Received: 17 October 2020 | Accepted: 13 May 2022

DOI: 10.1111/poms.13789

ORIGINAL ARTICLE

Toward servitization: Optimal design of uptime-guarantee maintenance contracts

Behzad Hezarkhani[1] | **Mahesh Nagarajan**[2] | **Chunyang Tong**[3]

[1]Brunel Business School, Brunel University London, Uxbridge, UK

[2]Sauder School of Business, University of British Columbia, Vancouver, British Columbia, Canada

[3]School of Economics and Management, Tongji University, Shanghai, China

Correspondence
Behzad Hezarkhani, Brunel Business School, Brunel University, London UB8 3PH, UK.
Email: behzad.hezarkhani@brunel.ac.uk

Handling Editor: Michael Pinedo

Funding information
National Science Foundation of China, Grant/Award Numbers: 72172104, 72021002

Abstract

This paper analyzes the contracting of maintenance services provided by an original equipment manufacturer (OEM) to an operator for a device. The service provider can exert different levels of maintenance effort during the course of the contract and the device's reliability (uptime) is influenced by these levels. However, the service provider's effort level is noncontractible. Our research seeks to find the optimal structures, as well as parameters, of performance-based maintenance contracts. We single out a unique uptime-guarantee contract structure that contains profit-maximizing contracts in many situations. Complete servitization is the essence of such optimal contract structures. With this contract structure, the service provider simply guarantees 100% uptime and compensates the operator's for any occurred downtime at a higher unit rate than it charges for maintenance services. Our findings show that some of the well-known performance-based contract structures used in practice (e.g., pay-for-performance contracts) can be suboptimal for the OEMs. We incorporate the customer's ability to affect the uptime and show that the optimal contract structures can also coordinate the customer's effort. We demonstrate the advantages and limitations of offering menus of contracts to increase the service provider's expected profits. Finally, through simulations using a sample data set, we find that a contract designed using the key ideas in our paper shows very promising results for practitioners.

KEYWORDS

contracting, maintenance, noncooperative game theory, optimization, reliability

1 | INTRODUCTION

Consider a setting in which an operator of some equipment (buyer) contracts with a service provider who agrees to service and maintain the equipment under certain contractual terms. Maintenance contracts can lower the risk of prolonged operational disruptions for the buyer. At the same time, they can generate revenue for service providers and offer excellent opportunities for *servitization*. Not surprisingly, many traditional manufacturing companies, such as Boeing, Philips, Siemens, Xerox, have reinvented their maintenance services as key sources of revenue. For example, at Xerox only about 16% of the annual revenue comes from sales with the remaining 84% coming from contracted services, equipment maintenance, and consumable supplies (Xerox, 2012). In the aerospace sector alone, the maintenance, repair, and overhaul market is worth $62bn, and it is estimated that it will reach $90bn by 2024 (Hollinger & Powley, 2015). The service provider in our context happens to be an original equipment manufacturer (OEM). With capital-intensive and mission-critical devices, for example, MRIs, heavy machinery, and jet engines, it is commonplace for the OEMs to offer service contracts as the expertise needed to preserve the operable conditions of a device is held by its manufacturer. In some cases, such as MRIs and jet engines, the manufacturer monitors the device's performance remotely on a real-time basis.

The interactions between a provider of maintenance services and an operator are often complex in nature. The availability of the operator's device in working condition, which is necessary for his/her revenue generation, is, to a large extent, dependent on the level of effort that the service provider exerts into maintaining it. The level of effort

---

Accepted by Michael Pinedo, after two revisions.

can manifest in several ways. The service provider, for example, can adhere to a schedule of preventive maintenance, keep a sufficient level of spare-parts inventory, set up dedicated service centers, and invest in reactive capacity (manpower, and so on) to deal with disruptions and restore operations very quickly. However, typically, the operator cannot fully verify the service provider's intended effort level or understand its link to operational outcomes. Performance-based contracts that tie the transfer of payments in the contract with the availability-related key performance indicators (KPIs) of the device (such as uptime) can provide self-enforcement mechanisms that elicit effort from the service provider.

In this paper, we focus on the class of uptime-guarantee (UG) maintenance contracts. In this contract structure, for a transactional price the service provider promises an uptime level, that is, a minimum ratio of time that the device will be available over the total workable time during the contract period. If the actual uptime of the device falls short of the promised level, the service provider compensates the operator for the underperformance at a prenegotiated rate. While the promise of uptime conveys the OEM's confidence in device's reliability to the customer, the penalty clause in such contracts implies that the guaranteed uptime level does not have to actually materialize. The devices subject to contract in our context are mission critical, capital-intensive, and highly reliable. UG contracts in these settings commonly proceed an initial warranty period wherein both parties have the chance of calibrating their perception of reliability, effort levels, and costs.

Take the example of *Philips Healthcare*, a leading manufacturer of high-tech medical imaging equipment. The company offers a range of service contracts to its operators to meet their requirements for reliable and high-performance medical devices. In some instances, the company offers a *menu* of maintenance contracts with different guaranteed uptime levels (98%, 99%, and so on) that operators can choose from Philips Healthcare (2014). If the device fails to deliver its guaranteed uptime, the company reimburses a percentage of the contract price to the operator.

While UG contracts are ubiquitous, substantially different guaranteed uptime levels manifest themselves in practice. Table 1 illustrates some of the guaranteed uptime levels offered in practice. Although several companies implement UG contracts with guaranteed uptime levels that are very close, but not equal, to 100%, "OEMs that adopt machines as a service differentiate themselves from competitors by guaranteeing 100% uptime" (Sundblad, 2018). In fact, 100% uptime contracts are what the industry is referring to as servitization (Sundblad, 2018).

The choice of 100% uptime in UG contracts gives rise to some specific performance-based maintenance contracts in practice. The class of *pay-for-performance contracts* works on a fixed cost per operating hour basis. The operator pays the service provider only if the device is in working condition (which is equivalent to guaranteed 100% uptime). Otherwise, there are no payments from the operator to the service provider. The most well-known instance of this contract structure is the *power-by-the-hour* contract launched in 1962 by *Rolls-Royce*, (Rolls-Royce, 2012). As another example, this contract structure is used between Philips and Schiphol airport for the airport's lighting system. The light-as-a-service contract requires the airport to pay only for the light it uses, while Philips, jointly with its partner Cofely, is responsible for the performance and durability of the system and, ultimately, its reuse and recycling at the end of its life (Philips, 2015). The power-by-the-hour model has recently been offered within the cloud computing services as well (e.g., Volta in Table 1).

TABLE 1 Some examples of guaranteed uptime levels offered in practice

Company	Industry	Guaranteed uptime (%)
Rolls Royce	Power generation	100
Volvo Trucks[1]	Auto	100
Toshiba Medical Systems[2]	Health care	100
Hewlett Packard Enterprise[3]	Could computing	100
Volta[4]	Cloud computing	100
US Cloud[5]	Cloud computing	100
Google Cloud[6]	Cloud computing	99.99
Amazon AWS[7]	Cloud computing	99.95
Philips Healthcare[8]	Health care	99
Canon Medical Systems[9]	Health care	98
BHGE[10]	Petrochemical	98
Siemens Gamesa[11]	Wind energy	98

In *downtime compensation contracts*, a service provider receives an up-front payment for the maintenance contract from an operator. If the operator's device encounters any downtime at all, the service provider compensates the operator at a prenegotiated penalty rate. For instance, *Volvo Trucks*, the second largest producer of heavy trucks, offers a 100% UG contract called Gold Contract, which promises operators that whenever their trucks are not in working condition, they will be either reimbursed at predetermined rates or supplied with replacement trucks (Volvo Trucks, 2017). In cloud computing sector, the company US Cloud (see Table 1 and references in the Notes section) also offers 100% UG contracts with "a policy of discounting its monthly subscription rate by 20% for each 3 hours of downtime per month." This large penalty for downtime signals to the customer of the service provider's intention on her effort to maintain the infrastructure and increase availability.

The first goal of this paper is to determine the optimal maintenance contract structure from the perspective of the service provider. In investigating this problem, we seek to obtain some clarity on when each of the many options for maintenance contracts described above should be utilized. That is, what advantage one form may have over another and whether one structure dominates others. We formulate a mathematical problem for obtaining optimal contracts that maximize the service provider's revenue while taking into account the operator's decision regarding the purchase of the contract, as well as his/her anticipation of the service provider's effort level. Despite the possibility of ending up with multiple revenue-equivalent contracts, we show that there is a unique contract structure that contains best single contracts in many relevant situations (even under asymmetric information). This unique contract structure guarantees 100% uptime, irrespective of the actual expected uptime of the device, and compensates the operator for his/her downtime at a unit rate that exceeds the unit rate charged for maintenance services (with a factor equal to the inverse of the device's expected downtime under no contract). This shows that complete servitization of maintenance operations is the best strategy for the service providers. Subsequently, any contract that guarantees an uptime less than 100% may fail to be optimal for the service provider in some situations. This means, for example, the pay-for-performance contracts implemented commonly in the aviation sector may result in the underperformance of the device and suboptimal profits for the service provider. We show that the optimal contract structure can also align the operators' incentives to exert their efforts in looking after the device.

The second goal of the paper is to find the optimal variables to include in a maintenance contract. We do this by first assuming that all information is symmetric. In this case, finding the optimal variables requires only the knowledge regarding the operator's revenue rate from the device. At optimality, the service provider must fully compensate the operator for any lost revenue due to downtime and should charge the operator a per-unit-time price equal to his/her revenue multiplied by the expected downtime under no contract. No other parameter (costs, effort-related reliability, and so forth) are needed to obtain optimal contracts under symmetric information. Subsequently, we relax this assumption by incorporating the fact that the service provider does not know how much value the operator really places on uptime. In fact, the operators have strong incentives to conceal and/or underrepresent their valuations because, otherwise, the service provider can extract all of their surplus. In this case, we provide a closed-form formula which obtains the optimal price and penalty rates. Although the moral hazard in the choice of service provider's effort level can be resolved with 100% UG contract, the asymmetry of information regarding operator's type could render single-optimal contracts inferior to menus of contracts.

The final goal of the paper is to generalize the analysis to contract menus and show the advantages and limitations of using menus instead of single contracts. We find that while the service provider without economies of scale in maintenance costs can benefit from offering multiple contracts through a menu, contract menus in practice are less useful for service providers enjoying economies of scale in their maintenance operations, that is, for these providers, a single-optimal contract has an expected profit that is as large as that of any contract menu.

We then use the main insights from this analysis to examine a real-life setting in which a manufacturer of diagnostic imaging equipment contracts on maintenance terms with a health care system. We show, using their data and simulations, that implementing the class of contracts described above can improve the payoffs for both the service provider and the operator.

The rest of this paper is organized as follows. In Section 2, we review the relevant literature briefly. In Section 3, we formally introduce the basic model and discuss the centralized version of the problem. Section 4 focuses on UG contracts as a generic class of availability-based maintenance contracts and analyzes the contracting game under the assumption of symmetric information. We formulate and solve the service provider's profit maximization problem and discuss the optimal choice of contract structure. Section 5 examines the setting with asymmetric information, that is, when the operator's valuation of revenue during uptime is unknown to the service provider. In Section 6, we expand the model to allow for contract menus. Section 7 tests the main findings of the paper through a numerical example. We then explore some extensions of the basic model in Section 8. In Section 8.1, we generalize the possible effort set of the service provider, and in Section 8.2, we incorporate the customer's effort in maintaining the device and check the robustness of the optimal contract structure in aligning operators' incentives. Section 9 applies our findings to the real-life setting described above. Section 10 presents the main managerial insights and we end with concluding remarks in Section 11. All proofs are provided in the Supporting Information Appendix.

2 | LITERATURE REVIEW

Decision problems involving both operators and service providers have been studied in the extant literature. Under the assumption of complete information, Murthy and Yeung (1995) investigate the optimal strategies for an operator as well as a service provider under two types of maintenance services: preplanned and immediate. They calculate the optimal contract prices based on the operator's known unit revenue rate. Murthy and Asgharizadeh (1999) and Asgharizadeh and Murthy (2000) study the optimal decisions of one or more equipment owners and a service provider in terms of the right choices of contract, contract prices, and service channels—also assuming complete information.

Unlike warranty contracts, which are indicative of the product reliability that is unknown to the operator, UG contracts are often seen in settings in which the operator has experience using the device. Rinsaka and Sandoh (2006) extend the work of Murthy and Asgharizadeh (1999) to include time after the initial warranty period. Within a multistage decision-making framework, Hartman and Laksana (2009) examine the optimal strategies for equipment owners with regard to the types of extended warranty contracts as well as pricing policies. They show that by offering multiple contracts, a service provider can increase his/her profit dramatically. Tong et al. (2014) discuss the pricing strategies for a provider of two-dimensional warranty contracts. Esmaeili et al. (2014) study the choices of various attributes in a three-level warranty service contract among a manufacturer, a service provider and an operator. Gallego et al. (2014)

analyze residual value extended warranty contracts, where an operator receives a bonus if no claims are made during the contract term. They study the pricing problem associated with a single contract and a menu of contracts for strategic operators with different risk attitudes. The contracts studied in this group of papers are not, however, performance based.

Another stream of research underlines the fact that interacting parties may not be fully aware of each other's attributes. Taking into account the random nature of an operator's attitude concerning risk, Huber and Spinler (2012) formulate a model that allows a service provider to manage his/her revenue by setting the prices for full-service and on-call maintenance contracts. They give a closed-form formula for the service provider's optimal contract prices under the assumption that the operator's attitude toward risk is uniformly distributed. Another approach to managing the relationship between operators and service providers seeks to coordinate the parties' efforts to optimize the system-wide profit—as opposed to focusing on either operators' or service providers' decision-making problems. Within a deterministic framework, Tarakci et al. (2006a, 2006b) analyze several mechanisms, including a pricing scheme for maintenance contracts, to ensure that the optimal intervals for preventive maintenance from the perspectives of both service provider and equipment owner coincide. Tseng and Yeh (2013) extend the single processor case discussed in Tarakci et al. (2006a) for risk-averse operators. Some authors draw upon demand elasticity to study the effects of different contracts on a service provider's profit. So and Song (1998) construct a model to address the pricing, delivery time guarantee, and capacity expansion decisions of a service provider. Lieckens et al. (2015) construct a multinomial logit model to approximate the operators' choice among full-service, on-call, or noract options. The reaction of the operator is not directly incorporated in these models.

In an alternative setting in which the operators design the maintenance contracts, Jain et al. (2013) study the double moral hazard problem between the operator (who designs the maintenance contract) and the service provider. There are two types of contracts investigated: in linear contracts, there is a uniform penalty applied for downtime; in tiered contracts, different penalty rates are applied for different levels of downtime. The authors examine the gap between the realized outcome under a given contract and that under the first-best outcome. They assume convex increasing costs and that the operator's valuation is known. In UG contracts, however, it is the service provider who designs and offers the contracts so the problem from the point of view of the service provider is not analyzed here.

Kim et al. (2007) study an operator's maintenance outsourcing contract design problem with suppliers who provide repair and spare-part services. The maintenance contract comprises a fixed fee, a ratio determining the operator's share of the service provider's repair costs, and a penalty for backorders incurred by the supplier. The supplier's decision regarding his/her inventory level is private action. The cost associated with the service provider's effort is common knowledge. They show the benefits of performance-based contracts compared to cost-plus and fixed-price contracts when aligning the supplier's incentives. Kim et al. (2010) introduce contractual structures that mitigate the moral hazard service providers face that are associated with their capacity investments and highlight the challenges due to the infrequent nature of device failures. As in the case of our paper, they consider the case where the suppliers' effort is noncontractible. Zeng and Dror (2015) extend the analysis of this problem in several directions. Unlike our model, these models assume that operator's valuation, that is, revenue rate, is common knowledge. Moreover, the structure of the contracts is not investigated.

Bakshi et al. (2015) develop a game theoretic model to study the interactions between the service provider and operators in maintenance contracts. The service provider has private information about the reliability of the device—which is either high or low. The uptime of the device is tied directly to the amount of inventory kept by the service provider, which has a probability distribution function which satisfies the increasing hazard rate property. They study both resource-based as well as performance-based contracts. The resource-based contract is characterized by a fixed fee and a ratio determining the service provider's share of the repair costs. The performance-based contract is characterized by a fixed fee and a penalty rate for units of the device's downtime. The cost of the contract is associated with product availability as well as the level of inventory kept by the service provider. The authors study the signaling problem that arises between the parties. Assuming that the service provider knows the operator's valuation, they show that when the inventory levels are not verifiable, which resembles our case of noncontractible effort levels, the performance-based contract puts more focus on reliability and overinvests in inventory. The initial reliability in our situation is known to both parties as our contracts usually follow the initial warranty period wherein the operator has the chance to experience firsthand the reliability of the device. In this paper, we relax the assumption that the service provider knows the operator's valuation and study the structures of the performance-based contract that could achieve optimality in all situations.

Li et al. (2016) construct a model for a situation in which the operator is better informed about product reliability and the supplier offers performance-based or transaction-based maintenance contracts. Similar to the majority of the papers reviewed so far, this model assumes that the operator's valuation as well as the supplier's cost structure are common knowledge. They characterize the settings that result in the operator or the supplier preferring either of the contract types.

In an empirical study of maintenance service plans offered by a manufacturer of capital-intensive medical imaging equipment, Chan et al. (2014) highlight the possible misalignment of incentives in full-protection versus pay-per-service plans. The full-protection plan includes the preventive maintenance services offered in a pay-per-service plan plus the charges for labor and material. As they report, suboptimal full-protection plans may increase the costs of both manufacturer and operator, implying that the intended benefits of innovative maintenance contracts could not be achieved without fine-tuning the parameters.

In a related paper, Hezarkhani (2017) incorporates the asymmetry of information regarding the operator's valuation to solve the revenue maximization problem for a service provider offering UG contracts. However, the corresponding model excludes the incentive issues between the contracting parties and assumes that the service provider's effort level is contactable, always delivered, and that the device's uptime is deterministic with regard to the effort. Thus, this work disregards the most important aspect of UG contracts, that is, the mechanisms that assure the operator about service provider effort after the purchase of contract. Our paper explicitly considers the possible incentive misalignment between the parties and the operator's adverse-selection, due to private information, as well as the service provider's moral hazard, due to noncontractibility of maintenance effort. Still, we find that with the right choice of contract structure, the incentives can be aligned to the extent that the service provider can achieve the maximum revenue afforded by the optimal contract.

3 MODEL

The focal point of the problem is a device owned by an operator for which an OEM provides maintenance services. We assume both parties are risk-neutral. We refer to the service provider and the operator with the pronouns he and she, respectively. The device's reliability is imperfect, and the ratio of the workable time that it will be available during a period of time, that is, its uptime $x \in [0,1]$, is a random variable. Let F be the distribution of the device's uptime during a given period of time. We fix the length of the contract period to one unit of time. The uptime distribution is a function of the maintenance effort of the service provider. The set of all possible levels of effort that the service provider can exert is denoted by $E \subset \mathbb{R}$. For the bulk of the paper, we assume $E = \{0, L, H\}$, with $0 < L < H$. This choice of effort set succinctly conveys the important aspects of our problem. The service provider can choose not to exert any maintenance effort, and when he does exert effort, he can do so at either low or high levels. We discuss the setting with more choices of effort levels in Section 8.1. Let F_e be the distribution of the uptime under the service provider's effort level $e \in E$. The expected uptime under this choice of effort level is $\mathbb{E}_e[x] = \mu_e \in [0,1]$. We assume, naturally, that μ_e is increasing on E. The expected status quo for the uptime, that is, the running time without any effort from the service provider, is μ_0. Note that μ_0 is effectively the expected uptime with no contract.

For every unit of working time that the device is available, the operator can generate revenue. We denote the average revenue rate of the operator by $r \geq 0$ and normalize operational costs to 0. The expected revenue of the operator under the choice of effort level $e \in E$ by the service provider is thus $r\mu_e$.

The cost of maintenance services to the provider is represented with the function c. We denote the service provider's cost under the choice of effort level $e \in E$ by c_e. We assume c is nondecreasing on E. That is, as the service provider increases his level of effort, his associated costs never decrease. We further normalize $c_0 = 0$. A maintenance service *situation* is represented by the tuple $\Gamma = (E, (F_e)_{e\in E}, (c_e)_{e\in E}, r)$. The set of all situations are denoted by $\boldsymbol{\Gamma}$. In the remainder of the paper, unless explicitly stated, we assume that the parameters of the situation are fixed.

3.1 Benchmark: Centralized system

If the service provider exerts effort level e, the total expected value generated in the system would be $r\mu_e - c_e$. With no interactions between the parties, the total expected value generated in the system would be $r\mu_0$. Hence, the extra value generated as the result of interactions between the service provider and the operator under effort level $e \in E$ is $U_e = r(\mu_e - \mu_0) - c_e$. We assume that contracting of maintenance services between the two parties in aggregate creates positive value. That is, $\max\{U_L, U_H\} > 0$, so the extra revenue generated as the result of the service provider exerting his best effort is positive overall.

It is useful to distinguish between two possibilities with regard to maintenance service situations. We say a situation Γ is of type I if $\frac{c_L}{\mu_L-\mu_0} \geq \frac{c_H}{\mu_H-\mu_0}$. Alternatively, a situation Γ is of type II if $\frac{c_L}{\mu_L-\mu_0} < \frac{c_H}{\mu_H-\mu_0}$. The set of all type I and II situations are denoted by $\boldsymbol{\Gamma}^I$ and $\boldsymbol{\Gamma}^{II}$, respectively. Note that with $E = \{0, L, H\}$, we have $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}^I \cup \boldsymbol{\Gamma}^{II}$. In $\boldsymbol{\Gamma}^I$ situations, the uptime-adjusted cost of effort is nonincreasing, implying that it would take progressively less investment to increase the expected uptime of the device. In $\boldsymbol{\Gamma}^{II}$ situations, the uptime-adjusted cost of effort is increasing so it is relatively more costly to increase the uptime from low to high effort compared with increasing it from zero to low effort. Thus, it would take progressively greater investments to increase the expected uptime of the device.

Let e^C be the first-best effort level such that $U_{e^C} \geq U_e$ for all $e \in E$. The first lemma in this paper characterizes e^C in maintenance service situations.

Lemma 1. *For every type I situation,* $\Gamma \in \boldsymbol{\Gamma}^I$*, we have* $e^C = H$*. For a type II situation,* $\Gamma \in \boldsymbol{\Gamma}^{II}$*, if* $\frac{c_H-c_L}{\mu_H-\mu_L} > r$ *then* $e^C = L$*, and if* $\frac{c_H-c_L}{\mu_H-\mu_L} \leq r$ *then* $e^C = H$*.*

In $\boldsymbol{\Gamma}^I$, it is always best that the service provider exerts a high effort level, as the additional uptime justifies the cost. The service provider's choice of effort in $\Gamma \in \boldsymbol{\Gamma}^{II}$, however, is determined by the magnitude of operator's revenue r. As

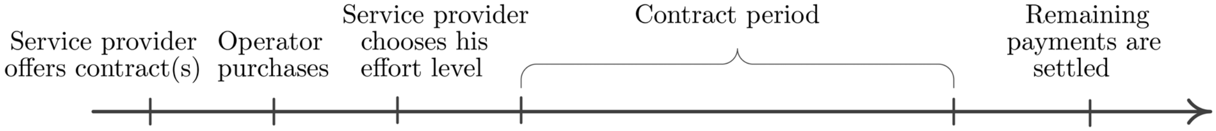

FIGURE 1 Sequence of events

a result, both low and high effort levels can be the first bests depending on whether r is low or high, respectively.

4 | DECENTRALIZED SYSTEM AND CONTRACTING OF MAINTENANCE SERVICES

In decentralized systems, the parties are self-interested. The service provider offers the contract, exerts the maintenance effort, and incurs the associated costs. The revenue due to the availability of the device is obtained by the operator.

We construct a game played between the service provider and the operator to model the strategic interactions in the maintenance relationship. The sequence of events is depicted in Figure 1. In this game, the service provider's choice in offering a contract is followed by the operator's response in terms of purchasing the contract or rejecting it. Since the effort is not contractible, the service provider still can decide the extent of his maintenance effort after the contract is offered and purchased.

Considering the moral hazard on the service provider's side when it comes to exerting his maintenance effort, the only way that the operator can be convinced of the service provider intention to exert effort is with a performance-based contract that conditions the payments on the actual uptime of the device. Thus, we focus our attention on performance-based contracts. In this context, a performance-based contract can be specified by the payments from the operator to the service provider under any realized device's uptime at the end of contracting period.

Due to the risk-neutrality of the parties, one can encapsulate the dynamics of a contract by transfer payments, which are calculated as the expected value of some function $h(x)$ of realized uptime which, in turn, is a stochastic function of the service provider's effort. We define the expected transfer payment of the contract under the choice of effort level $e \in E$ with $y_e = \mathbb{E}_e[h(x)]$. A (performance-based) contract is, accordingly, defined as the collection of transfers under all possible choices of effort, that is, $Y = (y_e)_{e \in E}$.

The expected profits of the operator and service provider under the contract Y, given the choice of effort $e \in E$ by the latter, are, respectively, $\pi_e(Y) = r(\mu_e - \mu_0) - y_e$, and $\Pi_e(Y) = y_e - c_e$.

The corresponding sequential contracting game is illustrated in Figure 2. There are three stages. In the first stage, the service provider offers the contract. In the second stage, the operator decides whether to purchase it or not. Finally, in the third stage, the service provider decides what his level of effort will be. We analyze the sequential contracting game using the concept of subgame perfect equilibrium (SPE) and backward induction.

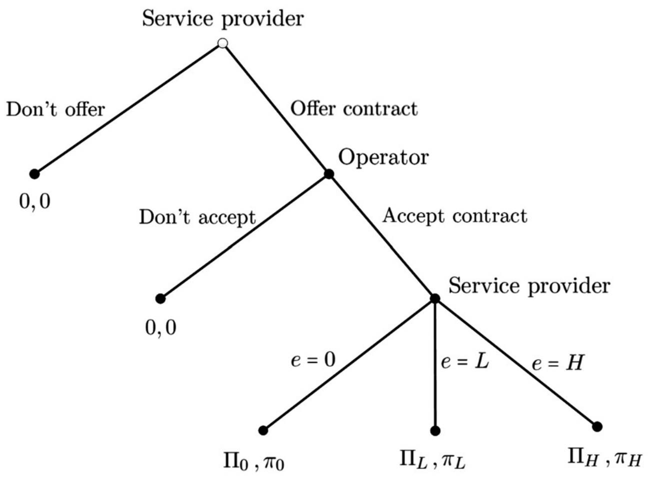

FIGURE 2 Contracting game tree

Given a contract Y, in the third stage of the game the service provider evaluates his expected profit under different effort levels. We say that the contract Y induces $e \in E$ if $\Pi_e(Y) \geq \Pi_{e'}(Y)$ for every $e' \in E$. The induction condition in this definition captures the ex post optimality of service provider's effort and is necessary to provide assurance to the operator that the effort level e is at least as good as other possible effort levels for the service provider. This condition is independent of r (the operator's type), so the intended effort level can be inferred before purchasing the contract.

In our setting, the induced effort level of a contract in the third stage can be inferred by the operator. Hence, in the second stage of the game the operator can make her decision regarding the acceptance of the contract under the associated induced effort level of the contract. Contract Y is individually rational for the operator under effort level e if $\pi_e(Y) \geq 0$. This corresponds to the standard participation constraint.

In the first stage, the service provider can determine if offering a contract which induces effort level e and is acceptable to the operator is profitable to him at all. Contract Y is individually rational for the service provider under effort level e if $\Pi_e(Y) \geq 0$.

4.1 | Symmetric information

Assume that the operator types, manifested by her unit revenue rate r, are common knowledge. To find the most profitable contract that induces effort level $e \in E$, the service provider must maximize his expected profit under the conditions that the contract is acceptable to the operator, it induces effort level e, and that his expected profit satisfies individual rationality. When these conditions hold, we have the strategy profile (offer contract, accept contract, e) as an SPE. In order to find the optimal contract, the service provider's maximum expected profits under all choices of effort level must be compared. Accordingly, the service provider's optimization problem can be compacted into the following:

$$\max_{(y_e)_{e\in E}} \; y_{e^*} - c_{e^*} \tag{1}$$

$$s.t. \quad y_{e^*} - c_{e^*} \geq y_e - c_e \quad \forall e \in E \tag{2}$$

$$y_{e^*} \leq r(\mu_{e^*} - \mu_0). \tag{3}$$

Note that in the program above the objective function gives the expected profit of the contract under the service provider's optimal effort. Also, $\Pi_{e^*}(Y^*) = y_{e^*} - c_{e^*}$ is the maximum expected profit to the service provider.

Under symmetric information, a maintenance contract Y implements the first best if $e^* = e^C$ and $y_{e^C} = r(\mu_{e^C} - \mu_0)$. Thus, the first requirement for implementation of the first best is for the service provider to extract all extra profit from the operator at optimality, which is in line with the predication of standard principal-agent models (Laffont & Martimort, 2002). The second requirement is for the contract to induce the first-best effort level that is $y_{e^C} - c_{e^C} \geq y_e - c_e$ for all $e \in E$. We discuss the latter condition in the next section.

4.2 | Uptime-guarantee and related contract structures

In this paper, we focus on the class of UG maintenance contracts, which includes several well-known performance-based contract structures as special cases. As we show later in the paper, this choice is without loss of generality.

A UG contract is characterized by triplet (u, p, v). The parameter $u \in [0, 1]$ is the uptime guaranteed by the service provider. The price of the contract is $p \in \mathbb{R}^+$ which is the amount that the operator pays to the service provider if she accepts the contract. The contract's penalty $v \in \mathbb{R}^+$ is how much the service provider pays back to the operator for every unit of realized under-performance with regard to u. The transfer payment of a UG contract (u, p, v) is thus $h(x) = p - v\max\{0, u - x\}$, where $\max\{0, u - x\}$ is the under-performance of the device with regard to the guaranteed uptime level u. The expected transfer payment of a UG contract (u, p, v) under effort level $e \in E$ is

$$y_e = \mathbb{E}_e[h(x)] = p - v\lambda_e^u, \tag{4}$$

where $\lambda_e^u = \mathbb{E}_e[\max\{0, u - x\}]$ is the expected underperformance of the contract when the service provider exerts effort e. In case of continuous distribution of uptime, we have $\lambda_e^u = \int_0^u F_e(x)dx$. A desirable feature of UG contracts is that the effort inducibility of a contract is independent of the operator's type. That is, a UG contract (u, p, v) induces effort level e if

$$v \geq \frac{c_e - c_{e'}}{\lambda_{e'}^u - \lambda_e^u} \qquad \forall e' \in E. \tag{5}$$

Hence, the magnitude of the penalty signals the service provider's intended effort level. Some special examples of UG contracts which are used widely in practice are given next.

4.2.1 100% *UG* (*downtime compensation*) *contract*

This contract structure can be represented with the triplet $(1, p, v)$ that corresponds to a UG contract with $u = 1$. In this case, we have $\lambda_e^1 = 1 - \mu_e$ so the expected underperformance is same as the expected downtime under the choice of effort level e. Thus, these UG contracts fully compensate any downtime experienced by the customer. The expected transfer payment in this contract structure is $y_e = p - v(1 - \mu_e)$. Volvo Truck's *Gold Service* contract (Volvo Trucks, 2017) is an example of these contracts.

4.2.2 *Pay-for-performance contract*

This contract structure can be represented with the triplet $(1, p, p)$. The operator pays p for every unit of time that the device has been available. When the device is down, there are no payments. Drawing upon a single parameter p, the expected transfer of this contract structure is $y_e = p\mu_e$. These contracts simplify the *power-by-the-hour* maintenance contracts offered by Rolls-Royce for servicing jet engines (Rolls-Royce, 2012).

Not all structures could implement the first best. In fact, well-known structures such as pay-for-performance contracts may fail to implement the first best. To verify this, note that with the expected transfer $y_e = p\mu_e$, one can choose $p^* = r(1 - \mu_0/\mu_{e^*})$ in a pay-for-performance contract to obtain the maximum transfer $y_{e^*} = r(\mu_{e^*} - \mu_0)$. Suppose $e^C = L$ and recall that this requires $r < (c_H - c_L)/(\mu_H - \mu_L)$. The L-inducibility constraint in this case requires $p^* \leq (c_H - c_L)/(\mu_H - \mu_L)$, which for the contract $(1, r(1 - \mu_0/\mu_L), r(1 - \mu_0/\mu_L))$ yields $r(\mu_L - \mu_0) - c_L \geq r(1 - \mu_0/\mu_L)\mu_H - c_H$. One can check that for any r such that $(c_H - c_L)/(\mu_H - \mu_L)(1 - \mu_0/\mu_L) < r < (c_H - c_L)/(\mu_H - \mu_L)$, the constraint is violated implying that for this range of operator's valuation low effort level cannot be induced and the first best cannot be attained.

Despite the plethora of possible contract structures, the main result of this section introduces a unique universally optimal contract under symmetric information.

Proposition 1. *Under complete information, the contract* $(1, r(1 - \mu_0), r)$ *is optimal in every situation* $\Gamma \in \mathbf{\Gamma}$. *Furthermore, it is the unique UG contract that implements the first best and achieves optimal revenue in every situation.*

Based on the above, the service provider can always maximize his expected profit using a UG contract so focusing on UG contracts in this context is without loss of generality. However, there is only one contract that can always be optimal. We next discuss the features of this optimal contract.

The universally optimal contract introduced in Proposition 1 guarantees 100% uptime. The problem with UG contracts that use any other uptime level is that one can always find situations in which the first-best effort level of the service provider is $e^C = L$, but a contract with $u < 1$ would be unable to induce $e^* = L$ at optimality. As the proof of Proposition 1 exhibits, for some distributions of device reliability, the feasible interval for a UG contract's penalty to induce low effort level would be empty when the guaranteed uptime falls below 100%. This will never be a problem for 100% UG contracts. Therefore, 100% UG contracts can tackle the moral hazard of the service provider with regard to signaling the optimal choice of effort levels in every situation.

Furthermore, in order to ensure that a UG contract can always satisfy the optimality conditions the unit penalty must coincide with the operator's valuation. If the first-best effort level of the service provider is $e^C = L$, then, for any 100% uptime contract $(1, p, v)$ with $v > r$, there exists a situation wherein the low effort level cannot be induced. Similarly, if the first-best effort level of the service provider is $e^C = H$, then for any 100% uptime contract $(1, p, v)$ with $v < r$, there exists a situation wherein the high effort level cannot be induced. Therefore, only one choice of universally optimal contracts remains, that is, $v = r$. Hence, the universally optimal contract under symmetric information has $v = r$, that is, it compensates the operator fully for her downtime.

The universally optimal contract is equivalent to a downtime-compensation contract structure. These contracts are the simplest to design and implement. The choice of $u = 1$ makes the profit functions of the parties dependent only on the expected uptime; thus, the operator's knowledge regarding the distribution of the device's uptime under different effort levels (i.e., reliability function) becomes irrelevant to the analysis. Hence, this contract is rather parsimonious with regard to its required level of commonly known information. Another appealing feature of this class of contracts is that the parties do not have to settle all payments at the end of the contract period. The reason is that with $u < 1$ the total workable time during the contract period must be fully realized before the actual ratio of uptime to workable time can be calculated. But, with $u = 1$ penalties associated with any downtime during workable times can be settled immediately. We discuss this issue further in Section 10.

5 | ASYMMETRIC INFORMATION

As seen in the case with complete information, the optimal contracts draws upon the operator's valuation r to implement the first best. In this section, we consider the case in which the service provider does not know the operator's type r. Since the service provider would extract all the additional revenue of the operator via an optimal contract under symmetric information, the operator has incentives to understate her valuation—this is the case in standard adverse-selection model (Laffont & Martimort, 2002). We assume that the operator's valuation is distributed in the range $[0, r_{max}]$ according to the distribution function G and probability function g. The distribution of operator's types is common knowledge. To ensure that there is a chance for maintenance contracts to be beneficial, we assume $r_{\max}$ is sufficiently large, that is, we assume $\max\{r_{\max}(\mu_L - \mu_0) - c_L, r_{\max}(\mu_H - \mu_0) - c_H\} > 0$.

Under asymmetric information, the service provider cannot verify ex ante the participation of the operator. The service provider's problem to maximize his expected revenue, when offering a single contract, is[12]

$$\max_{(y_e)_{e\in E}} \quad (y_{e^*} - c_{e^*})\bar{G}\left(\frac{y_{e^*}}{\mu_{e^*} - \mu_0}\right) \tag{6}$$

$$s.t. \quad y_{e^*} - c_{e^*} \geq y_e - c_e \qquad \forall e \in E \tag{7}$$

$$y_{e^*} \leq r_{\max}(\mu_{e^*} - \mu_0). \tag{8}$$

In the above formulation, the term $\bar{G}(\frac{y_{e^*}}{\mu_{e^*} - \mu_0})$ is the probability that contract Y, inducing effort level e^*, being individually rational to the operator. Accordingly, the program above maximizes the expected profit of the contract while satisfying the conditions that the choice of effort results in the highest expected profit among different choices of effort (6), and that the chosen effort level can be induced (7). The constraint (8) preserves the feasibility of the contract with regard to the highest possible revenue rate.

Note that although there can be many best single contracts—with different parameters—they all have equal expected profits. Thus, in the asymmetric valuation case, the degrees of freedom in choosing the three parameters of the contract discussed in the previous section exist as well. The main result in this section singles out a unique contract structure that always contains the best single contracts in the asymmetric information case.

Proposition 2. *For every situation* $\Gamma \in \mathbf{\Gamma}$*, there exists a UG contract of the form* $(1, p, \frac{p}{1-\mu_0})$ *that has the highest expected profit among all possible single contracts. Furthermore,* $(1, p, \frac{p}{1-\mu_0})$ *is the unique UG contract structure that contains a best single contract in all situations.*

The first outcome of the above result is in line with the observation in the previous section that among all possible contract structures, a 100% uptime contract is versatile enough to contain best single contracts in asymmetric information case. Also, the contract structure found in Proposition 2 always contains the best single contract in asymmetric information case. The expected transfer payment of this contract structure under the choice of effort level $e \in E$ is $y_e = p(\mu_e - \mu_0)/(1 - \mu_0)$.

In the contract structure introduced in Proposition 2, the ratio of price to penalty is the expected status quo downtime of the device. Hence, in order to achieve optimality, the service provider must set the penalty at a higher rate than the unit price—with a factor equal to the inverse of the status quo expected downtime. The ratio of penalty to price grows exponentially as the status quo uptime of the device increases. Subsequently, for high-availability devices, the penalty rate in the optimal contract structure can be considerably larger than the price rate. In practice, the service provider charges the customer a lump sum equal to the unit price rate multiplied by the entire contracting period (say, a year), and then announces a penalty rate for downtime at much smaller unit of time (say, an hour). This is apparent in the excerpt from the US Cloud SLA example mentioned in the introduction, that is, 20% off of monthly subscription fee for each three hours of downtime per month. If expressed in terms of equal unit time rates, this penalty amounts to 50 times of transaction price (720 working hours per month for a data center). Within our optimal contract structure, this ratio of penalty to price corresponds to a device with status quo expected uptime of 98%.

In the asymmetric information case, the first-best results may not be attained. The reason is twofold. First, it might be the case that the optimal contract deters a customer (with low valuation) from buying the contract even though in the symmetric information case there is an opportunity for generating value through a maintenance contract for that customer. This issue is the well-known distortion from the bottom in mechanism design games (Laffont & Martimort, 2002). One can verify that with a contract (u, p, v) that induces e, any customer with valuation $r < \frac{p-v\lambda_e^u}{\mu_e-\mu_0}$ would not buy the contract. However, if $\frac{c_e}{\mu_e-\mu_0} < r$ we have $U_e > 0$. The following remark illustrates the degree of distortion from bottom for the optimal contract structure.

Remark 1. For every situation $\Gamma \in \boldsymbol{\Gamma}$ and given a UG contract with the structure $(1, p, \frac{p}{1-\mu_0})$, the customer buys the contract if and only if $r \geq \frac{p}{1-\mu_0}$.

Second, an optimal contract may not induce the first-best optimal effort level. As we establish in the proof, the latter does not happen in $\Gamma \in \boldsymbol{\Gamma}^{I}$, that is, optimal contracts in these situations always induce a high effort level. But, in $\Gamma \in \boldsymbol{\Gamma}^{II}$, an optimal contract may induce the low effort level while the first-best effort level is high.

In order to obtain a closed-form solution for the optimal price, and subsequently, optimal expected transfer payment, we assume that the operator's type distribution G is an increasing generalized failure rate (IGFR) distribution. A probability distribution satisfies the IGFR condition if $tg(t)/\bar{G}(t)$ is nondecreasing everywhere in its support such that $G(t) < 1$ (Lariviere, 2006). The family of IGFR distributions includes uniform, exponential, normal, log-normal, logistic, Weibull, gamma, beta, and Cauchy distributions in addition to others as well as their one- or two-sided truncations (Banciu & Mirchandani, 2013).

Proposition 3. *The optimal transfer of a single contract under effort level e, that is, y_e^*, is attained via*

$$\frac{\bar{G}\left(\frac{y_e^*}{\mu_e-\mu_0}\right)}{y_e^*-c_e} = \frac{g\left(\frac{y_e^*}{\mu_e-\mu_0}\right)}{\mu_e-\mu_0}. \tag{9}$$

The closed-form solution in the last proposition resembles the optimal solution to the service provider's revenue maximization problem in Hezarkhani (2017) in absence of incentive problems. This, in conjunction with Proposition 2, highlights the power of our proposed contract structure to screen out any potential incentive misalignment among the parties and achieves the best possible outcomes for the service provider.

6 | OPTIMAL CONTRACTS

In previous sections, we considered the scenario wherein the service provider offers a single contract to the operator. However, when the operator's type is unknown, the service provider might be able to increase his expected profit by offering a menu of contracts that can stratify the operator's type space and extract more profits in the case where the operator has a high valuation while still ensuring that a low-valuation operator will purchase a contract. In this section, we address the optimal design of contract menus.

A contract menu $\{Y^1, \ldots, Y^m\}$ is a collection of single contracts. The number of contracts included in a menu determines its size. When offered a contract menu, the operator evaluates her options and chooses the contract which yields the highest expected profit to her. As our next lemma shows, with noncontractible effort levels the service provider can differentiate his contract offerings up to a maximum number.

Lemma 2. *The choice of the contract menu of size two $\{Y^1, Y^2\}$, without loss of generality, maximizes the expected profit of the service provider in every situation.*

The two contracts in the menu above must induce low and high effort levels. If optimal contract menus of the latter form exist, then they would be optimal among all other possible contract menus of any size as well. Subsequently, the service provider cannot accrue additional benefits from contract menus of size three or more.

In order to construct the optimization program corresponding to the aforementioned contract menu, we first need to analyze the acceptance regions for the single contracts included in it.

Lemma 3. *A necessary condition for a contract menu of size two $\{Y^1, Y^2\}$, inducing low and high effort levels, respectively, to increase the service provider's expected profit compared to the case with single contracts is to have $y_L^1(\mu_H - \mu_0) \geq y_H^2(\mu_L - \mu_0)$.*

Let $\{Y^1, Y^2\}$ be a contract menu with the single contracts inducing L and H effort levels, respectively. Whenever a contract menu of size 2 satisfies the necessary condition expressed in Lemma 3, the contract menu would split the operator's type space into three regions. For the operator with $r < \frac{y_L^1}{\mu_L - \mu_0}$, none of the contracts in the menu satisfies the individual rationality condition; thus, the operator does not purchase a contract. Whenever we have $\frac{y_L^1}{\mu_L-\mu_0} \leq r \leq \frac{y_H^2 - y_L^1}{\mu_L - \mu_0}$, the operator will prefer contract (u^1, p^1, v^1) that also satisfies the individual rationality condition, so it would be her choice of contract. Finally, for $\frac{y_H^2 - y_L^1}{\mu_L-\mu_0} \leq r \leq r_{\max}$, the operator prefers the individually rational contract Y^2. To design a contract menu of size 2 optimally, the service provider needs to solve the following program:

$$\mathbb{E}\left[\Pi^*_{\{L,H\}}\right]$$

$$= \max_{y^1, y^2} \left(y_L^1 - c_L\right)\left[G\left(\frac{y_H^2 - y_L^1}{\mu_H - \mu_L}\right) - G\left(\frac{y_L^1}{\mu_L - \mu_0}\right)\right]$$

$$+ \left(y_H^2 - c_H\right)\bar{G}\left(\frac{y_H^2 - y_L^1}{\mu_H - \mu_L}\right) \tag{10}$$

$$s.t. \quad y_L^1(\mu_H - \mu_0) \geq y_H^2(\mu_L - \mu_0) \tag{11}$$

$$y_H^2 \leq r_{\max}(\mu_H - \mu_0) \tag{12}$$

$$y_L^1 \leq r_{\max}(\mu_L - \mu_0) \tag{13}$$

$$y_H^1 - y_L^1 \leq c_H - c_L \tag{14}$$

$$y_L^1 - y_0^1 \geq c_L \tag{15}$$

$$y_H^2 - y_L^2 \geq c_H - c_L \tag{16}$$

$$y_H^2 - y_0^2 \geq c_H. \tag{17}$$

The objective function of the program above contains the profits obtained from the two contracts multiplied by their corresponding acceptance probabilities. Constraint (11) reflects the condition in Lemma 3. Constraints (12) and (13) guarantee that the single contracts in the menu have positive chances of being purchased by the operator. Finally, constraints (14)–(15) and (16)–(17) are necessary effort-compatibility conditions to ensure that contracts Y^1 and Y^2 induce effort levels L and H, respectively. To determine the best option, the service provider must compare the expected profits of optimal single contract as well as the optimal contract menu. Thus, the service provider's optimal choice satisfies $\mathbb{E}[\Pi^*] = \max\{\mathbb{E}[\Pi^*_{\{L,H\}}], \mathbb{E}[\Pi_{e^*}(Y^*)]\}$.

The main result of this section is contained in the following proposition.

Proposition 4. *The following statements obtain the optimal contracts in different situations:*

- *For a situation* $\Gamma \in \mathbf{\Gamma}^I$*, no contracts menu can improve upon the best single contract as described in Proposition 2.*
- *For a situation* $\Gamma \in \mathbf{\Gamma}^{II}$*, the optimal contract is a menu, with the transfer payment of the low effort inducing contract satisfying:*

$$\frac{\bar{G}\left(\frac{y_L^{1*}}{\mu_L - \mu_0}\right)}{y_L^{1*} - c_L} = \frac{g\left(\frac{y_L^{1*}}{\mu_L - \mu_0}\right)}{\mu_L - \mu_0}, \tag{18}$$

and the transfer payment of high effort inducing contract satisfying:

$$\frac{\bar{G}\left(\frac{y_H^{2*} - y_L^{1*}}{\mu_H - \mu_L}\right)}{y_H^{2*} - y_L^{1*} - c_H + c_L} = \frac{g\left(\frac{y_H^{2*} - y_L^{1*}}{\mu_H - \mu_L}\right)}{\mu_H - \mu_L}. \tag{19}$$

Both contracts can be found within UG contracts with the structure $(1, p, v)$*.*

As revealed in Proposition 4, the type of situation has a critical effect on the optimality of single contracts versus menus. In $\mathbf{\Gamma}^{II}$, the optimal contract menu can be obtained from optimizing the unconstrained version of the program in (10). The optimality condition for the transfer payment of the low-effort contract is same as that for a single contract. For obtaining the best H-inducing contract, the condition in (19) finds the optimal point for the transfer payments at which the operator's preference in choosing among the contracts in the menu switches. In $\mathbf{\Gamma}^I$, the optimal solution is rather different. As it turns out, the solution to the relaxed problem is infeasible because it violates the condition in constraint (11). Thus, it can be inferred that at optimality, the latter constraint is binding, meaning that for the optimal contract menu, the switching point of the operator's preference over the two contracts is the same as the point at which the L-inducing contract becomes acceptable. But if this is the case, then the probability of operator choosing the low effort level becomes zero. Therefore, in $\mathbf{\Gamma}^I$, contract menus are always dominated by single contracts inducing a high effort level. As a result, the service provider cannot benefit from offering multiple contracts within a menu in order to increase his expected profits. The outcome of Proposition 4 extends the observation in the previous sections that among all possible contract structures, focusing on 100% UG contracts is without loss of generality as one can always find optimal contracts there.

7 | NUMERICAL EXAMPLES

In this section, we elaborate on the results of previous sections via some numerical examples.

Consider the situation with $\mu_0 = 0.8$, $\mu_L = 0.85$, and $\mu_H = 0.9$, $c_L = 1$, $c_H = 6$, and assume $r = 150$. This is a

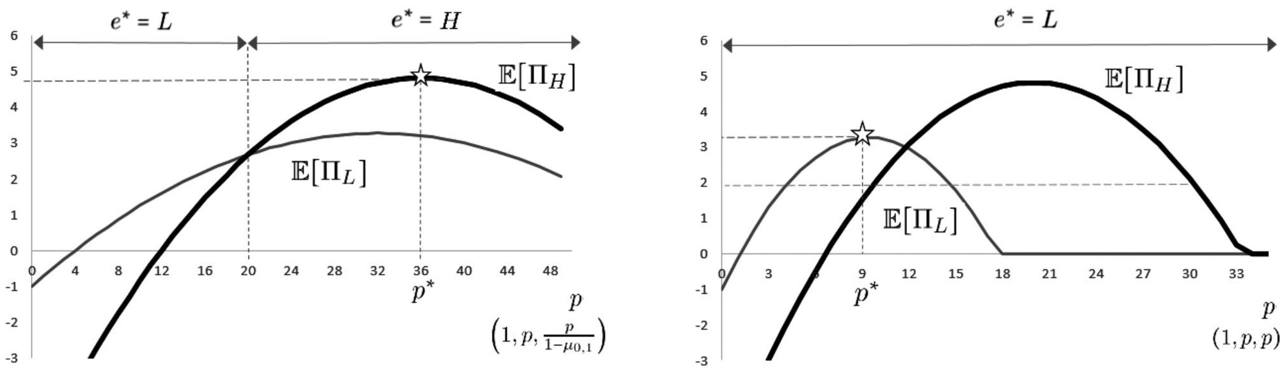

FIGURE 3 Expected profits and induces effort levels in different contract structures (left: $(1, p, \frac{p}{1-\mu_0})$, right: $(1, p, p)$)

situation in $\Gamma \in \boldsymbol{\Gamma}^{II}$. In the centralized system, we are comparing $U_L = 7.5 - 1 = 6.5$ and $U_H = 15 - 6 = 9$. So $e^C = H$ and $U_{e^C} = 9$. The optimal contract under symmetric information, according to Proposition 1, is (1,30,150)—this can be interpreted as a \$30 per hour maintenance contract, which compensates the operator for any downtime at \$150 per hour penalty rate. The contract induces the first-best (high) effort level and has an expected profit of \$9 for the service provider.

Under asymmetric information, suppose the operator's revenue rate is distributed uniformly between 0 and 300. With uniform distribution, the equation for obtaining the optimal transfer payment of the contract that induces e can be written as $y_e^* = \frac{1}{2}(r_{\max}(\mu_e - \mu_0) + c_e)$. The optimal contract induces a high effort level, which results in the expected profit of $\mathbb{E}[\Pi_H^*] = 4.80$. The optimal contract is (1,36,180).

In this example, optimality cannot be attained using a pay-for-performance contract structure. In order to induce effort level $e^* = H$, the contract's penalty must be high enough to signal the service provider's high effort; in this case it must satisfy $v \geq 100$. Figure 3 demonstrates the expected profits under effort levels H and L as functions of the contract price under two contract structures: the optimal structure $(1, p, \frac{p}{1-\mu_0})$ (left) and the pay-for-performance $(1, p, p)$ (right). These figures also indicate intervals during which either H or L is induced. As can be seen in the left-hand side of Figure 3, the service provider can increase p to maximize his expected profit under effort level H. With the pay-for-performance contract structure (right), the outcome is strikingly different. Although with price $p = 20$, that is, the contract (1,20,20), the same expected profit of 4.80 is attained assuming that the effort level is fixed at H, this contract induces a low effort level (as mentioned in Section 4.2, in order to induce high effort level, a 100% UG contract must satisfy $v \geq (c_H - c_L)/(\mu_H - \mu_L) = 100$ thus a pay-for-performance contract with $p = v = 20$ would induce the low effort level). Thus, the service provider cannot attain the optimal expected profit. In this case, the best possible outcome is to choose contract (1,9,9), which obtains the maximum expected profit of $\mathbb{E}[\Pi_L^*] = 3.26$. Therefore, our proposed contract structure outperforms the pay-for-performance contracts since the latter structure fails to induce the first-best effort level in this situation.

Next, we examine the use of contract menus. As this is a type II situation, by Proposition 4, the use of contract menus can increase the service provider's expected profit. The optimal low effort inducing contract in this case is (1,14,40). The optimal high effort inducing contract in this case is (1,36,180). The contract menu consisting of these two contracts will generate the expected profit of $\mathbb{E}[\Pi_{L,H}^*] = 4.93$. From the results above, the service provider with a single contract can obtain, at best, the expected profit of $\mathbb{E}[\Pi_H^*] = 4.8$ (which is obtained by contract (1,36,180)). Therefore, the optimal contract menu can increase the service provider's expected profit.

8 | EXTENSIONS

8.1 | General effort sets

In our setting so far, the service provider can choose from a set of three distinct effort levels. We generalize this setting by allowing effort levels to take any value from the set $E = \{0, 1, 2, \ldots, H\}$, $H \in \mathbb{N}^+$, $|E| \geq 3$, with larger integers representing higher levels of effort. It can be assumed naturally that higher effort levels are progressively more costly, that is, $c_0 < c_1 < \cdots < c_H$, and that they progressively increase the expected uptime, that is, $\mu_0 < \mu_1 < \cdots < \mu_H$. The optimal centralized effort levels with general effort levels can be obtained by finding $e^C \in E$ such that

$$\max_{e<e^C} \frac{c_{e^C} - c_e}{\mu_{e^C} - \mu_e} \leq r \leq \min_{e>e^C} \frac{c_e - c_{e^C}}{\mu_e - \mu_{e^C}}. \tag{20}$$

The above requires searching the space of $|E|^2$ possibilities. In the case of symmetric information, the results of our previous analysis can be extended to the case with general effort sets.

Proposition 5. *Contract $(1, r(1-\mu_0), r)$ is the unique UG contract that implements the first best under complete information for every situation under extended effort sets.*

In order to generalize our results to the case of asymmetric information regarding the optimal contract structure to the situations with extended effort sets, we need additional conditions. These conditions are extensions of type I and II situations, that is, $\Gamma \in \mathbf{\Gamma}^{I}$ if $c_e/(\mu_e - \mu_0)$ is nonincreasing on e, and $\Gamma \in \mathbf{\Gamma}^{II}$ if $c_e/(\mu_e - \mu_0)$ is increasing on e. Although the latter two classes of situations do not cover all possible situations with general effort sets, unless $|E| = 3$, they cover a wide range of situation with reasonable structures in their cost and effort functions. For example, the class of $\mathbf{\Gamma}^{II}$ situations contain any situation with the values of c_e and μ_e being taken from convex and concave functions, respectively, which implies that the cost of effort and the expected uptime grow at increasing and decreasing rates, respectively (see Lemma 4 in the Supporting Information). The optimal centralized effort levels with general effort levels is always $e^C = H$ in situations $\Gamma \in \mathbf{\Gamma}^{I}$ and can be any effort level, depending on the value of r, in situations $\Gamma \in \mathbf{\Gamma}^{II}$.

Under asymmetric information, the key result of the paper regarding optimal contract structures holds for type I and II situations under extended effort sets as well.

Proposition 6. *With the general effort set, for every situation* $\Gamma \in \mathbf{\Gamma}^{I} \cup \mathbf{\Gamma}^{II}$ *there exists an optimal contract of the form* $(1, p, \frac{p}{1-\mu_0})$. *Furthermore, the contract structure* $(1, p, \frac{p}{1-\mu_0})$ *is the unique structure that contains an optimal contract in all type I and II situations.*

Note that the above results highlight the contract structure $(1, p, \frac{p}{1-\mu_0})$ among all possible maintenance contracts and not just the class of UG contracts. With regard to menus of contracts, the result of Lemma 2 is also extendable for general effort levels. That is, contract menus containing as many individual contracts as the number of nonzero elements of E is without loss of generality. Although the use of contract menus for situations in $\mathbf{\Gamma}^{I}$ remains fruitless, for other situations one can use the menus of contracts successfully to increase the service provider's expected profit.

The challenge to extend our results to situations in $\mathbf{\Gamma} \setminus (\mathbf{\Gamma}^{I} \cup \mathbf{\Gamma}^{II})$ is the inability to induce potentially optimal effort levels using UG contracts in all situations. But this also implies that there might not be a contract structure that can always contain optimal contracts in general situations. Nevertheless, the use of 100% UG contracts with the highlighted structure in such situations can be a good practical proxy while the structure of optimal contracts in general situations remains an open question.

8.2 | Impact of customer's effort

In our models so far, the customer has had no impact on the uptime of the device. However, in most real-life instances the customer, as the user of the device, can influence the uptime as well as the cost of maintenance. In this section, we include this important aspect into our analysis.

Suppose the customer can exert effort for proper utilization and taking care of the device. We denote the customer's effort with $a \in [0, 1]$ with 0 and 1 representing the minimum and maximum levels of effort that she can exert, respectively. We extend the definition of cost and uptime functions to include customer's effort. The maintenance cost to the provider can be expressed as a function of both his as well as the customer's effort levels. We denote this function with $c_{e,a}$. We already established that exerting higher effort levels by the service provider corresponds to higher costs, that is, $c_{e,a}$ is nondecreasing on e. Exerting higher effort levels by the customer has no significant cost for her, however, it would result in lower costs for the vendor. This is because the service provider is fully responsible for maintenance costs and the customer's effort in taking care of the device reduces his maintenance costs, that is $c_{e,a}$ is nonincreasing on a. The expected uptime of the device under effort levels e and a is denoted with $\mu_{e,a}$. To reflect the fact that higher levels of both efforts do not decrease the expected uptime, we assume that $\mu_{e,a}$ is nondecreasing on a as it is on e. Furthermore, we assume that the service provider's effort has more impact on the device's uptime than that of the customer, that is, we let $\mu_{L,0} > \mu_{0,1}$. With no interaction between the parties, $\mu_{0,1}$ would be the expected uptime of the device. This serves as the base performance of the system.

The expected profits to the customer and the vendor if the contract is offered, accepted and parties exert effort levels a and e are, respectively:

$$\begin{aligned}\pi_{e,a}(u, p, v) &= r(\mu_{e,a} - \mu_{0,1}) - p + v\lambda^{u}_{e,a},\\ \Pi_{e,a}(u, p, v) &= p - v\lambda^{u}_{e,a} - c_{e,a}.\end{aligned} \tag{21}$$

A preliminary observation in this case is that in the centralized mode, the first-best effort level of the customer is to exert her maximum effort.

Remark 2. In the first-best solution for every situation $\Gamma \in \mathbf{\Gamma}$, it holds that $a^* = 1$.

This remark follows from the fact that expected uptime is nondecreasing and cost is nonincreasing on device user's effort level. For the service provider, the first-best effort level depends, as before, on the relationship between the μ and c. We then define the notion of effort-coordinated contracts.

Definition 1. A contract (u, p, v) is called *effort-coordinated* if for every situation $\Gamma \in \mathbf{\Gamma}$, whenever the customer accepts the contract, the choice of $a = 1$ be the best strategy for her irrespective of the vendor's effort level.

The definition of effort-coordinated contract mandates that exerting high effort be a weakly dominant strategy for the customer in case she chooses to accept it. The appeal of effort-coordinated contracts is that once they are in place, one can take the customer's utmost effort for granted and exclude it from the analysis of the problem. The main result of this section illustrates the coordinating ability of our optimal contracts.

Proposition 7. *Every contract with the structure* $(1, p, \frac{p}{1-\mu_{0,1}})$ *is effort-coordinated.*

The above result clarifies yet another advantage of our suggested UG contract structures. That is, the threat of customer's malicious intent to hinder the performance of the device in order to take advantage of downtime penalties as an alternative source of revenue is irrelevant. Therefore, with this contract structure in place, any customer whose valuation for the device's uptime r is low enough to make the downtime penalty v more attractive than the revenue from the uptime, that is, $r < v$, in effect will not buy the contract in the first place.

9 | SIMULATION ON IMPLEMENTATION

We explore how our findings can be applied to a real-life setting. To do so, we use data from a large supplier of equipment for diagnostic services to a large health care network that comprises public hospitals and private diagnostic centers. The setting is as follows. The supplier contracts with an entity that purchases medical-imaging equipment (computed tomography [CT] and magnetic resonance imaging [MRI] machines). The contracts include a service provision that involves maintenance of the equipment and contingencies in the event of a breakdown or disruption due to equipment failure. To keep things simple, we consider only the two parties, the supplier (service provider) and the buyer (operator). In our setting, the above contracts govern 32 CT and 26 MRI machines that conducted approximately 600,000 scans for fiscal year 2017–2018.

9.1 | Parameter estimation

We first consider the operator's parameters. The first task is to fit the values and distribution of r. There is a substantial variance in the type of jobs a machine may perform as well as the variance in the machines themselves. For example, the MRI machines in our study were mostly used for musculoskeletal, neurological and oncological exams. We note that in our theoretical model, although we call r the revenue rate obtained by operating the equipment per unit time, in reality, equipment has a significant operating cost per unit of time. To ensure that the data fit with the spirit of our model, a measure of "profit" per unit of operating time is needed. To calculate this, we looked at the distribution of the modalities and procedures of specific machines that were used historically and calculated an average pay rate. Next, we compute the operating cost per unit of time for a piece of equipment. We considered only the variable costs involved in operation per unit time. The main components of the variable cost were labor costs (radiologists, technicians, and nurses, among others). We calculated the weighted average (across procedures, time, and location) cost per unit of time per equipment. Having obtained these financial inflows and outflows, we assign the contribution margin of a piece of equipment per unit of time as the value of r, the revenue rate for the equipment. Recall that facilities have a mix of equipment and, as discussed, equipment varies in values of r, providing us with a natural range for r and a distribution of its values. It is realistic to assume that individual operators know their value of r and that the service provider knows only the distribution of these values.

We next turn our attention to estimations of the reliability of the machines and the impact of efforts expended by the service provider. Typically, diagnostic imaging machines that are new purchases are very reliable at the outset of their operations. This reliability deteriorates after a few years. The following are typical issues, observed from our data, which caused equipment downtime: (i) chiller issues that affect the MRI but are separate from the MRI; (ii) recalibration due to power disruption; (iii) unclear images; and (iv) malfunctioning cold-head on the scanner. These four reasons covered more than 95% of the significant downtime incidents. To compute the base rate of uptime, that is, μ_0, we need to calibrate what effort actually looks like. Service providers offer contracts that provide various levels of preventative maintenance (PM). For example, a PM contract may check the heating, ventilation, and air conditioning (HVAC) unit and the chiller periodically. Another example features PM done on the cryocooler during nonoperating hours. Other examples of effort include positioning of spare coil loaner inventory, which is useful to keep the equipment running when a coil fails. We assume that μ_0 is the uptime of a machine when there is no PM contract in place but rather only a standard maintenance contract that does not compensate for downtime. We note that μ_0 varies by machine and age and location. Calculating μ_e was more challenging. We grouped types of effort into discrete buckets. More precisely, we examined PM contracts, their schedules and specifications. PM contracts that were frequent and involved detailed and critical functions were grouped separately from routine PM contracts. We also looked at maintenance contracts and noted variations in the effort by the service provider depending on the location and the contract. For example, certain large hospitals in urban centers were offered more flexible schedules. They had inventory positioned close by so that the impact of downtime was decreased. We note that these costs were given to us once we proposed the buckets. Next, we performed a simple linear regression with the dependent variable being the average uptime and effort being one of the discrete independent variables. This approach allowed us to ballpark the impact of effort on the uptime and compute μ_e. Costs were normalized and standard maintenance contracts are assumed to represent zero effort level. A standard PM contract was deemed as a low effort level, that is, $e = L$, and a contract that combined a PM with a higher level of scheduling flexibility as well as a local positioning of spare parts was deemed as a high effort level, that is, $e = H$. In terms of effort costs, a PM contract's effort level, which required prebooked crew to examine and maintain machines, costs two to seven times more on average per machine compared to effort level of zero. The highest level of effort level that was deemed to be available for critical centers with advanced diagnostic capabilities incurred costs 2–11 times the cost of no effort.

Although we are interested only in the averages for the purpose of calibration, we also computed the distribution of time between failures, that is, the downtime distribution. We need this to compare our contracts, and will discuss this aspect in more detail later.

9.2 | Evaluation of uptime contract

Now that we have all the parameters, albeit approximately, for our theoretical model, we explore a consistent approach to study the impact of what would have happened if the operators and service providers had used an "optimal" uptime contract rather than the status quo maintenance contract. Needless to say, the ideal scenario would have been to do a field study where we implemented the uptime contract and observed the effects in a carefully controlled experiment. Such an experiment, as can be imagined, is not trivial to accomplish. Instead, our approach was as follows. Using the parameters from our calibration, we searched for a contract of the form $(1, p, p/(1-\mu_0))$ that is both incentive compatible and individually rational. That is, the service provider chooses the optimal effort, and the contract is individually rational to both players in the sense that the reservation profit is the contribution margin minus the contract fee for the operator, and the contract fee minus contract cost incurred due to downtime for the service provider with the existing contract. A few observations are in order. Note that although we are searching for the above contract form, we know it is not optimal unless we verify the explicit type under which the situation falls in. The question we are asking is: If we pick an operator at random, will this contract perform better, on average, than the status quo? To answer this question, we cast it as an optimization problem in p, as in (6)–(8) of the paper, and solve by enumeration. We check if the contract is optimal by verifying which type the situation falls in and adjust the optimality conditions. We find that this optimization problem has a solution (and our procedure generates the optimal contract), which, by definition of the (IR) constraints, implies that if we had implemented an uptime contract as per our calculation, on an average, operators and service providers would have been better off than they were historically in our data set. We note that not all operators will benefit, and, indeed, some will not choose the contract.

With the optimal values in hand, the optimal contract terms are as follows. In the event of a failure, the service provider will compensate the operator at the rate of $p/(1-\mu_0)$ for the duration of the downtime. We know the optimal effort bucket. We also know the distribution of the downtime from the data.

TABLE 2 Simulation results

r Value	Machines (%)	Increase in operator's payoff (average) (%)	Increase in service provider's payoff (average) (%)
0–0.397	18	0	0
0.397–0.483	23	1.86	7.81
0.483–0.69	28	2.5	11.6
0.69–1	31	2.7	14.4

We then did two comparisons. We compared, based on simulated downtimes and the optimal contract in place, what the payoffs to both players are. We then compared the payoffs again by simulating downtimes and their performance in the existing maintenance contract regime.

These simulations provide greater confidence in our predictions and are summarized in Table 2, where the r values are normalized within $(0, 1]$. The increase in the operator's payoff is computed as simply the payoff with the uptime contract that is the contribution margin at the contract fee minus the payoff with maintenance contract (status quo) divided by the payoff of the status quo and a similar calculation for the service provider.

As one would expect, for low values of r, the contract is not attractive (not IR) for the operator. This finding is consistent with the findings in the earlier sections. The operator and service provider's improvements are positive and monotonic. We also note that the above information does not give a sense of the size of the increases in the payoffs. Although specific contribution margins are proprietary information, the above increases are significant. The magnitudes of the increases are significant due to a few facts. First, compensating for downtime entails a stiffer contracting fee compared to the status quo, but it also decreases downtime. In our data, although downtime is not frequent, the results are skewed by incidences of severe downtime at a few different high r-valued locations. Clearly, if the overall downtime is low to begin with, the increases will also be modest.

With a larger data set, it may be possible to calculate more precise improvements and contract terms based on the exact profile of the equipment. Our approach here is rather approximate, as we have grouped together a varied set of operators, equipment, and contracts. However, the above analysis does provide a glimpse of how benefits can accrue by moving to an optimal uptime contract.

10 | DISCUSSION

We have shown that if UG contracts are designed appropriately, they can assure the operator about service provider's effort to increase the device's uptime. The crucial lever for assuring the operator of the service provider's effort is the penalty that the latter pays the former when the device fails to provide the promised uptime level. Our analysis of SPEs in this paper shows that the operator can always anticipate the service provider's optimal choice of effort level, which would be selected only after the contract is accepted.

An important factor in the analysis is whether the effort level of the service provider can be contracted. The possibility of formalizing the service provider's effort level within the contract makes it easier for the parties to speculate on their actions and make the best decisions. However, with noncontractible effort levels, which is consistent with most real-life situations, new challenges surface. In this case, the lack of clarity on the service provider's effort level might render the purchase of a contract too fuzzy a decision for the operator. We showed, nevertheless, that, under certain conditions, appropriate contracts exist. The appropriate contracts not only must consider the profit-maximizing criteria for the service provider but they should also signal the service provider's intention to exert effort clearly. With this clarity in place, the operator can make her purchasing decision with confidence. Such clarity can be achieved with a single UG contract structure that promises 100% uptime to the operator and reimburses downtimes at the rate equal to the contract payment inflated by the factor of $\frac{1}{1-\mu_0}$. When the operator's valuation for uptime is known, our best single-contract charges the operator an amount equal to her valuation discounted by the factor $1 - \mu_0$ and reimburses her entire valuation in case of downtime. In order to illustrate a more realistic picture regarding the interactions between the parties, we allow the service provider to have incomplete information about the operator's type.

Our analysis sheds light on some of the contracts implemented in practice and enables us to comment on their optimality. The first example is the UG contracts utilized by Philips Healthcare for servicing medical imaging devices such as MRI machines. The analysis reveals that the use of 100% UG contracts could be as good as any other uptime (e.g., 95% or 99%) that the company offers currently. In fact, it is the combination of the promised uptime level, price, and the penalty that determines the service provider's profit. In the case of Volvo Trucks, the Gold Service contracts have the optimal structure that we obtained in our analysis. However, when monitory payments to operators are employed, it is possible to encounter the problem that the operator's valuation may not be accurately known to the service provider. Still, when an alternative device (truck in this example) is offered to the operator, she can recoup almost all her entire lost revenue so the lack of knowledge about her type can be circumvented. The last example is the pay-for-performance service contracts used by Rolls-Royce. We showed that the penalty rate in this contract structure might be too low and may not signal the higher effort levels of the operator, resulting in a suboptimal contract.

We conclude this section by discussing the subtle difference between performance-based contracts that promise uptime versus availability. Although the UG contracts prove to be effective means to increase the availability of the device, the availability is not an appropriate KPI to be guaranteed.

To understand why this may be the case, we need to have a better understanding of workable and nonworkable times during the contracted period. As shown in Figure 4, the total time during the contracted period can be seen as workable or nonworkable. For example, normal business hours during weekdays are normal workable hours, and the rest can be considered as nonworkable hours. Although these "normal" workable/nonworkable hours can be foreseen in advance, the actual workable/nonworkable hours may be quite different. For example, suppose the device is down before an unforeseen event, such as a fire or a strike. Such unpredictable events turn normal workable hours into nonworkable hours and vice versa. In this case, the nonperational hours of the device would not be counted toward its promised uptime. However, changes in workable/nonworkable hours might be significant enough to make promises on availability detrimental to the service provider, who may end up paying huge penalties to the operator for reasons beyond his control or irrelevant to the device's reliability. The UG and availability guarantee would be the same if the total workable time during the contracted period could have been completely predictable by both service provider and operator. But this predictability would likely be an assumption that is hard to justify. In our model, it is only assumed that the expected workable time during the contract period is known and accepted by the parties. Nevertheless, the service provider must be able to verify downtimes as well as nonworkable times. Such verification is needed, as the service provider pays the operator whenever the machine is down during the workable time (no payments during nonworkable times). The discussion above reveals an advantage of pay-for-performance contracts. Since, in this contract structure, the service provider pays nothing to the operator during the nonworkable time and whenever the machine is down, the parties do not need to monitor or reach consensus regarding the nonworkable times. Therefore, disagreements concerning this value are irrelevant to the parties and have no bearing on payments and profits.

11 FINAL REMARKS

In this paper, we study maintenance contracts that use uptime as their performance criterion. Broadly speaking, maintenance service contracts fall into two main categories: resource-based contracts and performance-based contracts. While the resource-based contracts guarantee the supply of parts or a number of annual maintenance visits within a negotiated time frame, performance-based contracts that tie directly to specific KPIs of the device under contract (e.g., availability, reliability, restoration time, and so on). We assume that the relationship between the required investments

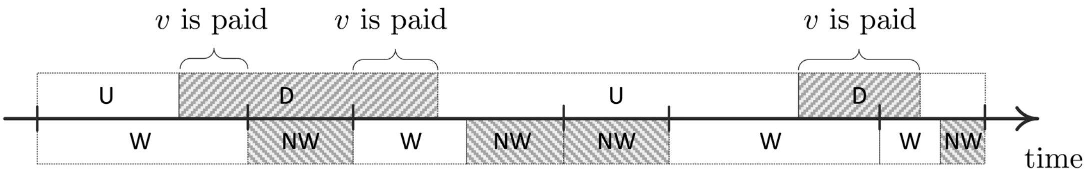

FIGURE 4 Workable and nonworkable times versus device's uptime (W, workable time; NW, nonworkable time; U, device is up; D, device is down)

of the service provider and associated uptime distributions are common knowledge. The basis for this assumption is that the UG contracts usually proceed the standard warranty period so the operator has already been observing the device's performance and is somewhat clear about its reliability. This assumption is rather common in the literature (see, e.g., Bakshi et al., 2015; Kim et al., 2007; Li et al., 2016, among others).

Our stylized model has several limitations. This setting naturally begs the question of what the contract structure would look like if the service provider were to be risk-averse. In this case, we hypothesize that the information set needed to arrive at an optimal contract would be richer than what we have shown. Examining general cost structures, assuming costs are private information and so forth are other natural extensions. These are left for future research.

ACKNOWLEDGMENTS

The authors are grateful to the associate editor and three anonymous referees for suggestions that substantially improved the paper. C. Tong was supported by the National Science Foundation of China (grants 72172104 and 72021002).

ORCID

Behzad Hezarkhani https://orcid.org/0000-0003-3439-3474

ENDNOTES

[1] https://www.volvotrucks.co.uk/en-gb/trucks/volvo-fh/uptime.html.

[2] https://uk.medical.canon/wpent/uploads/sites/8/2015/08/c215baea3e.pdf.

[3] https://www.hpe.com/psnow/doc/a00074521enw.

[4] https://www.voltadatacentres.com/pay-as-you-go-data-centre/.

[5] https://www.uscloud.com/microsoft-unified-enterprise-support/.

[6] https://cloud.google.com/compute/sla?hl=en.

[7] https://aws.amazon.com/compute/sla/.

[8] https://www.philips.co.uk/healthcare/services/maintenance-services/rightfit-service-agreements.

[9] https://uk.medical.canon/service-and-support/service-and-supportservice-and-supportmvs/.

[10] https://www.bakerhughes.com/ts/total-lifecycle-services/longterm-service-agreement-ltsa.

[11] https://www.siemensgamesa.com/en-int/products-and-services/service-wind/multibrand.

[12] Given that the operator knows her type privately, this problem can be alternatively formulated using the standard adverse-selection model, the service provider offers a screening contract (Bolton & Dewatripont, 2005) and maximizes his expected profit subject to the operator's incentive compatibility as well as his own individual rationality constraints—accompanied by the *e*-inducibility constraint.

REFERENCES

Asgharizadeh, E., & Murthy, D. N. P. (2000). Service contracts: A stochastic model. *Mathematical and Computer Modelling*, *31*(10), 11–20.

Bakshi, N., Kim, S. H., & Savva, N. (2015). Signaling new product reliability with after-sales service contracts. *Management Science*, *61*(8), 1812–1829.

Banciu, M., & Mirchandani, P. (2013). Technical note–new results concerning probability distributions with increasing generalized failure rates. *Operations Research*, *61*(4), 925–931.

Bolton, P., & Dewatripont, M. (2005). *Contract theory*. The MIT Press.

Chan, T., De Vericourt, F., & Besbes, O. (2018). Contracting in medical equipment maintenance services: An empirical investigation. *Management Science*, *65*. https://doi.org/10.1287/mnsc.2017.2993

Esmaeili, M., Shamsi Gamchi, N., & Asgharizadeh, E. (2014). Three-level warranty service contract among manufacturer, agent and customer: A game-theoretical approach. *European Journal of Operational Research*, *239*(1), 177–186.

Gallego, G., Wang, R., Hu, M., Ward, J., & Beltran, J. L. (2014). No claim? Your gain: Design of residual value extended warranties under risk aversion and strategic claim behavior. *Manufacturing & Service Operations Management*, *17*(1), 87–100.

Hartman, J. C., & Laksana, K. (2009). Designing and pricing menus of extended warranty contracts. *Naval Research Logistics*, *56*(3), 199–214.

Hezarkhani, B. (2017). Optimal design of UG contracts under IGFR valuations and convex costs. *European Journal of Operational Research*, *256*(2), 556–566.

Hollinger, P., & Powley, T. (2015). Lucrative aircraft maintenance market scrutinised. *Financial Times*, October 14, 2015.

Huber, S., & Spinler, S. (2012). Pricing of full-service repair contracts. *European Journal of Operational Research*, *222*(1), 113–121.

Jain, N., Hasija, S., & Popescu, D. G. (2013). Optimal contracts for outsourcing of repair and restoration services. *Operations Research*, *61*(6), 1295–1311.

Kim, S.-H., Cohen, M. A., & Netessine, S. (2007). Performance contracting in after-sales service supply chains. *Management Science*, *53*(12), 1843–1858.

Kim, S.-H., Cohen, M. A., Netessine, S., & Veeraraghavan, S. (2010). Contracting for infrequent restoration and recovery of mission-critical systems. *Management Science*, *56*(9), 1551–1567.

Laffont, J.-J., & Martimort, D. (2002). *The theory of incentives: The principal-agent model*. Princeton University Press.

Lariviere, M. A. (2006). A note on probability distributions with increasing generalized failure rates. *Operations Research*, *54*(3), 602–604.

Li, D., Mishra, N., & Netessine, S. (2016). *Contracting for product support under information asymmetry* (Working Paper). https://doi.org/10.2139/ssrn.2881356

Lieckens, K. T., Colen, P. J., & Lambrecht, M. R. (2015). Network and contract optimization for maintenance services with remanufacturing. *Computers & Operations Research*, *54*, 232–244.

Murthy, D. N. P., & Asgharizadeh, E. (1999). Optimal decision making in a maintenance service operation. *European Journal of Operational Research*, *116*(2), 259–273.

Murthy, D. N. P., & Yeung, V. (1995). Modelling and analysis of maintenance service contracts. *Mathematical and Computer Modelling*, *22*(10), 219–225.

Philips. (2015). *Philips provides light as a service to Schiphol airport*. https://www.philips.com/a-w/about/news/archive/standard/news/press/2015/20150416-Philips-provides-Light-as-a-Service-to-Schiphol-Airport.html

Philips Healthcare. (2014). *Service agreement terms and conditions*. http://images.philips.com/is/content/PhilipsConsumer/Campaigns/CA20151021_Teamsite_Decommision/HC-terms-and-conditions/CA20160208_CO_001-en_US-for-Service-2014.pdf

Rinsaka, K., & Sandoh, H. (2006). A stochastic model on an additional warranty service contract. *Computers & Mathematics with Applications*, *51*(2), 179–188.

Rolls-Royce. (2012). *Rolls-Royce celebrates 50th anniversary of power-by-the-hour*. https://www.rolls-royce.com/media/press-releases/yr-2012/121030-the-hour.aspx

So, K. C., & Song, J.-S. (1998). Price, delivery time guarantees and capacity selection. *European Journal of operational research*, *111*(1), 28–49.

Sundblad, W. (2018). "Machine as a service": Industry 4.0 powers OEM aftermarket revenue growth. *Forbes.com*, August 13.

Tarakci, H., Tang, K., Moskowitz, H., & Plante, R. (2006a). Incentive maintenance outsourcing contracts for channel coordination and improvement. *IIE Transactions*, *38*(8), 671–684.

Tarakci, H., Tang, K., Moskowitz, H., & Plante, R. (2006b). Maintenance outsourcing of a multi-process manufacturing system with multiple contractors. *IIE Transactions*, *38*(1), 67–78.

Tong, P., Liu, Z., Men, F., & Cao, L. (2014). Designing and pricing of two-dimensional extended warranty contracts based on usage rate. *International Journal of Production Research*, *52*(21), 6362–6380.

Tseng, F., & Yeh, Y. (2013). Maintenance outsourcing coordination with risk-averse contractors. *Journal of the Operational Research Society*, *65*(11), 1760–1769.

Volvo Trucks. (2017). *Volvo gold contract*. http://www.volvotrucks.co.uk/en-gb/serviceandparts/genuinevolvoservice/volvoservicecontracts/volvogoldcontract.html

Xerox. (2012). *Annual report*. Xerox Corporation.

Zeng, S., & Dror, M. (2015). *Formulating principal-agent service contracts for a revenue generating unit*. Springer.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Hezarkhani, B., Nagarajan, M., & Tong, C. (2022). Toward servitization: Optimal design of uptime-guarantee maintenance contracts. *Production and Operations Management*, *31,* 3806–3822. https://doi.org/10.1111/poms.13789