. Methodology

5.1.1. Experiment Design We recruited 116 participants on Prolific; a total of 108 passed the comprehension checks. Participation was restricted to Prolific workers who did not participate in prior studies. The experiment protocol was identical to the *Context* treatment in Experiment 1, with the difference being that the role of the live agent was now played by an experimenter. To this end, we recruited two research assistants who had no prior knowledge of the hypotheses. We followed common practices for deploying confederates (research assistants) in experimental research (Kuhlen and Brennan 2013) and trained the assistants using a script to control for potential differences in communication patterns. The training materials are in EC.4.2.

The experiment consisted of a single treatment, which we will refer to as the *Live* treatment. As in Experiment 1, participants made 33 decisions, with three of these decisions being implemented after the decisions were submitted. To provide some context for the customer support interactions, participants were assigned an issue type at the beginning of each interaction. The issue type was represented through a letter of the alphabet. At the beginning of each service stage, participants were asked to enter that letter into the live chat window to begin service.

The live-chat interface is shown in Figure 5(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 (the 20 seconds in the example in the Figure 5(a) screenshot) 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; however, to keep the framing constant across channels, we also required participants to enter their issue type when starting the interaction (see Fig. 5(b)).

5.2. Results

Table 6 Participant Types in Experiment 2

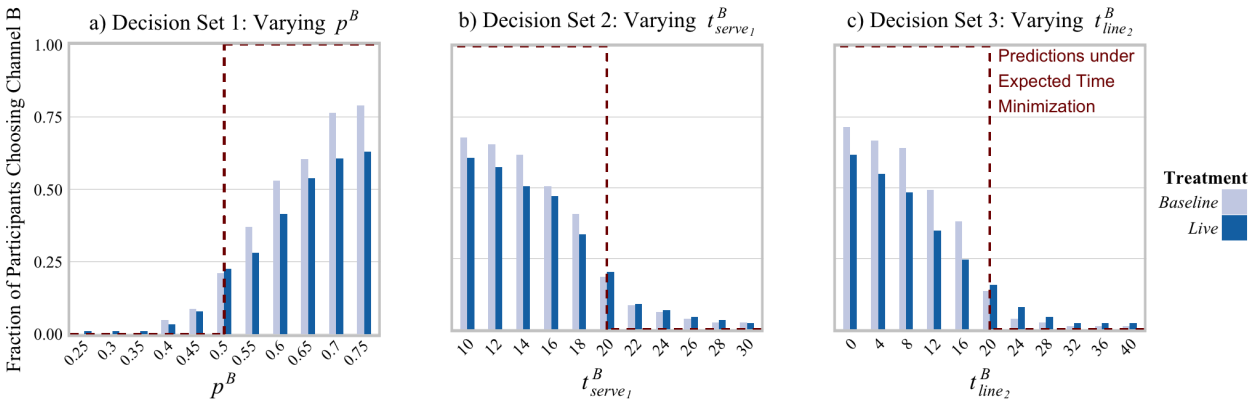
	No. of Channel A choices per decision set	No. of Channel B choices per decision set	Participant Type (1) Never chooses Channel B	(2) Underutilizes Channel B	(3) Expected time minimizer	(4) Other
Baseline treatment	7.72	3.28	17.28%	45.68%	18.52%	18.52%
Live treatment	8.20	2.80	28.09%	40.45%	7.87%	23.60%

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

5.2.1. Descriptive Statistics Figure 6 summarizes the choices. As before, we highlight the hypothetical decisions that would result from all participants minimizing expected time (dashed lines). Several observations are in order. First, in both treatments, Channel B uptake increases with Channel B performance. Second, there are some differences between treatments. In particular, there appears to be a consistent gap in chatbot uptake, especially for decisions where the chatbot channel results in expected time savings. Table 6 shows the prevalence of each participant type in the data. We observe that the number of participants never choosing Channel B jumps from 17.28% in *Baseline* to 28.09% in the *Live* treatment (proportion test $p = 0.047$), while the share of expected time minimizers drops from 18.52% to 7.87% (proportion test $p = 0.039$). This indicates that the human nature of the server prompts a significant portion of decision-makers to *never* consider the chatbot channel.

5.2.2. Hypothesis Tests To test for the presence of algorithm aversion (H2), we regressed channel choice on treatment while controlling for chatbot performance ($p^B, t^B_{serve_1}, t^B_{line_2}$). Table 7 presents the estimates. To examine the hypothesized effects of the human live agent relative to a neutral benchmark, we included the *Baseline* treatment. We also included the *Context* treatment, which presents participants with a similarly rich, contextualized decision without involving a real

Figure 6 Channel B Uptake in Experiment 2



human server. The results suggest that the *Live* treatment leads to a decrease in Channel B uptake. In columns (1) and (3), which focus on average treatment effects, the size of the effect ranges between 5.41 and 5.86 percentage points. The effect is marginally significant ($p = 0.074$ and $p = 0.076$).

Table 7 Channel Preferences in Experiment 2

Dependent Variable:	(1) Channel B	(2) Channel B	(3) Channel B	(4) Channel B
Baseline	-	-	-	-
Context	-0.0812	-0.0571	-0.243	-0.174
	(0.341)	(0.338)	(0.607)	(0.599)
Live	-0.611*	-0.606*	-1.067*	-1.033*
	(0.342)	(0.339)	(0.602)	(0.593)
Age	-0.0328***	0.00446	-0.0443**	0.0355
	(0.0114)	(0.0214)	(0.0200)	(0.0405)
Age × *Context*		-0.0297		-0.0654
		(0.0282)		(0.0516)
Age × *Live*		-0.0761***		-0.153***
		(0.0290)		(0.0522)
Channel B Performance Controls? ($p^B, t^B_{serve_1}, t^B_{line_2}$)	Yes	Yes	Yes	Yes
Demographic Controls?	Yes	Yes	Yes	Yes
Sample	All subjects	All subjects	Consistent subjects	Consistent subjects
Observations	10032	10032	8415	8415
Subjects	304	304	255	255
Marginal effect of *Live* at...		***p*−value:**		***p*−value:**
Age= 20		0.334		0.204
Age= 25		0.736		0.541
Age= 30		0.532		0.672
Age= 35		0.071		0.078
Age= 40		0.007		0.005
Age= 45		0.002		0.001
Age= 50		0.002		0.001
Age= 55		0.002		0.001
Age= 60		0.002		0.001

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

To obtain further insights we conducted additional post-hoc tests, specifically examining whether the treatment effect was more pronounced for certain demographics, particularly age and gender.[8] Our analysis reveals that participant age is indeed a significant moderator of the *Live* treatment effect. Columns (2) and (4) of Table 7 show that the interaction term between age and the *Live* treatment dummy is statistically significant at $p \ll 0.01$. Further, examining the marginal effect

[8] We note that our participants come from a variety of age groups, ranging from 18 to 84 years old, mean: 36.33.)

of the *Live* treatment conditional on age, we find statistically significant effects (at $p < 0.05$) for participants aged 40 and above. Conversely, the treatment effect is not significant for those under 40 (see the bottom of Table 7 for details).

Result 2 (Algorithm Aversion, cont.): *Hypothesis 2 is (weakly) supported. Channel B uptake is reduced under the Live manipulation. The effect is statistically significant for participants aged 40 and above.*

5.3. Discussion

The results of Experiment 2 provide several new insights. First, Channel B uptake remains below the expected time minimization benchmark, consistent with the findings from Experiment 1. Second, there is a marginally significant decrease in Channel B uptake relative to the *Baseline* case. This decrease is driven primarily by participants who completely reject Channel B and are unresponsive to performance improvements in the chatbot technology (i.e., type (1) in Table 6). Notably, the participants rejecting Channel B are significantly older than the rest. Third, even after accounting for age effects, the magnitude of algorithm aversion is small compared to the gap between the risk-neutral benchmark and the *Baseline* treatment identified in Experiment 1. Specifically, the difference in Channel A uptake between *Baseline* and *Live* is only $8.20 - 7.72 = 0.48$ (Table 6), which is less substantial than the difference between *Baseline* and the risk-neutral benchmark ($7.72 - 5.50 = 2.22$). In the following section, we will further unpack the mechanisms driving these results.

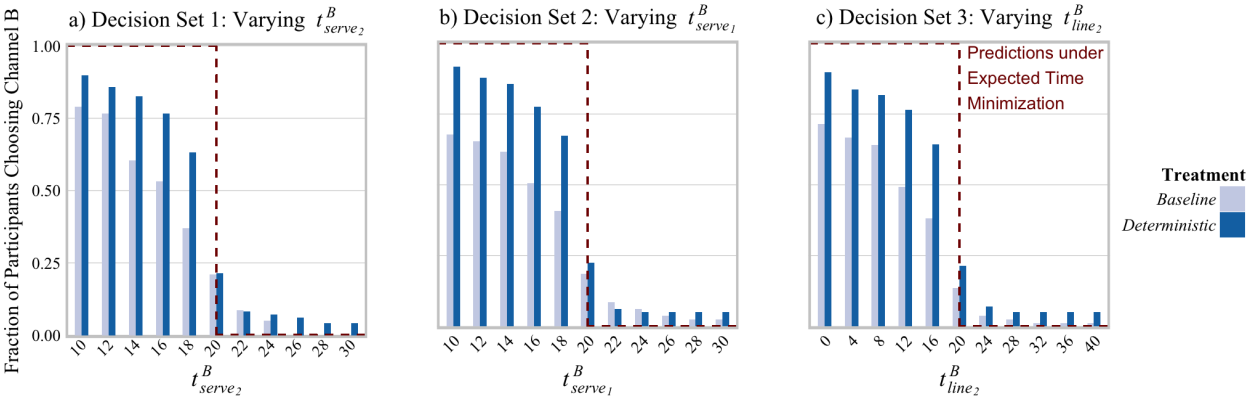
6. Experiment 3

We have so far seen that the bulk of Channel B (chatbot) underutilization is tied to process-related hurdles, rather than to the algorithmic nature of the chatbot channel. These hurdles include increased uncertainty, as well as the gatekeeper structure of the chatbot channel. To better understand the relative importance of each factor, we conducted a third experiment, in which *both* channels were deterministic in the amount of time spent by the customer in the system, so that any residual difference could be attributed to gatekeeper aversion.

6.1. Methodology

Experiment 3 consisted of a single treatment, which we will refer to as *Deterministic*. Similar to the *Baseline* treatment, participants chose between two channels: Channel A and Channel B. However, different from the *Baseline* treatment, the wait in both channels was deterministic. The experience in Channel A was identical to that of *Baseline*, consisting of a 20-second wait in line and a 20-second wait in service for all 33 decisions. In Channel B we adjusted the wait durations of each stage to match the expected waiting times in *Baseline* for each of the 33 decisions. In

Figure 7 Channel B Uptake in Experiment 3



particular, Channel B always involved three waits: a wait in service with the first server, a wait in line for the second server, and a wait in service with the second server. Thus, the *Deterministic* treatment maintained the gatekeeper structure of Channel B but removed the risk component.

As before, we varied $t^B_{serve_1}$ and $t^B_{line_2}$ to examine the response to different parametrizations. In addition, to keep the number of decisions consistent with the previous treatments, $t^B_{serve_2}$ was varied in one of the three decision sets (instead of p^B, which was now set to 1 across all decisions to make Channel B wait deterministic). The remaining parameters were unchanged. A total of 132 participants were recruited on Prolific, of which 121 passed comprehension checks and attention screens. Participation was restricted to workers who had not participated in prior studies.

6.2. Results

Figure 7 shows average Channel B uptake in the *Deterministic* treatment relative to the *Baseline* treatment. As in the previous experiments, the treatment differences are concentrated in the decisions where Channel B offers some time savings relative to Channel A. For these parametrizations, we observe a 10% to 25% increase in Channel B uptake. However, there remains a gap of approximately 10 to 40 percentage points (depending on parameters) between expected time minimization and observed decisions.

6.2.1. Descriptive Statistics Table 8 shows the average number of Channel B choices per decision set, as well as the split of types in the data. First, the average number of Channel B choices is 6.49 in the *Deterministic* treatment, which is approximately halfway between the risk-neutral benchmark of 5.5 and the average uptake of 7.72 in the *Baseline* treatment. Second, the share of participants of type (1), i.e, those who never choose Channel B, decreases from 17.28% to 6.49% (proportion test, $p = 0.018$). Further, the share of participants of type (2), i.e, those who sometimes underutilize Channel B, drops from 45.68% to 31.63% ($p = 0.054$). Finally, the share of participants of type (3), whose decisions are consistent with expected time minimization, increases from 18.52%

to 45.92% ($p \ll 0.001$). These summary statistics provide some preliminary evidence of both the effects of risk and gatekeeper structure on Channel B uptake.

6.2.2. Hypothesis Tests As before, we use $t-$tests and random effects logit regressions to test our hypotheses. Consider first H3.1, i.e., the hypothesis that chatbot uptake is not affected by risk. To test this hypothesis we regress chatbot uptake on chatbot performance and examine the coefficient of the *Deterministic* treatment. Because the expected times are exactly constant across the two treatments, the treatment coefficient provides a measure of the effect of risk. The regression coefficients are reported in Table 9. The results suggest that, contrary to H3.1, risk is a significant factor in determining channel preferences. In both specifications (column (1) and (2)) the effect is statistically significant at $p \ll 0.001$. To test H3.2., i.e., the hypothesis that Channel B uptake in the *Deterministic* treatment is consistent with expected time minimization, we compare average chatbot uptake with 5.5, the expected time minimization benchmark. Using $t-$tests, we are able to reject H3.2 at $p \ll 0.001$ for each of the three decision sets, with the full sample as well as with the restricted sample of consistent subjects.

Result 3: *H3.1 is rejected. Channel B uptake in the Deterministic treatment is significantly higher than in the Baseline treatment. H3.2 is also rejected. Channel B uptake in the Deterministic treatment is significantly below the expected time minimization benchmark.*

6.2.3. Discussion Experiment 3 allows us to separately identify risk aversion (preference against service formats with more uncertain waits) and gatekeeper aversion (preference against service formats with multiple service stages). Our results suggest that each of the two mechanisms plays a role in reducing Channel B (chatbot) uptake. Further, given that Channel B uptake in the *Deterministic* treatment is approximately halfway between the risk-neutral benchmark and that of the (risky) *Baseline* treatment, we can conclude that risk aversion and gatekeeper aversion have an approximately equal size in our data. In additional analyses we further examine the relationships between demographic variables and Channel B uptake and find no significant effects. Therefore,

Table 8 Participant Types in Experiment 3

			Participant Type			
	No. of Channel A choices per decision set	No. of Channel B choices per decision set	(1) Never chooses Channel B	(2) Underutilizes Channel B	(3) Expected time minimizer	(4) Other
Baseline treatment	7.72	3.28	17.28%	45.68%	18.52%	18.52%
Deterministic treatment	6.49	4.51	6.12%	31.63%	45.92%	16.33%

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

Table 9 Channel Preferences in Experiment 3

Dependent Variable:	(1) Channel B	(2) Channel B
Baseline	-	-
Deterministic	1.003***	1.999***
	(0.270)	(0.534)
Channel B Performance Controls? ($p^B, t^B_{serve_1}, t^B_{serve_2}, t^B_{line_2}$)	Yes	Yes
Demographic Controls?	Yes	Yes
Sample	All subjects	Consistent subjects
Observations	7227	5907
Subjects	219	179

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

unlike algorithm aversion, which we found to be strongly age-dependent (§5), both risk aversion and gatekeeper aversion are robustly present across various demographics.

7. Robustness, Integrative Summary and Structural Estimation

In this section we first present two robustness studies that help rule out alternative explanations. We then revisit our retrospective survey from §2 and discuss an external validity check on our main result. We conclude by re-analyzing our data using structural estimation of eq. (3.1) to evaluate all three constructs (risk aversion, gatekeeper aversion and algorithm aversion) within a common utility framework, and discuss the role of other factors that may influence behavior.

7.1. Robustness

Experiments 1-3 advance our understanding of the customer channel choice between an algorithmic chatbot and a live agent. Our data reject the null hypothesis that channel choices are well explained by expected time minimization. Rather, choices are strongly affected by three separate mechanisms: risk aversion, gatekeeper aversion and algorithm aversion. Among the three, algorithm aversion is the weakest contributor, and is only observed for a subset of the subject population. Before discussing the broader implications of these results, we present two robustness studies. Both robustness studies followed the same general procedures as Experiment 1 (subject pool, exclusions, elicitation method, incentives, sample sizes). We briefly describe the design of the studies and summarize the key findings. The stimuli are described in more detail in EC.4.3. Summary data are presented in Table EC.4.

Robustness Study 1: Making Algorithmic Nature of Server More Salient This study consisted of two experimental treatments. In the treatments, we replicated the *Baseline* and *Context* treatments from Experiment 1, but made the human/algorithmic nature of the server more salient. In particular, in Channel A (channel representing the live agent), participants were required to hold

down a button for the duration of service. In contrast, in Channel B participants continued using keystrokes to interact with the server, exactly like in Experiment 1. See Fig. EC.1 for an illustration. This robustness study allows us to re-examine H1.2 (algorithm aversion), by examining channel choices in a setting where service interactions have a more distinct, channel-specific component. If algorithm aversion continues to play only a small part in driving channel choices in this setting, then we can conclude that the process-related mechanisms (gatekeeper and risk aversion) are the primary driver of behavior.

Robustness Study 2: Adding Expected Time Information The results of Experiments 1-3 suggest risk aversion as an important mechanism, given that Channel B was chosen less frequently than expected time minimization would predict. However, this behavior may be driven not by the inherent riskiness of Channel B, but by subjects' inability or unwillingness to calculate expected times for Channel B.[9] To test this alternative explanation, we conducted two additional experimental treatments. The first one replicated the *Baseline* treatment with expected time information explicitly provided on the decision screen (see Fig. EC.2). The second one also presented subjects with expected time information but used the channel-specific interaction modes from Robustness Study 1. This robustness study allows us to re-examine H1.1, i.e., the role of risk aversion, by directly providing subjects with the expected time information.

Results Figure 8 presents an integrated summary of the results. First, consider panel (a) which summarizes the findings of Experiments 1-3. The x-axis shows the number of Channel A choices in the following treatments: *Deterministic*, *Baseline* and *Live*. Recall that, under expected time minimization, Channel A should be chosen in half of the decisions (i.e., in 5.5 out of 11 decisions per decision set). However, in the *Deterministic* treatment the number of Channel A choices was 6.49, in the *Baseline* treatment it was 7.72 and in the *Live* treatment it was 8.20. These values mark the three segments of the stacked bar in panel (a). By using the expected time minimization benchmark (5.5 Channel A choices) as the null point, we can compute the relative share of each mechanism in explaining the overall deviation from that benchmark. These calculations show that 36.67% of the difference is due to gatekeeper aversion, 45.56% is due to risk aversion, and only 17.78% is due to algorithm aversion.

Next, consider panel (b), in which we plot average Channel A uptake in Robustness Study 1. By making the algorithmic nature of the chatbot more salient in this treatment, we are able to increase the portion of behavior explained by algorithm aversion to 26.73%. However, the remaining two mechanisms continue to dominate behavior, jointly accounting for 73.27%. Finally, consider panel

[9] See, for example, Aimone et al. (2016b) and Aimone et al. (2016a) for evidence of anchoring on specific information like probability or best/worst case outcomes when choosing between risky and safe options.

(c), where we plot average Channel A uptake in Robustness Study 2. Here, the share of algorithm aversion is even smaller (at 13.76%) relative to the previous panels. Further, while the role of risk aversion is reduced, it continues to be present and accounts for 33.86% of behavior. Taken together, these comparisons suggest that algorithm aversion is a secondary hurdle that may reduce adoption in certain settings. In contrast, both gatekeeper aversion and risk aversion are robustly present across multiple decision environments and specifications, and account for much of the variation in the data.

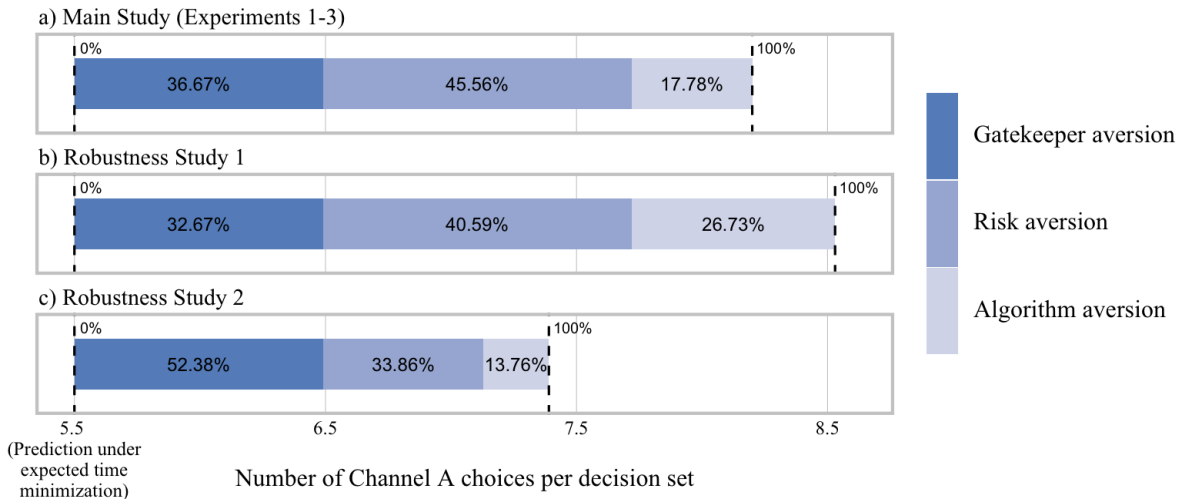
7.2. Additional Survey Question

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

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

The responses indicate that channel choices are more closely related to operational factors (speed and performance) than to the algorithmic/human nature of the server. In particular, a total of 71.78% of respondents invoke operational factors as the key determinant of their choice, with only

Figure 8 Gatekeeper Aversion, Risk Aversion and Algorithm Aversion



28.22% prioritizing the algorithmic/human nature of the server. Thus, the incentivized decisions made in our experiments are broadly consistent with the self-reported attitudes in the survey.

7.3. Structural Estimation

In §4-6 we tested our hypotheses by comparing average Channel B uptake across different treatments. To provide a structural test of our hypotheses and to develop further insights, we pool all four treatments and estimate the model in eq. (3.2)-(3.3) by following the random utility approach where the utility of participant i joining each channel (u_i^A, u_i^B) includes an idiosyncratic component $(\epsilon_i^A, \epsilon_i^B)$. In particular, we have:

$$u_i^A = r - \left(\alpha_{line_1}^A t_{line_1}^A + \alpha_{serve_1}^A t_{serve_1}^A\right)^\gamma + \epsilon_i^A \tag{7.1}$$

$$u_i^B = r - p^B \left(\alpha_{serve_1}^B t_{serve_1}^B\right)^\gamma + (1-p^B)\left(\alpha_{serve_1}^B t_{serve_1}^B + \alpha_{line_2}^B t_{line_2}^B + \alpha_{serve_2}^B t_{serve_2}^B\right)^\gamma + \epsilon_i^B \tag{7.2}$$

Note that the response to different types of servers $(\alpha_{serve_1}^B, \alpha_{serve_2}^B)$ may depend on the experimental implementation of the service process, which differed by treatment. To account for this, we will estimate these parameters separately for each type of the service interaction. We will use $\alpha_{serve}^{neutral}$ to denote the sensitivity to the time spent in the decontextualized service (*Baseline* and *Deterministic* treatments). We will use $\alpha_{human}^{context}$ and α_{human}^{live} to denote the sensitivity to the time spent in service with the animated human server (*Context* treatment) and live human research assistant (*Live* treatment), respectively. Finally, we will use α_{bot} to denote the sensitivity to the time spent with the bot (*Context* and *Live* treatments). By assuming that ϵ_i^A and ϵ_i^B are type-1 extreme-value distributed we can estimate (7.1)-(7.2) using logit probabilities.

We present the estimates in Table 10 (bootstrapped standard errors are available in Appendix EC.5.). The models are ordered from the most restrictive (column (1)) to the fully-specified model (column (5)). In (1) all waiting costs are restricted to be the same and γ is set to 1. Note that the log likelihood (LL) of this model is far lower than the other four, confirming that expected time minimization does not explain channel choices well. In (2) we relax the restriction on γ, which increases it to 2.104. In (3) we again restrict γ to be 1, but allow α_{line_1} to differ from α_{line_2} and set the remaining (in-service) α-parameters equal with each other. In this case we observe that the in-line waiting cost is higher for the second server than for the first one (0.258 vs. 0.222), indicating the presence of gatekeeper aversion. Note that the specification in (3) is somewhat better (in terms of LL and AIC) compared to the one in (2), suggesting that gatekeeper aversion explains a greater portion of variation in the data than risk aversion. In (4) we allow for both risk and gatekeeper aversion, which moderately increases the fit relative to (2) and (3). In (5), we relax all restrictions, which produces a nominally better fit relative to (4). The minimal improvement in fit suggests that, together, gatekeeper and risk aversion account for the bulk of the variation in the data.

The specification in column (5) allows us to perform several statistical tests. (All tests are based on bootstrapped standard errors, reported in EC.5.) First, we find that $\gamma > 1$ at $p < 0.001$, providing strong statistical evidence for risk aversion. Second, $\alpha_{line_2} > \alpha_{line_1}$ at $p = 0.029$, supporting the presence of gatekeeper aversion. Third, when examining algorithm aversion, we do not find evidence for it when the human server is contextualized, as in the *Context* treatment ($\alpha_{bot} > \alpha_{human}^{context}$ at $p = 0.1891$). However, we do find support for algorithm aversion when the human server is live ($\alpha_{bot} > \alpha_{human}^{live}$ at $p = 0.017$). Notably, each of these results robustly aligns with the corresponding reduced-form tests presented in Sections 4-6. We also find that the sensitivity to waiting in line (α_{line_1} and α_{line_2}) is significantly lower than to waiting in service ($\alpha_{serve}^{neutral}$) at $p < 0.001$; however, contextualizing the service process and showing that work is being performed (measured by the α_{human}^{live} parameter) reverses this difference. In particular, both α_{line_1} and α_{line_2} are larger than α_{live}^{human}, though the difference is not statistically significant.[10]

The main insight from the structural estimation exercise is that each of the three behavioral constructs – risk aversion, gatekeeper aversion, and algorithm aversion – explains a nontrivial portion of the variation, with algorithm aversion being the weakest mechanism. Conversely, other behavioral factors, such as the attitudes to waiting in line vs. in service, play only a minimal role. In EC.5, we repeat this analysis with exponential utility; we find that our results hold, but the overall fit is somewhat poorer relative to the power specification presented in the main text.

Table 10 Structural Estimates (Baseline, Context, Live, and Deterministic Treatments)

	(1) **Expected Time Minimization**	(2) **Risk Aversion**	(3) **Gatekeeper Aversion**	(4) **Risk + Gatekeeper Aversion**	(5) **Risk + Gatekeeper + Algorithm Aversion**
α_{line_1}	0.219***	0.054***	0.222***	0.084***	0.101***
α_{line_2}	0.219***	0.054***	0.258***	0.090***	0.109***
$\alpha_{serve}^{neutral}$	0.219***	0.054***	0.325***	0.105***	0.124***
$\alpha_{human}^{context}$	0.219***	0.054***	0.325***	0.105***	0.106***
α_{human}^{live}	0.219***	0.054***	0.325***	0.105***	0.090***
α_{bot}	0.219***	0.054***	0.325***	0.105***	0.116***
γ	1.000***	2.104***	1.000***	1.502***	1.407***
LL	-6124.49	-4880.03	-4832.38	-4745.64	-4719.29
AIC	12250.97	9764.06	9670.77	9499.28	9452.58

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

[10] This is consistent with Buell and Norton (2011), who show that customers respond positively to service experiences that allow them to observe the service being performed, relative to experiences that do not provide such visibility.

8. Concluding Remarks

AI-powered chatbots are becoming an increasingly integral part of online customer service. To successfully leverage chatbot technology, firms need to understand both the relevant customer choice trade-offs as well as their operational implications. In this paper we studied chatbot adoption by soliciting and analyzing testimonies from chatbot users, by using these user stories to formulate a key trade-off in channel choice, and by studying how users navigate this trade-off in incentivized experiments.

Summary and Contributions Together, our results suggest that standard economic analysis of channel choice (based solely on the analysis of waiting times) may oversimplify behaviors. Richer models of customer behavior that incorporate potential deviations from expected time minimization can have nontrivial implications for service design and costs. We identified three such deviations: one based on the inherent uncertainty of the time spent in the system ("risk aversion"); a second one based on the difference between a single service stage system and a multi-stage service system with a frequently failing gatekeeper and a more competent expert ("gatekeeper aversion"); and a third one, driven by customer attitudes regarding algorithms ("algorithm aversion"). Between the three, we found algorithm aversion to be the weakest driver of behavior, accounting for 14% to 27% (depending on specification) of chatbot underutilization.

Our contributions are threefold. First, we contribute to the algorithm aversion literature, which focuses primarily on algorithmic attitudes in human-AI collaboration. In contrast, we study algorithm aversion in a service context. We identify contingencies when algorithm aversion is present and measure its relative importance. Second, our methodological contribution is to show that experiments that rely on framing – the way information is presented to influence decision-making – used in much of the algorithm aversion literature (see Glikson and Woolley 2020, for a review) may prompt weaker levels of algorithm aversion than experiments that vary the human/robotic nature of the algorithm in a more natural, realistic manner (see, for example, Bastani et al. 2021). Third, we contribute to the behavioral queueing literature. We use a novel framework for examining queue joining behaviors in a complex gatekeeper service system, where service providers differ not only in duration but also in the sequence of stages and in the content of the wait. This framework can serve as a template for studying behaviors in service settings beyond just customer service, for example in e-commerce or healthcare.

Service Design Implications Our findings offer new practical insights for managers. The positive response to chatbot performance parameters suggests that managers should continue investing in chatbot performance. Such investments can target improvements in chatbot reliability and range, or reductions in the time needed to correctly diagnose and respond to customer requests. At the

same time, our data suggest that the returns to improvements in performance may be marginally decreasing in certain ranges – a nontrivial share of decision-makers in our data avoided the chatbot channel regardless of its performance. Second, our results suggest a simple operational lever that can increase chatbot adoption: prioritizing chatbot customers for quicker access to a live agent when the chatbot fails to resolve their issue. Because waiting in line after chatbot failure is especially painful, reducing these waits will prompt more customers to opt for the chatbot channel. The shift towards the chatbot channel can reduce the demand for live agents and free up staffing capacity, which can then be used to cut staffing or to improve service quality in the live agent channel. Third, our results on algorithm aversion suggest that managers need to be cognizant of age as a significant moderator of chatbot uptake and may need to customize the available channels to the type of product and target demographics.

Extensions and Outlook To keep the model and experiments focused on the key trade-offs, 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 industry and firm-specific vs. general-purpose chatbots, the role of privacy concerns and the use of customer data in these service interactions, the role of equity, agency and the right to access a human provider, as well as richer interaction environments, such as voice, video or virtual reality. As chatbot technology continues 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.

Afèche P, Baron O, Kerner Y (2013) Pricing time-sensitive services based on realized performance. *Manufacturing & Service Operations Management* 15(3):492–506.

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

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

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

Ata B, Olsen TL (2009) Near-optimal dynamic lead-time quotation and scheduling under convex-concave customer delay costs. *Operations Research* 57(3):753–768.

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

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

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

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

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

Buell RW, 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* ucad023.

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

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

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

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

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

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

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

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

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

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

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

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

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

Glikson E, Woolley AW (2020) Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals* 14(2):627–660.

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).

Hasija S, Pinker EJ, Shumsky RA (2005) Staffing and routing in a two-tier call centre. *International Journal of Operational Research* 1(1-2):8–29.

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

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

Janakiraman N, Meyer RJ, Hoch SJ (2011) The psychology of decisions to abandon waits for service. *Journal of Marketing Research* 48(6):970–984.

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

Kagel JH, Roth AE (2020) *The handbook of experimental economics, volume 2* (Princeton university press).

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

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

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

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

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

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

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.

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

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

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

Maynard N, Crabtree G (2020) 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.

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

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

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

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

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

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

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

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

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

Veeraraghavan S, Debo L (2009) Joining longer queues: Information externalities in queue choice. *Manufacturing & Service Operations Management* 11(4):543–562.

Electronic Companion

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

Below we provide details and data for the first wave of the survey described in §2 of the manuscript.

EC.1.1. 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-Q13). Participants were compensated with a show up fee of $2. They also received an additional payment of $2 at the end of the study (i.e., a total of $4 for completing the entire study).

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

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

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

Q2. *Please take a few minutes to describe, in as much detail as you can remember, a recent time when you had to contact customer support. Specifically, we are interested in a situation*

when you had to interact with a live (human) customer support agent [chatbot], either via phone or chat. Aiming for 1-2 sentences, please answer the following questions. What caused you to contact customer support? What type of service/issue did you need help with? What drove your decision to speak to a live agent [chatbot] (as opposed to, for example, looking at the FAQ)? How did the agent try to resolve your issue? How did your experience compare to your expectations? How did you feel about the decision to use live customer support?

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

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

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

Q5. *Of the two interaction types below* [Note: 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.2. Supporting Analysis for Survey (Wave 2) (§2)

Below we provide details and data for Wave 2 of the survey, described in §2 of the manuscript.

EC.2.1. Survey Setup and Questions

This wave of the survey was conducted in July 2023 on the Prolific platform. A total of 202 respondents were recruited from the US-based population (49% female, average age: 38). All respondents received a show up payment of $3.00 and an additional payment of $2.00 at the end of the study.[11] The survey included a replication of most of the questions asked in Survey 1, with the addition of several new questions.

EC.2.2. Replication of Questions from Survey 1

Below we reproduce the portion of Survey 2 which involved a replication of the Survey 1 questions. Note that participants saw different questions depending on the treatment (live agent vs chatbot). Summary statistics of responses for multiple choice questions are provided after each questions.

Q1. *In the past 12 months, have you interacted with a customer support agent [chatbot]?*
Live Agent: No (7.84%) Yes (90.20%) Not Sure (1.96%).
Chatbot: No (5.00%) Yes (93.00%) Not Sure (2.00%).

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

Q2. *Please take a few minutes to describe, in as much detail as you can remember, a recent time when you had to contact customer support. Specifically, we are interested in a situation when you had to interact with a live (human) customer support agent [chatbot]. Aiming for 1-2 sentences, please answer the following questions.*

- *What caused you to contact customer support? What caused you to contact customer support? What type of service/issue did you need help with?*
- *What drove your decision to speak to a **live agent [chatbot]** as opposed to, for example, using a chatbot [live customer support], or looking at the FAQ?*
- *How did the agent try to resolve your issue?*
- *How did your experience compare to your expectations? How did you feel about the decision to use live customer support [a chatbot]?*

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

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

Q3. *Were you given a choice between different options for customer support (e.g., a chatbot vs a live customer support agent)?*
Live Agent: No (49.91%) Yes (51.09%).
Chatbot: No (61.29%) Yes (38.71%).

Q4. *Were you given a time estimate for how much time it might take until you can use different support formats (e.g., waiting time for a chatbot vs waiting time for a live customer support* agent)? [Question dropped in Survey 2]

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

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

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

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

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

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

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

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

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

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

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

Live Agent: 3.27 on average.

Chatbot: 2.27 on average.

EC.2.3. Additional Questions

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

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

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

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

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

EC.3. Supporting Analysis for Experiments

In this section we present supporting analyses for §4-§7. Table EC.1 presents extended analysis for §4. This table is based on the same regression equation as Table 5, but includes additional specifications without demographic controls (columns 1 and 3), as well as the estimates for the channel parameters (p^{succ}, t^B_{serve} and $t^B_{line_2}$). Analogously, Table EC.2 and Table EC.3 presents extended analysis for §5 and §6. Finally, Table EC.4 presents summary data for the robustness studies in §7.1.

Table EC.1 Channel Preferences in Experiment 1

	(1)	(2)	(3)	(4)
Dependent Variable:	Channel B	Channel B	Channel B	Channel B
Baseline	-	-	-	-
Context	0.040	-0.082	0.003	-0.267
	(0.316)	(0.306)	(0.585)	(0.561)
p^B	12.600***	12.600***	29.380***	29.390***
	(0.545)	(0.545)	(1.303)	(1.302)
$t^B_{serve_1}$	-0.363***	-0.363***	-0.574***	-0.574***
	(0.015)	(0.015)	(0.026)	(0.026)
$t^B_{line_2}$	-0.195***	-0.195***	-0.331***	-0.331***
	(0.008)	(0.008)	(0.015)	(0.015)
Female		-0.094		-0.160
		(0.321)		(0.576)
Age		-0.013		-0.006
		(0.013)		(0.023)
Quiz errors		-0.509***		-1.098***
		(0.171)		(0.319)
Risk aversion		0.188**		0.309**
		(0.074)		(0.137)
Intercept	3.255***	3.418***	0.113	-0.028
	(0.495)	(0.767)	(0.884)	(1.394)
Observations	6567	6567	5478	5478
Subjects	199	199	166	166

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

Table EC.2 Channel Preferences in Experiment 2

	(1) Channel B	(2) Channel B	(3) Channel B	(4) Channel B	(5) Channel B	(6) Channel B
Baseline	-	-	-	-	-	-
Context	0.0450	-0.0812	-0.0571	0.0139	-0.243	-0.174
	(0.355)	(0.341)	(0.338)	(0.633)	(0.607)	(0.599)
Live	-0.551	-0.611*	-0.606*	-1.020	-1.067*	-1.033*
	(0.355)	(0.342)	(0.339)	(0.630)	(0.602)	(0.593)
p^B	12.50***	12.50***	12.50***	27.96***	27.96***	27.97***
	(0.451)	(0.450)	(0.450)	(1.029)	(1.028)	(1.028)
$t^B_{serve_1}$	-0.368***	-0.368***	-0.368***	-0.555***	-0.555***	-0.555***
	(0.0126)	(0.0126)	(0.0126)	(0.0207)	(0.0207)	(0.0207)
$t^B_{line_2}$	-0.194***	-0.194***	-0.194***	-0.314***	-0.314***	-0.314***
	(0.00672)	(0.00672)	(0.00672)	(0.0117)	(0.0117)	(0.0117)
Female		-0.228	-0.158		-0.365	-0.253
		(0.287)	(0.286)		(0.500)	(0.494)
Age		-0.0328***	0.00446		-0.0443**	0.0355
		(0.0114)	(0.0214)		(0.0200)	(0.0405)
Quiz errors		-0.383**	-0.355**		-0.698**	-0.627**
		(0.157)	(0.157)		(0.280)	(0.280)
Risk aversion		0.220***	0.216***		0.362***	0.356***
		(0.0673)	(0.0670)		(0.118)	(0.118)
Age × *Context*			-0.0297			-0.0654
			(0.0282)			(0.0516)
Age × *Live*			-0.0761***			-0.153***
			(0.0290)			(0.0522)
Intercept	3.247***	2.868***	2.810***	0.0905	-0.523	-0.643
	(0.440)	(0.544)	(0.542)	(0.763)	(0.936)	(0.932)
Sample	All subjects	All subjects	All subjects	Consistent subjects	Consistent subjects	Consistent subjects
Observations	10032	10032	10032	8415	8415	8415
Subjects	304	304	304	255	255	255
Marginal effect of *Live* at...			***p*−value:**			***p*−value:**
Age= 20			0.334			0.204
Age= 25			0.736			0.541
Age= 30			0.532			0.672
Age= 35			0.071			0.078
Age= 40			0.007			0.005
Age= 45			0.002			0.001
Age= 50			0.002			0.001
Age= 55			0.002			0.001
Age= 60			0.002			0.001

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

Table EC.3 Channel Preferences in Experiment 3

	(1)	(2)	(3)	(4)
Dependent Variable:	Channel B	Channel B	Channel B	Channel B
Baseline	-	-	-	-
Deterministic	0.992***	1.002***	2.035***	1.998***
	(0.272)	(0.270)	(0.541)	(0.534)
Expected time in Channel B	-0.344***	-0.344***	-0.685***	-0.685***
	(0.009)	(0.009)	(0.022)	(0.022)
Female		-0.370		-0.677
		(0.276)		(0.553)
Age		-0.011		0.003
		(0.012)		(0.025)
Quiz errors		-0.196		-0.376
		(0.155)		(0.313)
Risk aversion		0.103		0.167
		(0.064)		(0.128)
Intercept	11.880***	12.240***	23.930***	23.820***
	(0.405)	(0.666)	(0.912)	(1.434)
Observations	7227	7227	5907	5907
Subjects	219	219	179	179

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

Table EC.4 Data Summary for Robustness Studies

			Participant Type			
	No. of Channel A choices per decision set	No. of Channel B choices per decision set	(1) Never chooses Channel B	(2) Underutilizes Channel B	(3) Expected time minimizer	(4) Other
Experiment 1						
Baseline	7.72	3.28	17.28%	45.68%	18.52%	18.52%
Context	7.76	3.24	14.23%	52.94%	8.24%	24.71%
Robustness Study 1						
Rob1	7.99	3.01	17.95%	52.56%	7.69%	24.71%
Rob1+Context	8.53	2.47	29.63%	39.51%	11.11%	19.75%
Robustness Study 2						
Rob2	7.13	3.87	16.47%	28.24%	36.47%	18.82%
Rob2+Context	7.39	3.61	6.12%	31.63%	45.92%	16.33%

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

EC.4. Experimental Stimuli

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

The data were collected on the Prolific platform. We only used workers based in the United States (to avoid any country-specific effects) with an approval rating of at least 98%. The experiments were conducted on weekdays between 9am and 6pm Eastern Time. Participants were randomly assigned to a treatment upon signing up for the experiment. Instructions for the *Context* treatment are reproduced below. Instructions for the *Baseline* treatment are analagous, with the difference that the Live Agent Channel is referred to as "Channel A" and the Chatbot channel as "Channel B". Also see screenshots in Figure 3 for an illustration of the contextual cues.

Instructions

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

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

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

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

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

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

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

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

What will happen in the Chatbot Format? Unlike in the Live Agent Format, in the Chatbot Format you do not know the exact amount of time you will spend until you receive the reward. In this format you will first spend 20 with the chatbot. However, the chatbot is not always capable of resolving your request. If the chatbot fails to resolve your request you will need to interact with a live agent. Thus, different from Live Agent Format, in Format 2 there may be multiple service stages before service is completed.

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

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

Click "Next" to experience the Chatbot Format.

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

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

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

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

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

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

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

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

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

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

EC.4.2. Experiment 2: Script and Training

As in Experiment 1, we only used workers based in the United States (to avoid any country-specific effects) with an approval rating of at least 98%. The experiments were conducted in August 2023, on weekdays between 9am and 6pm Eastern Time.

Channel B was unchanged relative to Experiment 1 (*Context* 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 prior to making any choices to test the interface.

Participant: *[starts conversation]*

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

Participant: *[responds]*

Experimenter: "What is your issue type?"

Participant: *[enters issue type]*

Experimenter: "Got it. Advancing you."

[*Experimenter starts service process. Participant sees the progress bar fill which takes* $t^A_{serve_1}$ *seconds. Participant is able to move on to next page after* $t^A_{serve_1}$ *seconds*].

EC.4.3. Robustness Studies: Details and Instructions

Here, we provide supporting information for Robustness Study 1 and Robustness Study 2 introduced in §7.1. As in Experiments 1-3, we only used workers based in the United States with an approval rating of at least 98%. The experiments were conducted on weekdays between 9am and 6pm Eastern Time. Participants were randomly assigned to a treatment upon signing up for the experiment. We conducted a total of four between-subject treatments. A total of 381 participants completed the robustness studies (after exclusions based comprehension checks), with an approximately even split over the following four treatments.

Rob1 Treatment This treatment was identical to the *Baseline* treatment, with the difference that service interactions involved pressing down a button instead of using keystrokes in Channel A. Channel B was unchanged relative to the *Baseline* treatment. See Figure EC.1a)-b) for an illustration of the differences.

Rob1+Context Treatment This treatment was identical to the *Context* treatment, with the difference that service interactions involved pressing down a button instead of using keystrokes. Channel B was unchanged relative to the *Context* treatment. See Figure EC.1c)-d) for an illustration of the differences.

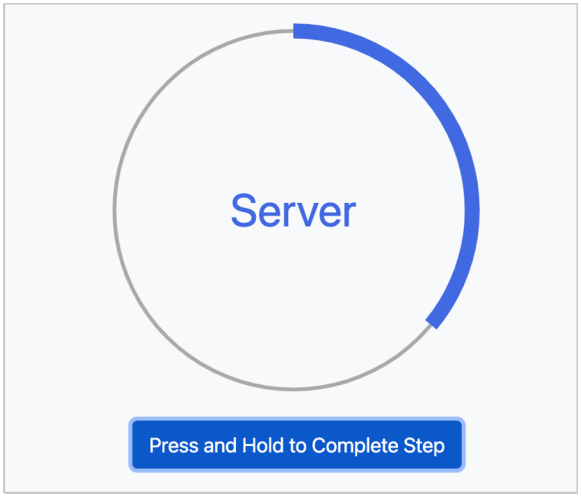
Rob2 Treatment This treatment was identical to the *Baseline* treatment, with the difference that participants saw expected waiting times for each alternative. See Figure EC.2 for an illustration of the differences.

Rob2+Context Treatment This treatment was identical to *Rob2*, with the difference that participants saw expected waiting times for each alternative, and the interactions involved holding down a button, as in the *Rob1 + Context Treatment*.

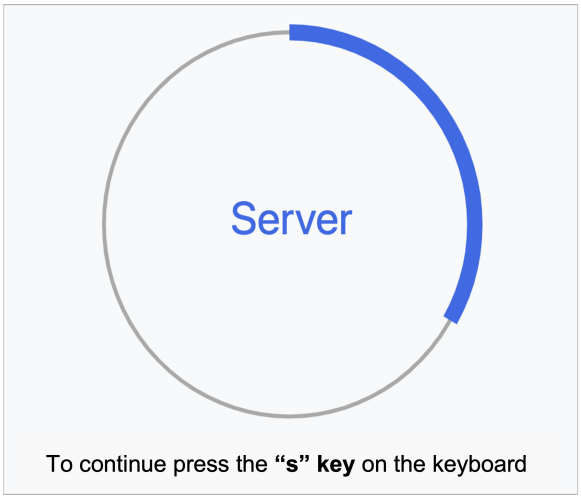
Instructions for *Rob1* and *Rob2* treatment were unchanged relative to the *Baseline* treatment. Instructions for *Rob1 + Context* and *Rob2 + Context* treatment were unchanged relative to the *Context* treatment. The screenshots of the key experimental manipulation in treatments *Rob1* and *Rob1 + Context* are reproduced in Figure EC.1. Screenshot of decision screen in *Rob2* and *Rob2 + Context*, as well as the original choice screens used in Experiments 1-3 are reproduced in Figure EC.2.

Figure EC.1 Waiting Experiences (Screenshots)

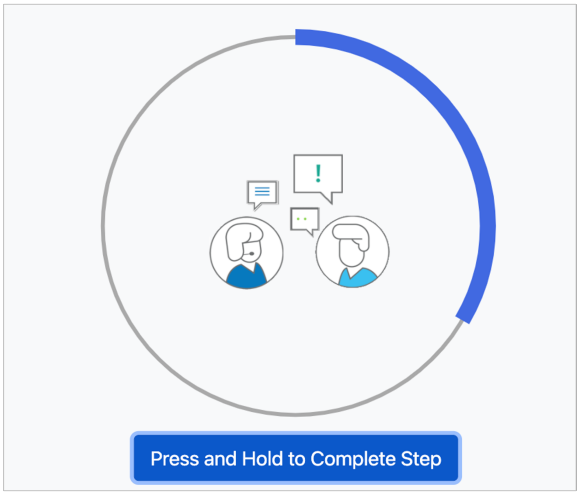
(a) *Rob1* Treat., Interaction w. "Live Agent"



(b) *Rob1* Treat., Interaction w. "Chatbot"



(c) *Rob1+Context* Treat., Interaction w. "Live Agent"



(d) *Rob1+Context* Treat., Interaction w. "Chatbot"

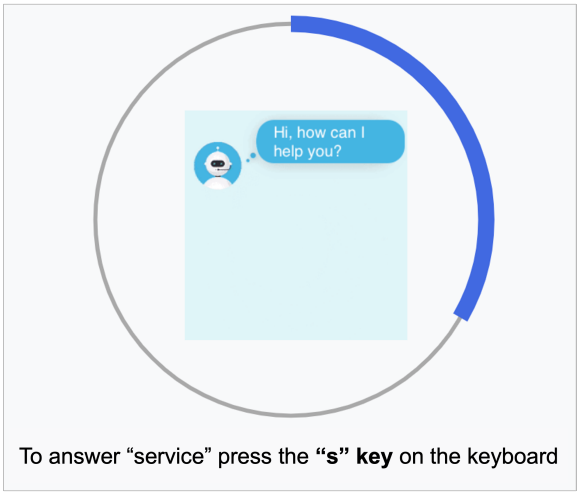


Figure EC.2 Choice Screens

(a) Example Choice Screen in Experiments 1-3

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

(b) Example Choice Screen in *Rob2* and *Rob2+Context* Treatments

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)**

Table EC.5 Structural Estimates (Baseline, Context, Live, and Deterministic Treatments)

	(1) Expected Time Minimization	(2) Risk Aversion	(3) Gatekeeper Aversion	(4) Risk + Gatekeeper Aversion	(5) Risk + Gatekeeper + Algorithm Aversion
α_{line_1}	0.219***	0.054***	0.222***	0.084***	0.101***
	(0.009)	(0.004)	(0.012)	(0.012)	(0.019)
α_{line_2}	0.219***	0.054***	0.258***	0.090***	0.109***
	(0.009)	(0.004)	(0.014)	(0.015)	(0.023)
$\alpha_{serve}^{neutral}$	0.219***	0.054***	0.325***	0.105***	0.124***
	(0.009)	(0.004)	(0.014)	(0.016)	(0.024)
$\alpha_{human}^{context}$	0.219***	0.054***	0.325***	0.105***	0.106***
	(0.009)	(0.004)	(0.014)	(0.016)	(0.020)
α_{human}^{live}	0.219***	0.054***	0.325***	0.105***	0.090***
	(0.009)	(0.004)	(0.014)	(0.016)	(0.018)
α_{bot}	0.219***	0.054***	0.325***	0.105***	0.116***
	(0.009)	(0.004)	(0.014)	(0.016)	(0.022)
γ	1.000***	2.104***	1.000***	1.502***	1.407***
	(0.000)	(0.097)	(0.000)	(0.102)	(0.118)
LL	-6124.49	-4880.03	-4832.38	-4745.64	-4719.29
AIC	12250.97	9764.06	9670.77	9499.28	9452.58

Estimates obtained using Maximum Likelihood Estimation. $^{***}p < 0.01$

EC.5. Structural Estimation

In Table EC.5 we provide the structural estimates from §7.3 with bootstrapped standard errors. In Table EC.6 we provide the estimates of the structural model under exponential utility, where

$$u_i^A = \exp(r - \left(\gamma + \alpha_{line_1}^A t_{line_1}^A + \alpha_{serve_1}^A t_{serve_1}^A\right)) + \epsilon_i^A$$
$$u_i^B = \exp(r - p^B \left(\gamma + \alpha_{serve_1}^B t_{serve_1}^B\right)) + (1 - p^B) \left(\gamma + \alpha_{serve_1}^B t_{serve_1}^B + \alpha_{line_2}^B t_{line_2}^B + \alpha_{serve_2}^B t_{serve_2}^B\right)) + \epsilon_i^B$$

Note that the log likelihood of the fully-specified model (5) is lower than that of the power utility model, but the results hold qualitatively.

Table EC.6 Structural Estimates (Baseline, Context, Live, and Deterministic Treatments)

	(1) Fully-Restricted	(2) Risk Aversion	(3) Gatekeeper Aversion	(4) Risk + Gatekeeper Aversion	(5) Risk + Gatekeeper + Algorithm Aversion
α_{line_1}	0.040***	0.028***	0.039***	0.010***	0.009***
α_{line_2}	0.040***	0.028***	0.037***	0.011***	0.010***
$\alpha_{serve}^{neutral}$	0.040***	0.028***	0.041***	0.013***	0.011***
$\alpha_{human}^{context}$	0.040***	0.028***	0.041***	0.013***	0.009***
α_{human}^{live}	0.040***	0.028***	0.041***	0.013***	0.007***
α_{bot}	0.040***	0.028***	0.041***	0.013***	0.010***
γ	0.000***	1.051***	0.000***	2.630***	2.972***
LL	-5009.30	-4901.26	-5002.68	-4756.38	-4725.26
AIC	10020.60	9806.52	10011.36	9520.69	9462.53

Estimates obtained using Maximum Likelihood Estimation. $^{***}p < 0.01$