Routing analyses for call centers with human and automated services

Benjamin Legros

Routing analyses for call centers with human and automated services

Benjamin Legros

Ecole de Management de Normandie, Laboratoire Métis, 64 Rue du Ranelagh, 75016 Paris, France

benjamin.legros@centraliens.net

Abstract

We analyze a call center with robot and human agents. Customers arrive over time and have a preference for being served by an agent. Thus, although robots are infinite in number, it might be preferable to let some customers wait so as to give them a human service. The aim of this paper is to determine how agents and robots should be placed in the call center architecture to deliver the lowest wait- and service-dissatisfaction. We prove that in terms of expected wait- and expected service-dissatisfaction, a preventive policy where scheduling to robots or to agents is made at arrival, outperforms corrective policies, where scheduling is made after a certain wait, or policies where the robot service can be interrupted. This conclusion is only true with respect to expectation, however. When considering higher moments of the wait, a corrective policy appears to be preferable. Service interruption leads to less variability in service-dissatisfaction, which can be appreciated when customer fairness is sought.

Keywords: Queueing; call center; robot agents; routing; scheduling

1 Introduction

The decreasing costs and improved capabilities of advanced technologies such as robotics make manufacturing and services more attractive in the US and in Europe than in economies whose major advantage is cheap labor. This evolution has led to a trend in reshoring activities. A report by BCG (2015) indicates that 56% of respondents perceived lower automation costs as a driver of U.S.-made products' competitiveness, and 75% said they would invest in additional automation or advanced technologies within the next five years. The same trend is observed in Europe (Storrie, 2019). An advantage of robots is that they are available 24/7 and do not need to be paid per hour of work. As robots become cheaper, an increasing amount of work will become automated, which offsets the advantage of cheap labor environments. However, this conclusion is only partially true as flexibility plays a big role in this trade-off. Manual labor remains more flexible and hence can respond more effectively to changes in demand.

The lack of flexibility in robot services can be particularly problematic in service systems like call centers. Already, when comparing real and virtual interaction between customers and agents, Bennington et al. (2000) showed that customers have slightly higher satisfaction levels with in-person services than with call center services. The most problematic aspects of virtual conversations are the lack of personalized/individualized service, uncaring communication, being given the "run around", and unreliable information and service. Callers predominantly report frustration as the emotion arising from a negative experience in a call center encounter (Hudson et al., 2017). These negative aspects would inevitably be worse in a conversation between a customer and a robot.

Although robots have not yet been widely deployed in call centers, a high number of patents indicates that robot services could become the future of call centers (Seeley et al., 2010; Lynch et al., 2016; Palakovich et al., 2017). The article by Dialer360 (2020) set out the positive and negative aspects of robot agents as compared to human ones. As would be expected, the advantages of robots are that they are trained instantly, they are cheaper as they do not need to be paid every month, they can serve as many customers as needed at the same time thereby eliminating waits, and the work can be done around the clock. The main negative aspect in the interaction is the lack of specific thinking. Humans have a creative mind which can easily invent and discover new things and utilize them to the benefit of a business, unlike a robot which cannot think for itself, and can only work according to the program and instructions it has been given. Recent studies have confirmed that the average consumer would rather interact with a live agent than a robot. CustomerServ (2018) noted that, if offered a choice, 83% of consumers said they would prefer to speak to a real person since human agents better understand their needs (78%) and can address multiple questions at once (57%).

Therefore, while the benefits of robot services are appealing, we are not yet at the stage of widespread

adoption in practice. This is confirmed by the lack of publications related to automated services in the academic call center literature. We believe that the future of call centers will not see a radical switch from human to robot services, but a combination of the two resources. The aim of this paper is to explore how agents and robots should be placed in the call center architecture to provide optimum management of customers' wait and service-dissatisfaction.

In terms of wait-dissatisfaction, we compare two architectures. In the first one, the routing decision to the agents' queue or to robots is made at customers' arrival based on the expected wait. This first architecture is viewed as a preventive policy where agents compete with robots. In the second one, all customers wait in the same queue and routing to robots is decided only if the wait exceeds a certain threshold. This second architecture is viewed as a corrective strategy where robots support agents. For both policies, we assume that the same proportion of customers is served by robots, and we compare the wait-dissatisfaction. We prove that the average speed of answer (i.e., the expected wait) is shorter with a preventive policy. This result is important as it shows that it is advisable to implement the first architecture when the global expected wait is the objective, as in many service systems. However, when considering higher moments of the wait, or wait-time percentiles, the corrective policy often shows better performance. Moreover, when only focusing on customers served by human agents, the second architecture outperforms the first one in all wait-related metrics. Therefore the corrective policy is advisable when the system manager cares more about customers served with humans than those served by robots. This may be the case when humans agents are involved in sales. In addition the reduction of the wait variability with the corrective policy can be appreciated as it reduces the occurrence of excessively long waits.

Next, we focus on service-dissatisfaction. While no empirical studies to date have indicated how service-dissatisfaction should be measured, we assume that dissatisfaction increases with the length of service, the number of questions asked, the number of retries, and the probability to abandon the service. We thus built a measure of service-dissatisfaction that accounts for these different metrics. First, we considered a model without human agents and show how service interruption can be efficiently implemented to reduce inefficient interactions between customers and robots. Next, we examined whether agents should interrupt a robot service or should be positioned as an alternative to the robot service. When considering expected service-dissatisfaction, the second strategy is found to be the best. It also allows the system to have a higher proportion of well-served customers, although improvement is limited for this metric. The main advantage of interrupting a service is to achieve lower variability in service-dissatisfaction, so that the system offers customers a fairer service among customers.

Section 2 presents the literature review. Section 3 compares preventive and corrective policies to manage

wait-dissatisfaction. Section 4 evaluates the service interruption strategy and Section 5 concludes the paper and provides avenues for future research. The proofs of the main results and a table of notations are provided in the appendix.

2 Literature review

First, since our study considers routing issue in a queueing context, we examine prior studies in the related field of admission control. Second, in relation to the call center application considered in this study, we present the existing literature on the implementation of routing strategies. Finally, as our analysis examines service interruption, we detail the literature on service interruptions due to server vacations or service failure.

Comparison between preventive and corrective policies can be understood as a routing issue in a queueing context. In particular, employing robots when the human queue is congested relates our study to an analysis of overflow policies whereby, for a given optimization problem where a trade-off between system congestion and rejection flow has to be established, a controller may choose whether or not to reject a customer from the system. This issue has been extensively addressed in various ways by many authors (Ku and Jordan, 2003; Maglaras and Van Mieghem, 2005; Ward and Kumar, 2008; Xu, 2015; Niyirora and Zhuang, 2017; Bountali and Economou, 2017). One main characteristic of the optimal admission policies is that these are of the threshold type (Koole, 2007). This result allows us to define the preventive policy in Section 3 in our paper. Extensions of the result can be found with general interarrival times for a loss system (Örmeci and van der Wal, 2006), batch arrivals (Örmeci and Burnetas, 2005; Yildirim and Hasenbein, 2010), and state-dependent service rates (Xia et al., 2017).

The control of customers' admission is often used in service systems like hospitals or call centers. In a queueing network of service facilities, Cosyn and Sigman (2004) analyzed the admission control question with waiting and reneging from a revenue perspective. Using orbiting as an approximation of queueing, they proved that a particular tracking policy may be close to optimal. Lin and Ross (2004) studied a loss queueing system with a single server where a gatekeeper decides whether to admit a customer without knowing the server's status (idle or busy). They found that a threshold policy which rejects arrivals for a certain time interval after each admission and next accepts the next customer is optimal. Bassamboo et al. (2005) considered a multi-class service system with many teams of agents and doubly stochastic arrivals. A twofold control system was used, with rejection control at arrival and routing control to a given team after a given wait. In the context of outsourcing, Gans and Zhou (2007) studied a call center with two types of calls to investigate routing mechanisms for outsourcing part of the low value call. Gurvich and Perry (2012) considered a service network operated under a threshold-type overflow scheme. If the waiting room is full, the call is overflowed to an

outsourcer. In another outsourcing context, Legros et al. (2020) showed that rejecting customers after letting them experiment some wait leads to an improvement in generated revenue for the call center as compared to rejection on arrival. The study in this paper provides another application of admission control where both human and automated services are implemented. It shows that the threshold admission policy -found to be optimal for most applications- is not necessarily optimal when considering certain functions of the wait-dissatisfaction.

Section 4 of the paper examines service interruption. Compared to our paper where service interruption is decided by the system manager, this field of the literature considers exogenous interruptions due to server vacations or service failure. Jain et al. (2019) recently conducted a survey to obtain an overview of this literature stream. Mitrany and Avi-Itzhak (1968) considered an M/M/N queue where each server is subject to random breakdowns of exponentially distributed duration and derived the moment generating function of the queue size. Baba (1986) investigated an $M^X/G/1$ queue with server vacation and derived the queue length distribution at an arbitrary time. Moreover, a transient analysis with service interruption can be found in Lee (1997). Choudhury and Deka (2008) provided an extensive analysis of the steady state behavior of a retrial queue with an additional second phase of service subject to breakdowns occurring randomly at any instant while serving the customers. This model generalizes both the classical retrial queue subject to random breakdown as well as queue with second optional service and server breakdowns. Also employing a retrial queue, Gao and Wang (2014) investigated the case with non-persistent customers. They presented the necessary and sufficient condition for the system to be stable and the joint queue length distribution in steady state. Multi-server queues have been investigated less frequently. We mention Dudin et al. (2015) who analyzed a queueing model with Markovian Arrival process and Markovian interruptions. One major difference between this field of literature and our study is that service interruption is an active decision of the system which can possibly be optimized in our study whereas service interruptions are analyzed as uncontrolled events in the aforementioned literature.

3 Robots as a support or an alternative to agents

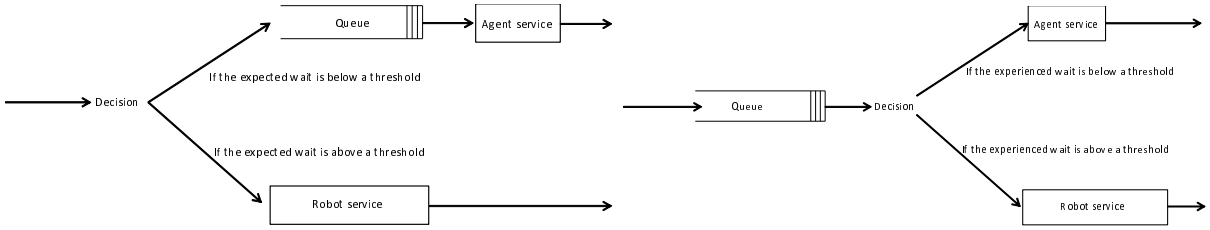
Here, we investigate how agents should be placed in the queueing architecture to provide the best management of customers' *wait-dissatisfaction*. First, in Section 3.1, we derive the performance measures under a preventive and a corrective policy for the wait management. We also obtain a theoretical comparison between the two policies for the first and second moments of the wait and for waiting time percentiles. Next, in Section 3.2 we provide a numerical comparison for higher moments of the wait and numerical experiments to explain the conditions under which one architecture should be selected.

3.1 Evaluation of the preventive and corrective architectures

We consider a situation with an infinite number of robots and a finite number of s agents. The service can be conducted by either agents or robots. Service time by an agent is assumed to be exponentially distributed with rate μ and customers' arrival process is Poisson with rate λ. The system manager can decide whether a customer should be served by an agent or by a robot. We assume that agents are preferred by customers rather than robots. The opposite case may occur, but since we have an infinite number of parallel robots for service, potentially leading to no wait in the system, it would not make sense to hire an agent. Thus, studying the position of agents in the system is not especially useful. Furthermore, we assume that agents are insufficient in number to provide a good service quality to all arriving customers, leading the system manager to employ robots for the service of some customers. This assumption is also made to avoid the trivial case where robots are not needed. Finally, as we aim to investigate the competition between human and robot agents, we assume that there is at least one human agent being present such that the routing issue to either human or robot agents can be investigated.

Due to the potential unavailability of agents, a customer may have to wait in a first-come-first-served queue before starting service. The challenge for the system manager is to determine whether a service with lower quality ensured by a robot is preferable to waiting for a service with higher quality. As dissatisfaction increases with wait time, robots should be employed when the congestion is too high. We consider two architectures for managing customers' flow to agents and robots. In the first one, customers are routed in priority to agents but if the expected wait at arrival is too high, then an arriving customer is directly routed to a robot. In the second one, all customers are routed to the queue with the idea of being served by an agent. Robots are employed only for customers who have waited too long in the queue. In the first architecture, robots are viewed as an alternative to agents' service whereas in the second one robots support the agents' team as a complementary resource. For the system manager, the comparison between the two architectures can be understood as managing an overflow policy where either a preventive (first architecture) or a corrective (second architecture) policy should be employed. In both cases, we employ threshold policies for decision making as depicted in Figure 1. This policy type is shown to be optimal for various settings and problems with customers' rejection at arrival (Cosyn and Sigman, 2004; Bassamboo et al., 2005; Gans and Zhou, 2007; Gurvich and Perry, 2012) or customers' rejection after experimenting some wait (Koole et al., 2012; Legros et al., 2020; Legros, 2020).

We introduce the control parameters n and w, such that n is the maximum number of customers who can be present in the queue in the first architecture and w is the maximum authorized wait in the second one. Note that due to Little's law, this is equivalent to controlling the expected wait in the first architecture. Parameters



(a) Robots as an alternative to agents (preventive policy) (b) Robots as a support to agents (corrective policy)

Figure 1: The two architectures

n and w are adjusted such that the proportion of customers served by robots (and by agents) is identical in the two systems. The question is *which of the two systems provides the lowest wait-dissatisfaction.* To answer this question, we need to specify how wait-dissatisfaction should be measured. Koole (2003) provides a discussion on the subject. Expected wait (i.e., average speed of answer) or a percentile of the wait are considered as the two most common metrics used to evaluate a system's performance as they are simple for managers to understand. However, they may not reflect the reality of wait-dissatisfaction. The author concluded that average excess (i.e., expected wait above a threshold) could be a good metric to capture customers' behavior as the latter appear to care about the wait only after a certain threshold. Kumar and Krishnamurthy (2008)'s empirical study showed that customers' perception of the wait is also impacted by the variability of the service times. Subsequently, the second moment of the wait should also be considered when evaluating wait-dissatisfaction. These metrics will be included in the discussion.

First, we recall the performance measures in the two systems. The performance measures of the first architecture can be obtained from the Markov chain analysis of an M/M/s/s+n queue, while the performance measures of the second one are those of an M/M/s+D queue and can be deduced from the results in Baccelli and Hebuterne (1981). We denote by a, the ratio λ/μ. The probability of being served by agents, P_S, and the probability-density function of the waiting time, $f_W(t)$, are given by

$$\text{Architecture 1: } P_S = \frac{\epsilon + \frac{a}{s}\frac{1-\left(\frac{a}{s}\right)^n}{1-\frac{a}{s}}}{\epsilon + \frac{a}{s}\frac{1-\left(\frac{a}{s}\right)^{n+1}}{1-\frac{a}{s}}}, \text{ and } f_W(t) = \lambda \frac{e^{-s\mu t}\sum_{x=0}^{n-1}\frac{(\lambda t)^x}{x!}}{\epsilon + \frac{a}{s}\frac{1-\left(\frac{a}{s}\right)^{n+1}}{1-\frac{a}{s}}}\mathbb{1}_{t>0} + \delta_{t=0}\frac{\left(\frac{a}{s}\right)^{n+1}}{\epsilon + \frac{a}{s}\frac{1-\left(\frac{a}{s}\right)^{n+1}}{1-\frac{a}{s}}}, \text{ and,} \tag{1}$$

$$\text{Architecture 2: } P_S = \frac{\epsilon + \frac{a}{s}\frac{1-e^{-w(s\mu-\lambda)}}{1-\frac{a}{s}}}{\epsilon + \frac{a}{s}\frac{1-\frac{a}{s}e^{-w(s\mu-\lambda)}}{1-\frac{a}{s}}}, \text{ and, } f_W(t) = \lambda \frac{e^{-t(s\mu-\lambda)}}{\epsilon + \frac{a}{s}\frac{1-\frac{a}{s}e^{-w(s\mu-\lambda)}}{1-\frac{a}{s}}}\mathbb{1}_{t<w} + \delta_{t=w}\frac{\frac{a}{s}e^{-w(s\mu-\lambda)}}{\epsilon + \frac{a}{s}\frac{1-\frac{a}{s}e^{-w(s\mu-\lambda)}}{1-\frac{a}{s}}}, \tag{2}$$

where δ and $\mathbb{1}$ are the Dirac and the Indicator functions, and $\epsilon = \frac{\sum_{x=0}^{s-1}\frac{a^x}{x!}}{\frac{\left(\frac{a}{s}\right)^{s-1}}{(s-1)!}}$.

We capture the wait-dissatisfaction by a non-negative and increasing function, $g(t)$, for $t \geq 0$. As the

proportion of served customers by agents is identical in the two systems, we obtain the following relation between n and w:

$$e^{-w(s\mu-\lambda)} = \left(\frac{a}{s}\right)^n. \tag{3}$$

Thus, comparing the two systems results in comparing

$$I_1 = \int_{t=0}^{\infty} e^{-s\mu t} \sum_{x=0}^{n-1} \frac{(\lambda t)^x}{x!} g(t)dt + \frac{g(0)}{\lambda}\left(\frac{a}{s}\right)^{n+1}, \text{ for Architecture 1, and,}$$
$$I_2 = \int_{t=0}^{w} e^{-t(s\mu-\lambda)} g(t)dt + \frac{g(w)}{\lambda}\frac{a}{s}e^{-w(s\mu-\lambda)}, \text{ for Architecture 2.}$$

It should be noted that these expressions are proportional to the wait-dissatisfaction of all customers (served both by agents and by robots). By removing the terms proportional with $g(0)$ and $g(w)$, we reduce the comparison to the wait-dissatisfaction of customers served by agents. In the case where the function $g(t)$ is a continuous and infinitely differentiable function in t, we can show that

$$I_1 = \sum_{x=0}^{n-1} \frac{\lambda^x}{x!} \sum_{k=0}^{\infty} \frac{g^{(k)}(0)(k+x)!}{(s\mu)^{k+1+x}k!} + \frac{g(0)}{\lambda}\left(\frac{a}{s}\right)^{n+1}, \text{ and, } I_2 = \sum_{k=0}^{\infty} \frac{g^{(k)}(0^+) - e^{-w(s\mu-\lambda)}g^{(k)}(w)}{(s\mu-\lambda)^{k+1}} + \frac{g(w)}{\lambda}\frac{a}{s}e^{-w(s\mu-\lambda)}.$$

These expressions will be used in Theorem 1 to compare the two architectures for different performance measures.

Theorem 1. *The following holds:*

- *For all customers, the probability of delay and expected wait are lowest for Architecture 1. The second moment of the wait is lowest for Architecture 2 if and only if $n \leq 2e$, or $n > 2e$ and $a/s \geq x_2^n$ or $a/s \leq x_1^n$, where x_1^n and x_2^n are the two roots of $n\left[(x-1)^2 - x\left(\ln(x)\right)^2\right] + 2\ln(x) + (x-1)(x-3) = 0$, with $0 < x_1^n \leq x_2^n < 1$.*

- *For customers served by agents, the probability of delay, expected wait, wait percentiles, and the second moment of the wait are lowest for Architecture 2.*

Theorem 1 gives us a conclusion for most performance measures. For the expected wait, customers served by agents have a lower wait in the second architecture than in the first one, while the opposite is true when all customers are considered (served by robots and by agents). This result is important as it shows that it is advisable to implement the first architecture when the global expected wait is the objective, as in many service systems. However, in a system where human agents have sales goals in addition to respond to customers'

demand, the second statement of the theorem shows that the second architecture provides better performance measures for served customers, possibly leading to more revenues for the call center.

3.2 Numerical comparison

Comparison for higher moments of the wait and wait time percentiles for all customers can only be made numerically. For higher moments of the wait for all customers, we consider the function $g(t) = t^k$, from which we derive

$$I_1 = \frac{1}{(s\mu)^{k+1}} \sum_{i=0}^{n-1} \frac{(a/s)^i (k+i)!}{i!} = \frac{k!}{(s\mu)^{k+1}(1-a/s)^{k+1}} \left[1 - (a/s)^n \sum_{i=0}^{k} \frac{(-1)^i (n+k)!}{(n+i)i!(n-1)!(k-i)!} \left(\frac{a}{s}\right)^i \right], \text{ and,}$$

$$I_2 = \frac{k!}{(s\mu)^{k+1}(1-a/s)^{k+1}} \left[1 - (a/s)^n \sum_{i=0}^{k} \frac{(-n\ln(a/s))^i}{i!} + (a/s)^n \frac{(-n\ln(a/s))^k}{k!}(1-a/s) \right].$$

In Table 1, we compare I_1 and I_2 for the k^{th} moment of the wait for different values of the arrival rate. The last column provides the relative difference between I_1 and I_2 computed as $RD = \frac{I_2 - I_1}{I_2}$. As proven in Theorem 1, we observe that $I_1 \leq I_2$, when $k = 1$. When $k \geq 2$, we instead have $I_2 \geq I_1$. The difference between I_1 and I_2 increases with the arrival rate and the parameter k.

Table 1: Comparison between I_1 and I_2 for the k^{th} moment of the wait ($s = 10$, $n = 10$, $\mu = 1$)

	λ	I_1	I_2	RD
$k=1$	1	0.010	0.010	0.000%
	5	0.040	0.040	-0.150%
	9	0.303	0.321	-5.627%
	11	1.000	1.126	-11.156%
	15	9.266	11.762	-21.218%
	20	92.170	131.727	-30.029%
$k=2$	1	0.003	0.003	0.000%
	5	0.016	0.016	0.019%
	9	0.222	0.220	0.998%
	11	0.853	0.835	2.162%
	15	8.980	8.542	5.125%
	20	94.206	86.247	9.228%
$k=3$	1	0.001	0.001	0.000%
	5	0.009	0.009	1.292%
	9	0.205	0.175	16.791%
	11	0.864	0.672	28.651%
	15	9.836	6.375	54.274%
	20	106.906	57.089	87.263%
$k=4$	1	0.000	0.000	0.000%
	5	0.007	0.007	4.909%
	9	0.222	0.152	45.590%
	11	0.993	0.565	75.777%
	15	11.922	4.841	146.280%
	20	132.956	38.080	249.147%

From extensive numerical studies, we observe that $I_1 \geq I_2$, for $k \geq 3$. This indicates that the second architecture provides lower moments for the wait than the first one for $k \geq 3$. For $k = 2$, there exists some situations where $I_1 < I_2$. From Theorem 1, these situations occur when $a/s < 1$ and for large values of n.

This means that the first architecture can be better than the second one only in situations where the volume of customers served by robots is very low. In such cases, the two architectures display similar performance measures. Therefore, the second architecture is the most robust one to reduce the second moment of the wait. This shows another value of the corrective policy. It allows to reduce the variability in the experimented wait and consequently provides more fairness among customers. This aspect can be important when a service provider gives delay announcements. Having a low variability in the wait indicates that most customers will be served with a wait that is close to the announced delay. The accuracy of the announced delay is perceived as a high service quality by customers (Guo and Zipkin, 2007; Allon and Bassamboo, 2011; Yu et al., 2017, 2018).

For the wait percentiles of all customers, we consider $g(t) = \mathbb{1}_{t \geq z}$, where z can be understood as the maximum acceptable wait. In the case $z > w$, we have $I_2 = 0$, so the second architecture is the best. For $z \leq w$, we obtain

$$\begin{aligned} I_2 - I_1 &= \frac{e^{-s\mu z}}{s\mu(1-a/s)} \left(e^{\lambda z} - \sum_{x=0}^{n-1} \frac{(\lambda z)^x}{x!} - \left(\frac{a}{s}\right)^n \left(\frac{a}{s} e^{s\mu z} - \sum_{x=0}^{n-1} \frac{(s\mu z)^x}{x!} \right) \right) \\ &= \frac{1}{s\mu} \left(\frac{\left(\frac{a}{s}\right)^m - \left(\frac{a}{s}\right)^{n+1}}{1-\frac{a}{s}} + \left(\frac{a}{s}\right)^{\frac{m}{1-\frac{a}{s}}} \sum_{x=0}^{n-1} \frac{\left(\frac{a}{s}\right)^n - \left(\frac{a}{s}\right)^x}{1-\frac{a}{s}} \frac{\left(\frac{m \ln(a/s)}{\frac{a}{s}-1}\right)^x}{x!} \right), \end{aligned}$$

where $m \leq n$ is defined by $\left(\frac{a}{s}\right)^m = e^{-z(s\mu-\lambda)}$. By derivation, we can show that $I_2 - I_1$ is decreasing in m. We note that for $m = 0$ (i.e., for the probability of delay), from Theorem 1, we get $I_2 - I_1 \geq 0$.

In Table 2, we compute the difference $I_2 - I_1$ for a percentile of the wait, $P(W > z)$, where $z \leq w$. The parameters z and w are related to the parameters m and n via $\left(\frac{a}{s}\right)^m = e^{-s\mu z(1-a/s)}$, and $\left(\frac{a}{s}\right)^n = e^{-s\mu w(1-a/s)}$, respectively. The difference $I_2 - I_1$ can become negative when the arrival rate is low and when m gets close to n (i.e., when z gets close to w). For $m = n$, we can find situations where either $I_2 - I_1 > 0$ or $I_2 - I_1 < 0$. As the ratio a/s increases, the former case is more likely to happen than the latter one (see Table 2). This means that the first architecture provides lower wait percentiles for all customers in highly loaded systems. When the ratio a/s is lower, we find situations where the second architecture is better than the first one, as parameter z gets close to w. The numerical analysis for the expected excess leads to the same conclusion as that for the wait percentiles. This conclusion is important for service systems where percentiles of the wait are preferred than the average speed of answer to measure the service quality. This is the case in call centers where it is expected to have 80% of customers served in less than 20 seconds or in hospitals where 90% of emergency patients should be operated in less than 4 hours (Legros, 2016).

To summarize, applying a preventive policy with robots employed as an alternative to agents is the best

Table 2: Difference $I_2 - I_1$ for a percentile of the wait $P(W > z)$ ($s = 10$, $n = 10$, $\mu = 1$)

	m	$I_2 - I_1$		m	$I_2 - I_1$
$\lambda = 1$	5	2.018E-12	$\lambda = 11$	5	2.551E-01
	6	1.032E-13		6	2.437E-01
	7	-7.251E-13		7	2.165E-01
	8	-1.006E-12		8	1.642E-01
	9	-1.086E-12		9	7.800E-02
	10	-1.106E-12		10	-4.874E-02
$\lambda = 5$	5	8.525E-05	$\lambda = 15$	5	5.733E+00
	6	6.713E-05		6	5.622E+00
	7	4.095E-05		7	5.306E+00
	8	1.082E-05		8	4.564E+00
	9	-1.839E-05		9	3.058E+00
	10	-4.323E-05		10	2.928E-01
$\lambda = 9$	5	3.381E-02	$\lambda = 20$	5	1.022E+02
	6	3.139E-02		6	1.014E+02
	7	2.627E-02		7	9.868E+01
	8	1.760E-02		8	9.102E+01
	9	5.103E-03		9	7.198E+01
	10	-1.088E-02		10	2.854E+01

policy to minimize the most common wait-dissatisfaction measures such as expected wait and the probability of delay. However, the drawback to this architecture, in comparison with that where robots are positioned as a support to agents, is that it leads to higher variability in the experimented wait as measured by the k^{th} moments of the wait with $k \geq 2$, and by percentiles of the wait with a high value for the maximal acceptable wait. Furthermore, for all wait-related performance measures, customers served by agents have a lower wait-dissatisfaction with the corrective policy. This aspect can be important when the system manager shows more interest in these customers than in those served by robots.

4 Value of service interruption

In this section, we investigate the value of service interruption. To this end, we provide a measure of service-dissatisfaction in Section 4.1. Next, in Section 4.2, we show how service interruption can be employed to reduce service-dissatisfaction in a context without human agents. Finally, in Section 4.3, we study the value of letting human agents interrupt an automated service and compare this mechanism with the preventive policy of Section 3.

4.1 Measuring service-dissatisfaction

In the previous section, we stated a preference for agents as compared to robots and discussed the way agents should be placed in the system to manage wait-dissatisfaction. In this section, we construct a measure of *service-dissatisfaction* for the interaction between a customer and a robot agent to explore the potential of service interruption by agents. Since robot call centers are not yet widely implemented, we do not have comprehensive agreement based on empirical analyses of how service-dissatisfaction should be measured. However,

from knowledge of interaction between customers and human agents, we can guess what could go wrong with such interaction. As robots give standardized answers to complex questions, the length of service could be long, with several questions and answers not leading to a satisfactory end of service. Thus, the length of service is likely to be positively correlated with customers' frustration. Another measure of dissatisfaction is the retry feature. After a series of questions and answers that do not lead to a solution, a customer may lose patience and restart a call from zero. The number of retried calls and the overall length of service can be easily measured by the system due to customer's identification. Finally, dissatisfaction can be due to an interaction that fails to lead to a solution for the customer. In this case, a customer abandons the service without answers. In practice, this metric is difficult to measure as it can be difficult to identify whether a service termination is due to an abandonment or to a customer being satisfied by adequate answers. Customers can subsequently be sent an email to ask about their service experience and get a better understanding of the probability of success. One way to control service-dissatisfaction is to disconnect a customer from the service interaction and to propose alternative channels like visiting a website, using a chat, or being routed to an agent. We explore the latter feature in this section. In short, our aim is to capture dissatisfaction through (i) the probability of abandoning service, (ii) the probability of being disconnected from the service, (iii) the expected time spent in service, and (iv) the expected number of retries.

To evaluate these performance measures, we modelled the interaction between a customer and a robot agent by a series of questions receiving an answer. Each question and its related answer is modelled by an exponential duration with rate γ. Having a random duration for questions means that customers are heterogeneous and that the call center cannot anticipate how long each customer will take to formulate the question. The speed of the robot agent's answer is also random due to the time needed to analyze the sentences pronounced by the customers. The first interaction is assumed to be an identification step which cannot lead to a service completion. If the call center decides to disconnect customers after a given number of questions, we introduce a parameter d, such that a customer is disconnected after receiving an answer to the d^{th} question. Having an infinite value for d means that disconnection is not permitted. Following the identification step, we assume that customers leave the system after an exponential duration with rate β. This duration can be the service time or the abandonment time. We assume that a service is successful with probability q, or leads to abandonment from the service with probability $1 - q$. Finally, during service, a customer may get frustrated, subsequently retrying to recall after spending a time in the service that is exponentially distributed with rate θ. The process representing the time spent by a customer in service can be represented by a continuous time Markov chain, where a state of the system, x, represents the number of questions asked by a customer. The transitions in the Markov chain are as follows:

1. A phase increase with rate γ, for $0 \leq x \leq d$, which changes the state to $x+1$ if $x < d$ and to a disconnection from service if $x = d$.

2. An end of service or an abandonment with rate β, for $1 \leq x \leq d$, which leads to a customer ending the interaction.

3. A callback with rate θ, for $1 \leq x \leq d$, which leads to a restart of service from state $x = 0$.

The Markov chain is depicted in Figure 2.

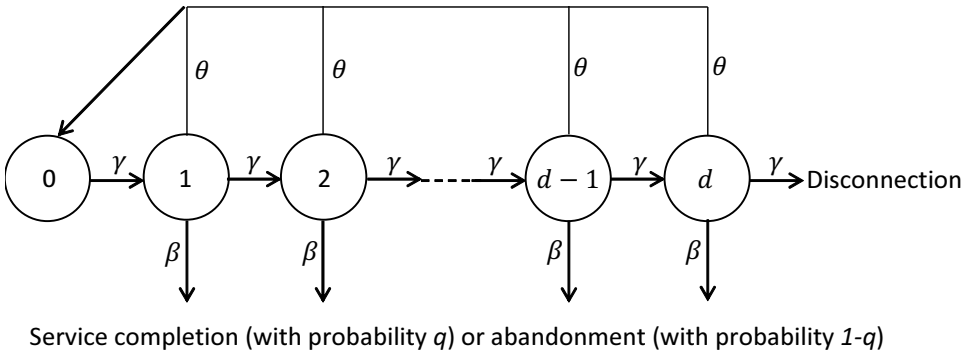


Figure 2: Markov chain for the customer-robot interaction

We denote by T_x^D, and T_x^S, the Laplace Transform (LT) in the variable t, of the Distribution Function (DF) of the first passage time from state x to state "disconnected" and "served or abandoned", respectively. We also denote by T_x, the LT of the DF from state x to state "disconnected, served or abandoned". This gives us $T_x = T_x^D + T_x^S$. Applying the *first-step analysis* to the discrete-time Markov chain (e.g., see Kulkarni (2016), p.162), we have the following finite sets of equations for $d \geq 1$:

$$
\begin{aligned}
&T_0^D(\gamma + t) = \gamma T_1^D, \;\; T_x^D(\theta + \gamma + \beta + t) = \gamma T_{x+1}^D + \theta T_0^D, \text{ for } 1 \leq x \leq d-1, \text{ and} \\
&T_d^D(\theta + \gamma + \beta + t) = \gamma + \theta T_0^D, \text{ for } x = d, \\
&T_0^S(\gamma + t) = \gamma T_1^S, \;\; T_x^S(\theta + \gamma + \beta + t) = \beta + \gamma T_{x+1}^S + \theta T_0^S, \text{ for } 1 \leq x \leq d-1, \text{ and,} \\
&T_d^S(\theta + \gamma + \beta + t) = \beta + \theta T_0^S, \text{ for } x = d, \\
&T_0(\gamma + t) = \gamma T_1, \;\; T_x(\theta + \gamma + \beta + t) = \beta + \gamma T_{x+1} + \theta T_0, \text{ for } 1 \leq x \leq d-1, \text{ and,} \\
&T_d(\theta + \gamma + \beta + t) = \beta + \gamma + \theta T_0, \text{ for } x = d.
\end{aligned}
\tag{4}
$$

For $d = 0$, we have

$$
T_0^D(\gamma + t) = \gamma, \;\; T_0^S(\gamma + t) = 0, \text{ and, } T_0(\gamma + t) = \gamma. \tag{5}
$$

In Proposition 1, we give the expressions of T_0^D, T_0^S, and T_0 found by solving Equations (4) and (5).

Proposition 1. *For $d = 0$, we have*

$$T_0^D = T_0 = \frac{\gamma}{\gamma + t}, \text{ and, } T_0^S = 0.$$

For $d \geq 1$, we have

$$T_0^D = \frac{\gamma(\theta + \beta + t)}{\gamma\theta + \left(\frac{\theta+\gamma+\beta+t}{\gamma}\right)^d [t(t + \theta + \beta + \gamma) + \gamma\beta]},$$

$$T_0 = \frac{\gamma^2(t + \theta) + \gamma^2\beta\left(\frac{\theta+\gamma+\beta+t}{\gamma}\right)^d}{\theta(t^2 + t(\beta + \gamma + \theta) + \gamma^2) + \gamma(t + \gamma)(t + \beta)\left(\frac{\theta+\gamma+\beta+t}{\gamma}\right)^d}, \text{ and, } T_0^S = T_0 - T_0^D.$$

In Corollary 1, we deduce the expressions of the probability to be disconnected, P_D, the probability to abandon service P_A, the expected time spent in service, $E(T)$, and the expected number of retries, $E(R)$.

Corollary 1. *For $d = 0$, we have*

$$P_D = 1, \quad P_A = 0, \quad E(T) = \frac{1}{\gamma}, \text{ and, } E(R) = 0.$$

For $d \geq 1$, we have

$$P_D = \frac{\theta + \beta}{\theta + \beta\left(\frac{\theta+\gamma+\beta}{\gamma}\right)^d}, \quad P_A = (1 - q)\frac{\beta\left(\left(\frac{\theta+\gamma+\beta}{\gamma}\right)^d - 1\right)}{\theta + \beta\left(\frac{\theta+\gamma+\beta}{\gamma}\right)^d},$$

$$E(T) = \frac{\theta^2 - \gamma^2 + \theta(\beta + \gamma) + \gamma(\beta + \gamma)\left(\frac{\theta+\beta+\gamma}{\gamma}\right)^d}{\gamma^2\left[\theta + \beta\left(\frac{\theta+\beta+\gamma}{\gamma}\right)^d\right]}, \text{ and, } E(R) = \frac{\theta\left(\left(\frac{\theta+\beta+\gamma}{\gamma}\right)^d - 1\right)}{\theta + \beta\left(\frac{\theta+\beta+\gamma}{\gamma}\right)^d}.$$

The case without disconnection can be deduced by letting d tend to infinity in the expressions of Corollary 1. This leads to $P_D = 0$, $P_A = 1 - q$, $E(T) = \frac{1}{\gamma} + \frac{1}{\beta}$, and $E(R) = 1 + \frac{\theta}{\beta}$. Using the expression of T_0, we can also find the other moments of T. We have $E(T^k) = (-1)^k \frac{\partial^k T_0}{\partial t^k}|_{t=0}$, for the k^{th} moment of T. In particular, when $d \longrightarrow \infty$, we get the standard-deviation of T, $\sigma_T = \sqrt{\frac{1}{\gamma^2} + \frac{1}{\beta^2}}$.

We define customer service-dissatisfaction, SD, by a linear combination of the different metrics, P_D, P_A, $E(T)$, and $E(R)$:

$$SD = c_D P_D + c_A P_A + c_T E(T) + c_R E(R). \tag{6}$$

The cost parameters c_D, c_A, c_T, and c_R represent the costs associated with a disconnection, an abandonment, a time unit spent in service, and a callback.

4.2 Optimizing the disconnection policy without human agents

One value of automated services is that they can be implemented 24/7, even when there is no agent available. In this first analysis of the service dissatisfaction, we investigate how the disconnection threshold can be optimized when there is no agent available. Without agents, disconnection is negatively perceived by customers as disconnected customers are not served at all. Therefore, having a large d is interesting as it increases the proportion of efficiently served customers. However, it may also reduce their satisfaction. We have

$$\frac{\partial SD}{\partial d} = (\theta+\beta)\left(\frac{\theta+\beta+\gamma}{\gamma}\right)^d \ln\left(\frac{\theta+\beta+\gamma}{\gamma}\right) \frac{-c_D\beta + (1-q)\beta c_A + c_T\left[1-\frac{\theta\beta}{\gamma^2}\right] + c_R\theta}{\left(\theta+\beta\left(\frac{\theta+\beta+\gamma}{\gamma}\right)^d\right)^2}.$$

This shows that depending on the sign of $-c_D\beta + (1-q)\beta c_A + c_T\left[1-\frac{\theta\beta}{\gamma^2}\right]$ either SD is strictly increasing or strictly decreasing in d. The cost for disconnection should be such that $c_D\beta > (1-q)\beta c_A + c_T\left[1-\frac{\theta\beta}{\gamma^2}\right]$ (i.e., having a large disconnection cost) to have a decreasing dissatisfaction in d. In such case, we should set $d=\infty$ as it minimizes SD and maximizes the proportion of served customers when there is no available human agent. Otherwise, the control parameter d should be set such that a sufficient proportion of customers is successfully served by an automated service while minimizing customers' dissatisfaction. This optimization question can be formulated as

$$\begin{cases} \text{Minimize } SD \\ \text{subject to } P_e \geq \overline{P_e}, \end{cases} \tag{7}$$

where P_e is the proportion of efficiently served customers by robots and $\overline{P_e}$ is the objective that the system manager wishes to achieve. We have $P_e = 1 - P_A - P_D$. This leads to $P_e = q\frac{\beta\left(\left(\frac{\theta+\gamma+\beta}{\gamma}\right)^d - 1\right)}{\theta+\beta\left(\frac{\theta+\gamma+\beta}{\gamma}\right)^d}$. Note that the proportion of disconnected customers is considered as non-efficiently served as there is no available human agents to take over the service.

The maximal value that P_e can achieve is $P_e = q$ when d tends to infinity. Therefore Problem (7) can be solved only if $\overline{P_e} \leq q$. Since we consider the case where SD is strictly increasing in d. The optimal value for d, d^*, is obtained by solving $P_e = \overline{P_e}$. The solution of this equation is given by

$$d^* = \frac{\ln\left(\frac{q\beta+\theta\overline{P_e}}{\beta(q-\overline{P_e})}\right)}{\ln\left(\frac{\theta+\beta+\gamma}{\gamma}\right)}, \tag{8}$$

provided that $\overline{P_e} \leq q$.

In summary, this first step of the analysis -when there is no human agent involved- shows that service disconnection can be helpful to interrupt an inefficient interaction between a customer and robot. By deciding to interrupt a conversation, the system stops the increasing frustration of a customer. Although the customer is then left without service, the frustration can be lower than if this customer wastes time in inefficient interactions. To improve the service quality, we explore in the next subsection how human agents can turn service disconnection into a positive end of service.

4.3 Employing agents to take over the robot service

We now explore how agents can help to reduce dissatisfaction from the robot service by taking over the end of service. To prevent customers from spending too long in the robot service, disconnection is used to send customers to an infinite capacity queueing system where they will be served by a group of s agents with a service time duration which is exponentially distributed with rate μ as in the previous section. Figure 3 depicts this strategy.

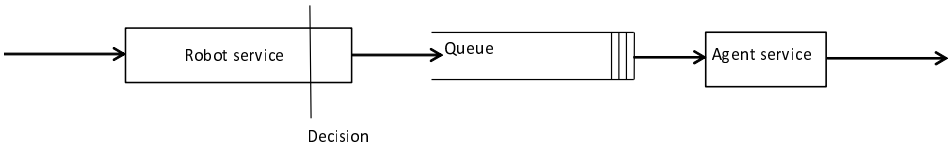


Figure 3: Agents taking over the end of robot service

Disconnection is thus no longer perceived as something negative but as a way to resolve customer's demand more efficiently. Therefore, with disconnection leading to a service from agents, we assume that $c_D = 0$. Choosing to propose an agent service to customers who have already spent time with robots provides a way to select the most dissatisfied customers for a positive end of service. To evaluate this policy, we need to distinguish the dissatisfaction of customers who are disconnected from those who are not as only the former will be sent to the agents' queue. We denote by D, the event being disconnected and by $\overline{D}$, the complementary event which means either being served or abandoning the system before disconnection. In Proposition 2, we provide the different components of dissatisfaction given D and $\overline{D}$.

Proposition 2. *For disconnected customers, we have*

$$P_{D|D} = 1, P_{A|D} = 0, E(T|D) = \frac{-\gamma\theta(\theta+\beta+\gamma) + \left(\frac{\theta+\beta+\gamma}{\gamma}\right)^d ((\theta+\beta)^3 + \gamma(2\theta^2+\beta^2+3\beta\theta+\gamma\theta) + d\gamma\beta(\theta+\beta))}{\gamma(\theta+\beta+\gamma)(\theta+\beta)\left(\theta+\beta\left(\frac{\theta+\beta+\gamma}{\gamma}\right)^d\right)},$$

$$\textit{and } E(R|D) = \frac{\theta\left[\left(\frac{\beta+\gamma+\theta}{\gamma}\right)^d - 1\right]}{\theta+\beta\left(\frac{\beta+\gamma+\theta}{\gamma}\right)^d}.$$

For customers who left the system before being disconnected, we have

$$P_{D|\overline{D}} = 0, P_{A|\overline{D}} = 1-q,$$

$$E(T|\overline{D}) =$$

$$\frac{\gamma\beta(\beta+\gamma)\left(\frac{\theta+\beta+\gamma}{\gamma}\right)^{2d} + \left[\theta^2(\beta-\gamma)+\beta^2(\theta-\gamma)-\gamma^2\beta\right]\left(\frac{\theta+\beta+\gamma}{\gamma}\right)^d - d\gamma\beta(\theta+\beta)\left(\frac{\theta+\beta+\gamma}{\gamma}\right)^{d-1} + \theta^2(\theta+\beta+\gamma)}{\gamma^2\beta\left(\left(\frac{\theta+\beta+\gamma}{\gamma}\right)^d - 1\right)\left(\theta+\beta\left(\frac{\theta+\beta+\gamma}{\gamma}\right)^d\right)},$$

$$\textit{and } E(R|\overline{D}) = \frac{\theta\left(\left(\frac{\theta+\beta+\gamma}{\gamma}\right)^d - 1\right)}{\theta+\beta\left(\frac{\theta+\beta+\gamma}{\gamma}\right)^d}.$$

The results of Proposition 2 indicate that disconnected customers are not necessarily those for whom service-dissatisfaction is the highest. Although the time spent in the system is the longest for disconnected customers, the probability of abandoning the system without being successfully served only exists for non-disconnected customers and the expected number of retries is independent of the event "being disconnected". This means that the main value for disconnecting customers after spending time in robot service is to reduce the expected service time. We deduce from Proposition 2 that the service-dissatisfaction for disconnected customers, SD^D, is given by

$$SD^D = c_T \frac{-\gamma\theta(\theta+\beta+\gamma) + \left(\frac{\theta+\beta+\gamma}{\gamma}\right)^d ((\theta+\beta)^3 + \gamma(2\theta^2+\beta^2+3\beta\theta+\gamma\theta) + d\gamma\beta(\theta+\beta))}{\gamma(\theta+\beta+\gamma)(\theta+\beta)\left(\theta+\beta\left(\frac{\theta+\beta+\gamma}{\gamma}\right)^d\right)}$$
$$+ c_R \frac{\theta\left[\left(\frac{\beta+\gamma+\theta}{\gamma}\right)^d - 1\right]}{\theta+\beta\left(\frac{\beta+\gamma+\theta}{\gamma}\right)^d}.$$

We can show that this function is increasing in d. Therefore, by disconnecting customers after a small number of interactions, the system manager reduces the service-dissatisfaction of customers who will be served later by agents. However, having a too small value for d may result in a too high volume of customers in the queue that

cannot be served with a sufficiently low wait. Therefore, we complete Problem (7) by considering a constraint for the expected wait in the queue. The optimization problem becomes

$$\begin{cases} \text{Minimize } SD \\ \text{subject to } P_e \geq \overline{P_e} \text{ and } E(W) \leq \overline{E(W)}, \end{cases} \tag{9}$$

where $E(W)$ is the expected wait in the queue and $\overline{E(W)}$ is the service level objective for the expected wait. The proportion of efficiently served customers by agents or by robots, P_e, is given by $P_e = q(1 - P_D) + P_D = q + (1 - q)P_D$, since disconnected customers are all served by agents. This means that the proportion of efficiently served customers decreases with d. Therefore, for both SD and P_e, the threshold d should be decreased. Each robot service is independent and the number of robots is infinite. Therefore, the robot service can be viewed as an M/G/∞ queue. The output of this queue is also Poisson (Mirasol, 1963). Therefore, the arrival process in the human queue is Poisson and the agents' queue behaves as an M/M/s queue. The arrival rate in the queue, λ_h, is given by $\lambda_h = \lambda P_D$. Therefore, the constraint $E(W) \leq \overline{E(W)}$ results in controlling the volume of customers routed to agents.

With $s = 1$, if with $d = 0$ (i.e., if all customers are disconnected), we get $E(W) \leq \overline{E(W)}$, then $d^* = 0$ is optimal. Otherwise, with $\lambda_h = \mu \frac{\mu \overline{E(W)}}{1+\mu \overline{E(W)}}$, we get $E(W) = \overline{E(W)}$. Therefore, the solution of Problem (9) is given by

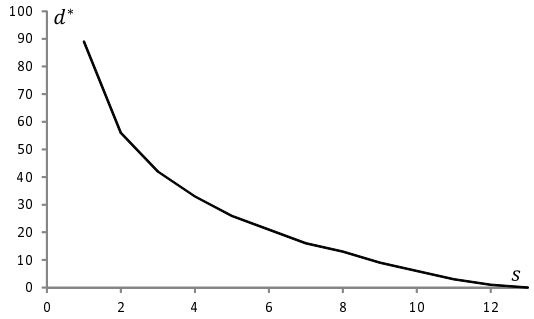
$$d^* = \frac{\ln\left(\frac{1}{\beta}\left[\frac{\lambda(\theta+\beta)(1+\mu\overline{E(W)})}{\mu^2 \overline{E(W)}} - \theta\right]\right)}{\ln\left(\frac{\beta+\theta+\gamma}{\gamma}\right)}.$$

With $s \geq 2$, the optimal value for d^* cannot be obtained in closed-form. If with $d = 0$, we get $E(W) \leq \overline{E(W)}$, then $d^* = 0$ is optimal. Otherwise, we should determine λ_h such that $E(W) = \overline{E(W)}$. In other words, we need to solve the equation in λ_h:

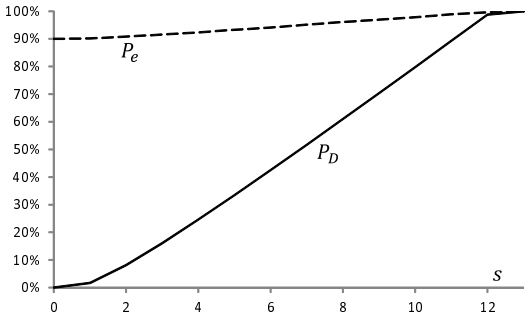
$$s\mu\left(1 - \frac{\lambda_h}{s\mu}\right)\left(\sum_{k=0}^{s-1} \frac{s!\left(1 - \frac{\lambda_h}{s\mu}\right)\left(\frac{\lambda_h}{\mu}\right)^{k-s}}{k!} + 1\right) = \overline{E(W)}^{-1}. \tag{10}$$

Equation (10) can be transformed into a polynomial equation of degree s which can be solved numerically. Once the solution of Equation (10) is found, the optimal threshold, d^*, can be obtained using $\lambda P_D = \lambda_h$, which leads to

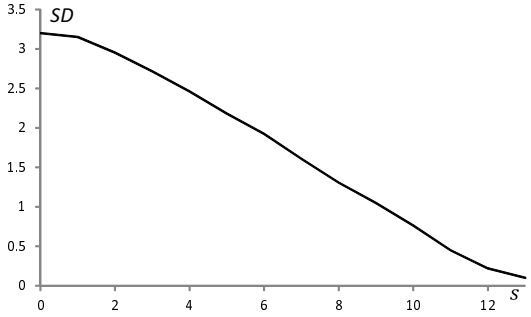
$$d^* = \frac{\ln\left(\frac{1}{\beta}\left[\frac{\lambda(\theta+\beta)}{\lambda_h} - \theta\right]\right)}{\ln\left(\frac{\beta+\theta+\gamma}{\gamma}\right)}.$$



(a) Optimal disconnection threshold



(b) Proportion of disconnected customers and proportion of served customers



(c) Customers' service-dissatisfaction

Figure 4: Impact of the number of agents ($\lambda = 10$, $\mu = 1$, $\beta = 0.4$, $\theta = 0.1$, $\gamma = 10$, $q = 90\%$, $\overline{E(W)} = 0.2$, $c_A = 1$, $c_T = 1$, and $c_R = 2$)

In Figure 4, we show the effect of the number of agents on the disconnection threshold, the proportion of served and disconnected customers, and the customers' service-dissatisfaction. Having more agents allows us to reduce the disconnection threshold, leading customers to spend a shorter amount of time in the robot service and to be less dissatisfied. We observe that the proportion of served customers increases less than the proportion of disconnected customers when the number of agents increases. This means that disconnection is more effective to reduce dissatisfaction than to obtain a higher volume of served customers.

We now evaluate the disconnection policy presented in this section with the policy where robots are placed as an alternative to agents (see Figure 1(a)). The latter policy was shown to allow a lower expected wait than the policy where robots are placed in support of the agents teams. The idea is to determine whether disconnecting customers from a robot service offers an improvement compared to making decisions on arrival. An illustration of this comparison is given in Table 3. In the first column, we provide the number of agents. The second, third, fourth and fifth columns provide the disconnection threshold, the proportion of successfully served customers, the expected service-dissatisfaction, and the standard-deviation of service-dissatisfaction under the policy with disconnection. The next four columns provide the same metrics for the preventive policy. The last three columns indicate the relative differences in P_e, SD, and σ_{SD}, defined as $RD(P_e) = \frac{P_e(\text{Preventive policy}) - P_e(\text{Policy with disconnection})}{P_e(\text{Preventive policy})}$, $RD(SD) = \frac{SD(\text{Preventive policy}) - SD(\text{Policy with disconnection})}{SD(\text{Preventive policy})}$, and $RD(\sigma_{SD}) =$

$\frac{\sigma_{SD}(\text{Preventive policy})-\sigma_{SD}(\text{Policy with disconnection})}{\sigma_{SD}(\text{Preventive policy})}$, respectively. In this illustration the main constraint is the expected wait. The thresholds d^* and n^* are adjusted such that $E(W) = 0.2$ in both systems. Note that the parameters d^* and n^* are real in the table. In practice, it means that randomization between two adjacent thresholds is made possible.

Table 3: Comparison between the two policies ($\lambda = 10$, $\mu = 1$, $\beta = 0.4$, $\theta = 0.1$, $\gamma = 10$, $q = 90\%$, $\overline{E(W)} = 0.2$, $\overline{P_e} = 90\%$, $c_A = 1$, $c_T = 1$, and $c_R = 2$)

	Policy with disconnection				Preventive policy						
s	d^*	P_e	SD	σ_{SD}	n^*	P_e	SD	σ_{SD}	$RD(P_e)$	$RD(SD)$	$RD(\sigma_{SD})$
0	∞	90.00%	3.200	2.502	–	90.00%	3.200	2.502	0.00%	0.00%	0.00%
1	88.627	90.16%	3.150	2.350	0.235	90.95%	2.897	2.265	0.86%	-8.03%	-3.77%
2	55.805	90.80%	2.952	1.980	0.521	91.92%	2.587	2.023	1.21%	-12.37%	2.08%
3	41.546	91.56%	2.716	1.648	0.841	92.89%	2.276	1.780	1.43%	-16.21%	7.43%
4	32.480	92.38%	2.463	1.358	1.195	93.85%	1.967	1.538	1.57%	-20.10%	11.75%
5	25.812	93.28%	2.182	1.104	1.593	94.80%	1.662	1.300	1.60%	-23.81%	15.03%
6	20.508	94.12%	1.924	0.884	2.051	95.74%	1.363	1.066	1.69%	-29.14%	17.10%
7	16.090	95.14%	1.608	0.693	2.599	96.65%	1.073	0.839	1.56%	-33.25%	17.41%
8	12.278	96.11%	1.307	0.531	3.286	97.51%	0.796	0.623	1.44%	-39.04%	14.78%
9	8.911	96.94%	1.050	0.396	4.213	98.31%	0.539	0.422	1.40%	-48.61%	6.10%
10	5.883	97.86%	0.764	0.288	5.490	98.97%	0.330	0.258	1.12%	-56.76%	-11.71%

We observe that the preventive policy provides a significantly lower service-dissatisfaction than the policy with disconnection. The difference is most significant when the proportion of customers sent to the queue is high (i.e., when the number of agents is high). This observation is in line with the one in the previous section which showed that the preventive policy was better to manage the expected wait. Moreover, we observe that the preventive policy allows the system to serve a higher volume of customers. Nevertheless, the difference between the two policies on this level remains small (less than 1.7% difference in our illustration). The main drawback of the preventive policy is the variability in service-dissatisfaction. As for the wait-dissatisfaction, the preventive policy leads to a high second order moment of service-dissatisfaction, which makes its standard-deviation higher than for a policy with disconnection in most cases. This aspect should be taken into account when the system manager cares about fairness among customers. For an overview of the notion of fairness in queues, we refer to Avi-Itzhak and Levy (2004). In our case, fairness among customers means that customers receive a similar service experience in terms of service-dissatisfaction.

5 Conclusion

Our paper investigated the positioning of agents in a service architecture where the service can be delivered by robots. Assuming that customers have a preference for a human service, we examined whether robots should be placed as a support or as an alternative to humans and whether a robot service should be interrupted. To answer these questions, we determined the effect of different policies on wait and service-dissatisfaction. When considering expected values like expected-wait or expected service-dissatisfaction, a preventive policy where

customers are either sent to robots or to agents on arrival is shown to outperform the other architectures. Nevertheless, when considering wait-dissatisfaction, higher moments of the wait can be better managed when the routing of customers to robots is decided after experimenting some wait. Moreover, a lower wait can be offered to customers who are served by agents when customers scheduling is made after waiting. Next, we built a measure of service-dissatisfaction to account for the time spent in service, the number of retries and the probability of abandonment due to a robotized service. This measure allowed us to compare the preventive policy with one where customers are disconnected from a robot service if it takes too long. Again, expected service-dissatisfaction is shown to be lower with a preventive policy but the standard deviation of service-dissatisfaction is higher in most cases. Therefore, choosing a disconnection strategy compared to a preventive policy depends on the importance given to fairness among customers.

This opens up several avenues for future research. Once robot services are more broadly adopted in call centers, empirical studies could help to assess how the wait and the service are perceived by customers. Qualitatively, this may not change our conclusions but it may help to quantify the improvement when selecting one architecture over another. We assumed that human agents are in constant number in our analysis. This makes our results valid for intervals of time of one or two hours. It may be interesting to investigate how service-dissatisfaction may evolve when the staffing level is changed. Considering a call center model, we assumed that robots are in infinite number. In other services, like restaurants, the number of robots may be limited. Therefore, one focus could be to investigate how agents can support robots in such cases. The customers' arrival process is assumed to be endogenously determined. It could be interesting to explore how customers may adjust their joining strategies to different architectures with robots. Finally, different model extensions can be considered such as having non-exponential distributions or time-dependent parameters.

References

Allon, G. and Bassamboo, A. (2011). The impact of delaying the delay announcements. *Operations Research*, 59(5):1198–1210.

Avi-Itzhak, B. and Levy, H. (2004). On measuring fairness in queues. *Advances in Applied Probability*, pages 919–936.

Baba, Y. (1986). On the $M^X/G/1$ queue with vacation time. *Operations Research Letters*, 5(2):93–98.

Baccelli, F. and Hebuterne, G. (1981). On queues with impatient customers. *Performance'81 North-Holland Publishing Company*, pages 159–179.

Bassamboo, A., Harrison, M., and Zeevi, A. (2005). Dynamic routing and admission control in high-volume service systems: Asymptotic analysis via multi-scale fluid limits. *Queueing Systems*, 51(3-4):249–285.

BCG (2015). *Reshoring of Manufacturing to the US Gains Momentum*. Boston Consulting Group. https://www.bcg.com/fr-fr/publications/2015/reshoring-of-manufacturing-to-the-us-gains-momentum, 2020-10-20.

Bennington, L., Cummane, J., and Conn, P. (2000). Customer satisfaction and call centers: an Australian study. *International Journal of Service Industry Management*.

Bountali, O. and Economou, A. (2017). Equilibrium joining strategies in batch service queueing systems. *European Journal of Operational Research*, 260(3):1142–1151.

Choudhury, G. and Deka, K. (2008). An M/G/1 retrial queueing system with two phases of service subject to the server breakdown and repair. *Performance Evaluation*, 65(10):714–724.

Cosyn, J. and Sigman, K. (2004). Stochastic networks: Admission and routing using penalty functions. *Queueing Systems*, 48(3-4):237–262.

CustomerServ (2018). *Humans vs. Robots: Why AI Won't Replace Humans in the Call Center*. CustomerServ. https://www.customerserv.com/blog/why-ai-wont-replace-humans-call-center, 2020-10-20.

Dialer360 (2020). *An expected future of robots in call center*. Call Center Software. https://www.dialer360.com/2016/11/an-expected-future-of-robots-in-call-center, 2020-10-20.

Dudin, A., Jacob, V., and Krishnamoorthy, A. (2015). A multi-server queueing system with service interruption, partial protection and repetition of service. *Annals of Operations Research*, 233(1):101–121.

Gans, N. and Zhou, Y. (2007). Call-routing schemes for call-center outsourcing. *Manufacturing & Service Operations Management*, 9(1):33–50.

Gao, S. and Wang, J. (2014). Performance and reliability analysis of an M/G/1-G retrial queue with orbital search and non-persistent customers. *European Journal of Operational Research*, 236(2):561–572.

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

Gurvich, I. and Perry, O. (2012). Overflow networks: Approximations and implications to call center outsourcing. *Operations research*, 60(4):996–1009.

Hudson, S., González-Gómez, H., and Rychalski, A. (2017). Call centers: is there an upside to the dissatisfied customer experience? *Journal of Business Strategy*.

Jain, M., Kaur, S., and Singh, P. (2019). Supplementary variable technique (SVT) for non-Markovian single server queue with service interruption (QSI). *Operational Research*, pages 1–44.

Koole, G. (2003). Redefining the service level in call centers. *Technical report, Department of Stochastics, Vrije Universiteit, Amsterdam*.

Koole, G. (2007). Monotonicity in markov reward and decision chains: Theory and applications. *Foundations and Trends in Stochastic Systems*, 1(1):1–76.

Koole, G., Nielson, B., and Nielson, T. (2012). First in line waiting times as a tool for analysing queueing systems. *Operations Research*, 60:1258–1266.

Ku, C. and Jordan, S. (2003). Near optimal admission control for multiserver loss queues in series. *European Journal of Operational Research*, 144(1):166–178.

Kulkarni, V. (2016). *Modeling and analysis of stochastic systems*. Crc Press.

Kumar, P. and Krishnamurthy, P. (2008). The impact of service-time uncertainty and anticipated congestion on customers' waiting-time decisions. *Journal of Service Research*, 10(3):282–292.

Lee, D. (1997). Analysis of a single server queue with semi-Markovian service interruption. *Queueing Systems*, 27(1-2):153–178.

Legros, B. (2016). Unintended consequences of optimizing a queue discipline for a service level defined by a percentile of the waiting time. *Operations Research Letters*, 44(6):839–845.

Legros, B. (2020). Late-rejection, a strategy to perform an overflow policy. *European Journal of Operational Research*, 281(1):66–76.

Legros, B., Jouini, O., and Koole, G. (2020). Should we wait before outsourcing? analysis of a revenue-generating blended contact center. *Manufacturing & Service Operations Management*.

Lin, K. and Ross, S. (2004). Optimal admission control for a single-server loss queue. *Journal of Applied Probability*, 41(2):535–546.

Lynch, T., Harper, K., and Radebaugh, C. (2016). Systems and methods for storing record of virtual agent interaction. US Patent 9,262,175.

Maglaras, C. and Van Mieghem, J. (2005). Queueing systems with leadtime constraints: A fluid-model approach for admission and sequencing control. *European journal of Operational Research*, 167(1):179–207.

Mirasol, N. (1963). Letter to the Editor-The output of an M/G/∞ queuing system is Poisson. *Operations Research*, 11(2):282–284.

Mitrany, I. and Avi-Itzhak, B. (1968). A many-server queue with service interruptions. *Operations Research*, 16(3):628–638.

Niyirora, J. and Zhuang, J. (2017). Fluid approximations and control of queues in emergency departments. *European Journal of Operational Research*, 261(3):1110–1124.

Örmeci, L. and Burnetas, A. (2005). Dynamic admission control for loss systems with batch arrivals. *Advances in Applied Probability*, 37(4):915–937.

Örmeci, L. and van der Wal, J. (2006). Admission policies for a two class loss system with general interarrival times. *Stochastic models*, 22(1):37–53.

Palakovich, J., Eigeman, J., McDaniel, C., Maringas, M., Chodavarapu, S., et al. (2017). Virtual agent proxy in a real-time chat service. US Patent 9,559,993.

Seeley, A., Kuzsma Jr, R., Shen, L., and Sturgeon, K. (2010). Testing using asynchronous automated virtual agent behavior. US Patent 7,822,803.

Storrie, D. (2019). The future of manufacturing in Europe. *Publications Office of the European Union, Luxembourg*.

Ward, A. and Kumar, S. (2008). Asymptotically optimal admission control of a queue with impatient customers. *Mathematics of Operations Research*, 33(1):167–202.

Xia, L., He, Q., and Alfa, A. (2017). Optimal control of state-dependent service rates in a MAP/M/1 queue. *IEEE Transactions on Automatic Control*, 62(10):4965–4979.

Xu, K. (2015). Necessity of future information in admission control. *Operations Research*, 63(5):1213–1226.

Yildirim, U. and Hasenbein, J. (2010). Admission control and pricing in a queue with batch arrivals. *Operations Research Letters*, 38(5):427–431.

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

Table 4: Table of notations

System parameters	
λ	Customers' arrival rate
s	Number of agents
μ	Service rate
a	Ratio λ/μ
γ	Phase-rate in the robot service
β	Service rate by a robot
θ	Retry rate
q	Probability of a successful robot service
c_D, c_A, c_T, and c_R	Cost parameters related to disconnection, abandonment, service time, and number of retries
Control parameters	
n	Queue-length threshold for the preventive policy
w	Wait threshold for the corrective policy
d	Disconnection threshold
Performance measures	
$f_W(t)$	Probability-density function for the wait, for $t \geq 0$
P_S	Probability of being served by agents
$g(t)$	Wait-dissatisfaction function, for $t \geq 0$
T_x, T_x^D, and T_x^S	Laplace Transforms of the distribution functions of the first passage time from state x to state "being out of the system", "disconnected", and "served or abandoned", respectively.
P_D	Probability of being disconnected
P_A	Probability of abandoning service
$E(T)$	Expected service time
$E(R)$	Expected number of retries
SD	Expected service-dissatisfaction
P_e	Probability of being successfully served either by a robot or an agent
σ_{SD}	Standard-deviation of the service-dissatisfaction
$E(W)$	Expected wait
$\overline{E(W)}$	Service-level objective for the expected wait
$\overline{P_e}$	Service-level objective for the probability of being successfully served

Yu, Q., Allon, G., Bassamboo, A., and Iravani, S. (2018). Managing customer expectations and priorities in service systems. *Management Science*, 64(8):3942–3970.

A Table of notations

B Proof of Theorem 1

Proof. We compare I_1 and I_2, to determine which architecture provides the lowest wait-related performance for all customers (served by agents and served by robots). By removing the terms proportional with $g(0)$ and $g(w)$ in the expressions of I_1 and I_2, we obtain $\tilde{I_1}$ and $\tilde{I_2}$, which will be used to compare the performance measures for customers served by agents.

Probability of delay. To measure the probability of delay, we have $g(t) = 1$, for $t > 0$. We obtain

$$I_1 = \frac{1}{s\mu}\frac{1-\left(\frac{a}{s}\right)^n}{1-\frac{a}{s}}, \text{ and } I_2 = \frac{1}{s\mu}\frac{1-\frac{a}{s}e^{-w(s\mu-\lambda)}}{1-\frac{a}{s}}.$$

Using Relation (3), we get

$$I_2 - I_1 = \frac{\left(\frac{a}{s}\right)^n}{s\mu} > 0.$$

Therefore, the probability of delay is better managed in the first system. For customers served by agents, the probability of delay is a special case of the wait percentile. Therefore, we only provide the comparison for the wait percentile in this proof.

Expected wait. We now consider the expected wait with $g(t) = t$. In this case, we get

$$I_1 = \frac{1}{(s\mu)^2}\sum_{x=0}^{n-1}(x+1)\left(\frac{a}{s}\right)^x = \frac{1}{(s\mu)^2}\frac{1-(n+1)\left(\frac{a}{s}\right)^n + n\left(\frac{a}{s}\right)^{n+1}}{\left(1-\frac{a}{s}\right)^2}, \text{ and}$$
$$I_2 = \frac{1}{(s\mu)^2}\frac{1-e^{-w(s\mu-\lambda)}\left(1+w\lambda(1-a/s)\right)}{\left(1-\frac{a}{s}\right)^2}.$$

Using Relation (3), we get

$$I_2 - I_1 = \frac{1}{(s\mu)^2}\frac{n\left(\frac{a}{s}\right)^n\left(\frac{a}{s}\ln\left(\frac{a}{s}\right)+1-\frac{a}{s}\right)}{\left(1-\frac{a}{s}\right)^2}.$$

The sign of this function is given by the sign of the function $f(x) = x\ln(x) + 1 - x$, for $x \geq 0$. We have $f'(x) = \ln(x)$. So, f has a minimum at $x = 1$, with $f(1) = 0$. This proves that $I_2 \geq I_1$. Therefore, again, the expected wait is lower in the first system. The comparison for customers served by agents can be deduced from the comparison of percentiles of the wait. Comparing percentiles of the wait means determining a first-order stochastic dominance between the wait in Architecture 1 and Architecture 2. As wait percentiles are the lowest for Architecture 2, the expected wait is also the lowest for Architecture 2.

Second moment of the wait. We now consider $g(t) = t^2$, for $t \geq 0$. We obtain

$$I_1 = \frac{A}{(s\mu)^3}\frac{2-(n+1)(n+2)\left(\frac{a}{s}\right)^n + 2n(n+2)\left(\frac{a}{s}\right)^{n+1} - n(n+1)\left(\frac{a}{s}\right)^{n+2}}{\left(1-\frac{a}{s}\right)^3}, \text{ and,}$$
$$I_2 = \frac{A}{(s\mu)^3}\frac{2-e^{-w(s\mu-\lambda)}\left(2+2ws\mu(1-a/s)+\frac{a}{s}w^2(s\mu)^2(1-a/s)^2\right)}{\left(1-\frac{a}{s}\right)^3}.$$

Using Relation (3), we get

$$I_2 - I_1 = \frac{1}{(s\mu)^3}\frac{n\left(\frac{a}{s}\right)^n\left(2\ln\left(\frac{a}{s}\right) - n\frac{a}{s}\left(\ln\left(\frac{a}{s}\right)\right)^2 + n + 3 - 2(n+2)\frac{a}{s} + (n+1)\left(\frac{a}{s}\right)^2\right)}{\left(1-\frac{a}{s}\right)^3}.$$

We need to study the sign of the function $f(x) = 2\ln(x) - nx(\ln(x))^2 + n + 3 - 2(n+2)x + (n+1)x^2$. We have for $x \geq 0$,

$$\begin{aligned}
f'(x) &= -n(\ln(x))^2 - 2n\ln(x) + 2(n+1)x + 2/x - 2(n+2), \\
f"(x) &= 2\frac{-nx\ln(x) + (n+1)x^2 - nx - 1}{x^2}, \text{ and,} \\
f^{(3)}(x) &= \frac{2}{x^3}\left(nx\ln(x) + 2\right).
\end{aligned}$$

The sign of $f^{(3)}(x)$ is given by the sign of $h(x) = nx\ln(x) + 2$. We have $h'(x) = n(\ln(x) + 1)$, for $x \geq 0$. Therefore, $h(x)$ has a minimum for $x = e^{-1}$. We have $h(e^{-1}) = 2 - ne^{-1}$.

- Case 1: $n \leq 2e$. In this case $h(x) \geq h(e^{-1}) \geq 0$, for $x \geq 0$ and $f^{(3)}(x) \geq 0$. Thus, $f"(x)$ is increasing in x. Since $f"(1) = 0$, then $f"(x)$ is negative for $x \leq 1$ and positive otherwise. Therefore, $f'(x)$ has a minimum for $x = 1$. We have $f'(1) = 0$, so $f'(x) \geq 0$, for $x \geq 0$. This prove that $f(x)$ is increasing in x. Again, $f(1) = 0$, so $f(x)$ is negative for $x \leq 1$ and positive otherwise. Therefore, the numerator in $I_2 - I_1$ is negative for $\lambda \leq s\mu$ and positive, otherwise. The denominator, $\left(1 - \frac{a}{s}\right)^3$, is positive for $\lambda \leq s\mu$ and negative, otherwise. This results in $I_2 - I_1 \leq 0$.

- Case 2: $n > 2e$. In this case $h(e^{-1}) < 0$. Therefore, there exist two unique roots of $h(x) = 0$, x_1 and x_2, such that $x_1 < e^{-1}$ and $x_2 > e^{-1}$. Since $h(x)$ is increasing for $x \geq e^{-1}$ and $h(1) = 2 > 0$, we also have $x_2 < 1$. This proves that $f"(x)$ is increasing on the intervals $[0, x_1]$ and $[x_2, \infty)$, and decreasing on the interval $[x_1, x_2]$. Given that $f"(1) = 0$, we have $f"(x) \geq 0$, for $x \geq 1$. So, $f'(x)$ is increasing for $x \geq 1$. Moreover, $f'(1) = 0$, so $f'(x)$ is also positive and $f(x)$ is increasing for $x \geq 1$. Finally, $f(1) = 0$, so $f(x) \geq 0$, for $x \geq 1$. Therefore, as in the case $n \leq 2e$, we have $I_2 - I_1 \leq 0$ if $\lambda \geq s\mu$. In the case $x < 1$, the function $f(x)$ can be written as a function of x and the parameter n as $f_n(x) = n\left[(x-1)^2 - x(\ln(x))^2\right] + 2\ln(x) + (x-1)(x-3)$. The function $x :\longrightarrow (x-1)^2 - x(\ln(x))^2$ is positive and decreasing in x. The function $x :\longrightarrow 2\ln(x) + (x-1)(x-3)$ is negative and increasing in n. Therefore, for each $x \in (0,1)$, if $n \geq -\frac{2\ln(x)+(x-1)(x-3)}{(x-1)^2 - x(\ln(x))^2}$, then the function $f_n(x) \geq 0$. Moreover, $f_n(0) = -\infty$, $f_n(1^-) = f_n'(1^-) = f_n''(1^-) = 0$ and $f_n^{(3)}(1^-) = 4 > 0$. This proves that either we have $f_n(x) \leq 0$, for $0 \leq x \leq 1$, or there exists two roots of $f_n(x) = 0$, x_1^n and x_2^n such that $f_n(x) \geq 0$ on the interval $[x_1^n, x_2^n]$.

We now consider customers served by agents, we have in this case

$$\tilde{I}_2 - \tilde{I}_1 = \frac{1}{(s\mu)^3}\frac{n\left(\frac{a}{s}\right)^n\left(2\ln\left(\frac{a}{s}\right) - n\left(\ln\left(\frac{a}{s}\right)\right)^2 + n + 3 - 2(n+2)\frac{a}{s} + (n+1)\left(\frac{a}{s}\right)^2\right)}{\left(1 - \frac{a}{s}\right)^3}.$$

We need to study the sign of the function $f(x) = 2\ln(x) - n(\ln(x))^2 + n + 3 - 2(n+2)x + (n+1)x^2$. We have for $x \geq 0$, $f'(x) = \frac{2(-n\ln(x)+(n+1)x^2-(n+2)x+1)}{x}$. The sign of $f'(x)$ depends on the sign of $n(x) = -n\ln(x) + (n+1)x^2 - (n+2)x + 1$. We have $n'(x) = -\frac{n}{x} + 2(n+1)x - (n+2)$ and $n"(x) = \frac{n}{x^2} + 2(n+1) > 0$. Therefore, $n'(x)$ is increasing in x. Since, $n'(1) = 0$, $n(x)$ has a minimum for $x = 1$. Since $n(1) = 0$, we have $f'(x) \geq 0$. Thus, $f(x)$ is increasing in x. Moreover, $f(1) = 0$, so $f(x) \leq 0$ if and only if $x \leq 1$. This proves that $\tilde{I}_2 \leq \tilde{I}_1$.

Percentile of the wait. For a percentile of the wait, we can make the comparison between $\tilde{I}_1$ and $\tilde{I}_2$, for customers served by agents. We consider the function $g(t) = \mathbb{1}_{t \geq z}$ to penalize the wait when it exceeds z time units. If $z > w$, we have $\tilde{I}_2 = 0$, so it is clear that $\tilde{I}_1 > \tilde{I}_2$. If $z \leq w$, we get

$$\tilde{I}_1 = \frac{e^{-s\mu z}}{s\mu}\sum_{x=0}^{n-1}(\lambda z)^x \sum_{k=0}^{x}\frac{1}{(s\mu z)^k (x-k)!} = \frac{e^{-s\mu z}}{s\mu}\sum_{x=0}^{n-1}\left(\frac{a}{s}\right)^x \sum_{k=0}^{x}\frac{(s\mu z)^k}{k!} = \frac{e^{-s\mu z}}{s\mu}\sum_{x=0}^{n-1}\frac{(\lambda z)^x}{x!}\frac{1-\left(\frac{a}{s}\right)^{n-x}}{1-\frac{a}{s}}, \text{ and}$$
$$\tilde{I}_2 = \frac{e^{-z(s\mu-\lambda)} - e^{-w(s\mu-\lambda)}}{s\mu - \lambda}.$$

This leads to

$$\begin{aligned}\tilde{I}_2 - \tilde{I}_1 &= \frac{e^{-s\mu z}}{s\mu(1-a/s)}\left(e^{\lambda z} - \sum_{x=0}^{n-1}\frac{(\lambda z)^x}{x!} - \left(\frac{a}{s}\right)^n\left(e^{s\mu z} - \sum_{x=0}^{n-1}\frac{(s\mu z)^x}{x!}\right)\right)\\ &= \frac{e^{-s\mu z}}{s\mu}\sum_{x=n}^{\infty}\frac{(s\mu z)^x}{x!}\frac{\left(\frac{a}{s}\right)^x - \left(\frac{a}{s}\right)^n}{1-\frac{a}{s}} \leq 0,\end{aligned}$$

and proves the result. □

C Proof of Proposition 1 and Corollary 1

Proof of Proposition 1.

Proof. We only present the non-trivial case $d \geq 1$. let us introduce $\Delta_x = T_{x+1} - T_x$, for $0 \leq x \leq d-1$. From Equation (4), we get

$$\Delta_x(\theta + \gamma + \beta + t) = \gamma\Delta_{x+1}, \text{ for } 1 \leq x \leq d-2.$$

Thus, Δ_x is geometric. We then deduce that

$$\Delta_x = \left(\frac{\theta + \gamma + \beta + t}{\gamma}\right)^{x-1}\Delta_1,$$

for $1 \leq x \leq d$. Using $T_x - T_1 = \sum_{k=1}^{x-1} \Delta_k$, we obtain

$$T_x - T_1 = (T_2 - T_1)\frac{\gamma}{\theta+\beta+t}\left(\left(\frac{\theta+\gamma+\beta+t}{\gamma}\right)^{x-1} - 1\right).$$

Finally, using the expression of T_1 as function of T_0 leads to

$$T_2 - T_1 = T_0\left[\frac{\gamma+t}{\gamma}\frac{t+\theta+\beta}{\gamma} - \frac{\theta}{\gamma}\right] - \frac{\beta}{\gamma}.$$

We deduce that

$$T_x = T_0\frac{(t+\gamma)\left(\theta + (t+\beta)\left(\frac{\theta+\gamma+\beta+t}{\gamma}\right)^{x-1}\right)}{\gamma(t+\beta+\theta)} - \frac{\beta\left(\left(\frac{\theta+\gamma+\beta+t}{\gamma}\right)^{x-1} - 1\right)}{t+\beta+\theta},$$

for $1 \leq x \leq d$. We also have

$$T_d = \frac{\beta+\gamma}{\theta+\beta+\gamma+t} + \frac{\theta}{\theta+\beta+\gamma+t}T_0.$$

This leads to the expression of T_0. The expression of T_0^D can be obtained in a similar way. □

Proof of Corollary 1.

Proof. We obtain the probability of being disconnected by setting $t = 0$ in the expression of T_0^D. The probability to abandon the system is obtained via $P_A = (1-q)T_0^S|_{t=0}$ and $E(T) = -\frac{\partial T_0}{\partial t}|_{t=0}$. The probability to interrupt a service to reenter later, P, is given by

$$P = \frac{\theta}{\beta+\theta}\left(1 - \left(\frac{\gamma}{\beta+\gamma+\theta}\right)^d\right).$$

Each cycle of service is independent. Therefore, the number of retries for a customer, R, is geometrically distributed with parameter P. We thus deduce the expression of $E(R)$ from $E(R) = \frac{P}{1-P}$. □

D Proof of Proposition 2

Proof. Given disconnection it is clear that $P_D = 1$ and $P_A = 0$. We obtain $E(T|D)$ from $E(T) = -\frac{\partial T_0^D}{\partial t}|_{t=0}\frac{1}{P_D}$. Let us denote by N the number of retries before being disconnected. We have

$$P(N = k \cap D) = \left(\frac{\gamma}{\theta+\beta+\gamma}\right)^d \left(\frac{\theta}{\beta+\theta}\left[1 - \left(\frac{\gamma}{\theta+\beta+\gamma}\right)^d\right]\right)^k.$$

Therefore, we have

$$P(N = k|D) = \frac{\beta + \theta\left(\frac{\gamma}{\theta+\beta+\gamma}\right)^d}{\theta+\beta}\left(\frac{\theta}{\beta+\theta}\left[1 - \left(\frac{\gamma}{\theta+\beta+\gamma}\right)^d\right]\right)^k.$$

Thus, we can deduce $P(N = k|D)$ from $P(N = k|D) = \frac{P(N=k\cap D)}{P_D}$. Finally, $E(R|D) = \sum_{k=0}^{\infty} kP(N = k|D)$.

For non-disconnected customers, we have $P_D = 0$, $P_A = 1 - q$. Again $E(T|\overline{D})$ is obtained via $E(T) = -\frac{\partial T_0^S}{\partial t}|_{t=0}\frac{1}{1-P_D}$, and $E(R|\overline{D})$ is deduced from $E(R) = P_D E(R|D) + (1 - P_D)E(R|\overline{D})$. □