Semantic Information Marketing in the Metaverse: A Learning-Based Contract Theory Framework

Ismail Lotfi, *Member, IEEE*, Dusit Niyato, *Fellow, IEEE*, Sumei Sun, *Fellow, IEEE*, Dong In Kim, *Fellow, IEEE*, and Xuemin Shen

***Abstract*— In this paper, we address the problem of designing incentive mechanisms by a virtual service provider (VSP) to hire sensing IoT devices to sell their sensing data to help creating and rendering the digital copy of the physical world in the Metaverse. Due to the limited bandwidth, we propose to use semantic extraction algorithms to reduce the delivered data by the sensing IoT devices. Nevertheless, mechanisms to hire sensing IoT devices to share their data with the VSP and then deliver the constructed digital twin to the Metaverse users are vulnerable to adverse selection problem. The adverse selection problem, which is caused by information asymmetry between the system entities, becomes harder to solve when the private information of the different entities are multi-dimensional. We propose a novel iterative contract design and use a new variant of multi-agent reinforcement learning (MARL) to solve the modelled multi-dimensional contract problem. To demonstrate the effectiveness of our algorithm, we conduct extensive simulations and measure several key performance metrics of the contract for the Metaverse. Our results show that our designed iterative contract is able to incentivize the participants to interact truthfully, which maximizes the profit of the VSP with minimal individual rationality (IR) and incentive compatibility (IC) violation rates. Furthermore, the proposed learning-based iterative contract framework has limited access to the private information of the participants, which is to the best of our knowledge, the first of its kind in addressing the problem of adverse selection in incentive mechanisms.**

***Index Terms*— Digital twin, semantic communication, contract theory, age of information, deep reinforcement learning.**

I. Introduction

DRIVEN by the Covid-19 pandemic, the Metaverse has gained huge interest recently from different industry and public sectors [1], [2]. Considered as the next generation of the Internet, the Metaverse enables users and objects to experience near real-life interaction with each other in the virtual environment through their avatars. The Metaverse is made up from different emerging technologies such as virtual reality (VR), augmented reality (AR) and haptic sensors. Furthermore, other emerging technologies such as beyond 5G and 6G are driving the Metaverse from imagination and fiction towards real world implementation as they enable users to access the Metaverse from anywhere, anytime instantly.

The first step towards realizing and exploiting the Metaverse is the replication of the physical objects into their respective digital twins. As the digital twins are required to replicate the physical real-world system to the finest details [2], generating an accurate 3D model of the physical system and constant update of the physical system in digital twin is the first step towards this goal. However, the creation of an accurate 3D digital copy is challenging for several reasons. First, in the upstream layer, i.e., between the VSP and the sensing IoT devices, the collected data by the sensing IoT devices is huge in size and the available bandwidth for data transmission will quickly exceed the system limitation. In addition, the delivered data by the sensing IoT devices needs to be delivered timely and should not be outdated. Second, in the downstream layer, i.e., between the VSP and the Metaverse users, to support a real-time interaction between the Metaverse users and the physical world, the rendered digital twin by the VSP needs to be delivered timely and with an acceptable quality to the Metaverse users. Therefore, to enable a real-time construction and delivery of the digital twin in the Metaverse, the communication system needs to be carefully designed so as to maximize successful data transmission with high data value while minimizing the latency of packet delivery.

However, it is challenging for the VSP to find a balance between the revenue, i.e., profit, and the cost in both layers. In the upstream layer, the VSP needs to find an optimal

Manuscript received 15 March 2023; revised 1 August 2023; accepted 31 August 2023. Date of publication 21 December 2023; date of current version 1 March 2024. This work was supported in part by the National Research Foundation, Singapore; in part by Infocomm Media Development Authority under its Future Communications Research and Development Program; in part by the Defence Science Organisation (DSO) National Laboratories under the Artificial Intelligence (AI) Singapore Programme (AISG) under Award AISG2-RP-2020-019; in part by the Energy Research Test-Bed and Industry Partnership Funding Initiative, Energy Grid (EG) 2.0 Programme, DesCartes and the Campus for Research Excellence and Technological Enterprise (CREATE) Programme; in part by Ministry of Education, Singapore (MOE) Tier 1 under Grant RG87/22; in part by the Ministry of Science and Information and Communication Technology (ICT) (MSIT), South Korea, under the ICT Creative Consilience Program under Grant IITP-2020-0-01821; and in part by the Information Technology Research Center (ITRC) Support Program Supervised by the Institute for ICT Planning and Evaluation (IITP) under Grant IITP-2023-RS-2023-00258639. *(Corresponding author: Dong In Kim.)*

Ismail Lotfi and Dusit Niyato are with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: ismail003@e.ntu.edu.sg; dniyato@ntu.edu.sg).

Sumei Sun is with the Institute for Infocomm Research, Agency for Science, Technology, and Research (A*STAR), Singapore 138632 (e-mail: sunsm@i2r.a-star.edu.sg).

Dong In Kim is with the Department of Electrical and Computer Engineering, Sungkyunkwan University (SKKU), Suwon 16419, South Korea (e-mail: dikim@skku.ac.kr).

Xuemin Shen is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: sshen@uwaterloo.ca).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JSAC.2023.3345402.

Digital Object Identifier 10.1109/JSAC.2023.3345402

strategy to maximize its revenue while minimizing the cost, i.e., prices, given to the sensing IoT devices for their data. In the downstream layer, the VSP also needs to find another optimal strategy for the prices to sell the digital twin service to the Metaverse users with the constraint on the delivery costs, i.e., computation cost for rendering and delivering the digital twin to the Metaverse users. The VSP needs to design incentive mechanism in both layers to motivate the sensing IoT devices and the users to participate in the Metaverse ecosystem. Specifically, the VSP uses private information of the participants to design a mechanism that satisfies the incentive compatibility (IC) and individual rationality (IR) properties for each participant. However, the uncertainty of the VSP about private information of the participants further hinders the derivation of the optimal strategy, causing the problem of information asymmetry [3]. The problem of information asymmetry encourages malicious participants to misreport their private information truthfully to gain higher profits. For instance, some sensing IoT devices with low ability to provide rich semantic information and fresh data might claim to have a higher level than their true type, causing the VSP to provide them with payments higher than what they truly deserve. This behaviour can similarly happen when the VSP intends to deliver the digital twin to the Metaverse users. The Metaverse users can misreport their private valuation about the delivered digital twin (which are based on their private types) to push the VSP to decrease the offered prices and hence, getting a higher utility than deserved.

To address the aforementioned challenges, we first propose the use of semantic information extraction algorithm on the raw data collected by the sensors on the IoT devices to minimize the size of the transmitted data [4]. Instead of transmitting the raw data to the VSP, the IoT devices transfer only relevant information (semantic) which can be used directly to create a 3D copy of the physical world. To formulate our problem, we then adopt contract theory which is regarded as an efficient tool to design incentive mechanism under asymmetric information scenarios [3]. Nevertheless, to properly design the contract and maximize its revenue, the VSP needs to be able to categorize the users based on their private types which can be multi-dimensional. In the upstream layer, different sensing IoT devices can have different semantic information abilities and different speed for data delivery. Similarly, in the downstream layer, different Metaverse users can have different private valuation towards different qualities of the digital twin delivered by the VSP, e.g., resolution and refresh rate. Therefore, the VSP needs to take these multi-dimensional private information of the participants of the Metaverse ecosystem while designing the incentive mechanism.

However, solving multi-dimensional contracts is not straightforward to address. Existing works proposed to reduce the multi-dimensional contract into a single dimensional contract [5], [6]. The derived single dimensional contract is then solved by using standard single-crossing condition technique. However, this conversion requires tedious mathematical transformations with proofs for IC and IR properties in addition to the computation efficiency. For instance, the single-crossing condition does not always hold when there are more than one type [5], [6]. Moreover, if the definition of the utility function of either the contract designer or the users changes slightly, the derived solution needs to be reformulated from scratch. Furthermore, all of the existing techniques to solve the single-dimensional contract requires certain assumptions about the used functions, e.g., monotonicity, which adds more limitations for the generality of the derived solutions. In this work, we address the problem of solving multi-dimensional contract from a totally different perspective. This is a key contribution of this paper. Specifically, we formulate the contract problem as a Markov Decision Process (MDP) and solve it using a new variant of multi-agent reinforcement learning (MARL) algorithm.

To the best of our knowledge, no previous work has attempted to solve the contract optimization problem using DRL. This is due to the fact that the nature of the problems addressed by DRL is different from those of contracts. For instance, a closely related work was proposed in [7] in which the authors adopted the DRL framework to solve the double Dutch auction (DDA). The DDA is conducted in many rounds and the DRL is suitable to learn the optimal "step size" over time. However, in contract, the contract bundles are delivered only once to the participants, and the participants select their preferred bundles only once. A major challenge for using DRL in contract is the generation of the training set. Typically, in problems where DRL is applied, the environment is highly repetitive, i.e., decisions are taken frequently, which enables the collection and evaluation of previous actions. However, in contract, the contract bundles are generated only once and they need to guarantee the optimality in addition to the IC and IR properties for all participants. These system settings make it non-trivial to adopt DRL for our problem. Nevertheless, motivated by the DRL's generality, we are able to create a learning environment for the proposed iterative contract problem.

In this paper, we extend our previous work in [8] to a multi-dimensional asymmetric information problem in a two-layer Metaverse system and develop a learning-based iterative contract to solve it. In summary, the main contributions of our work are as follows:

- We design a novel two layer Metaverse ecosystem where in the first layer, the VSP hires sensing IoT devices to collect data from the physical world, while in the second layer the VSP uses the collected data to create the digital twin of the physical world and delivers it to the Metaverse users. To minimize the data volume over the wireless link, we require the sensing IoT devices to extract and transmit only the semantic information from the raw data. The proposed design is shown to achieve the objectives of the Metaverse ecosystem, i.e., fast delivery and update of reliable information.
- We then use the contract theory framework to design an incentive mechanism to incentivize the participants in both layers, i.e., sensing IoT devices and Metaverse users, to engage in the Metaverse ecosystem and mitigate the adverse selection problem. We propose a novel iterative contract framework to solve the challenging multi-dimensional optimization problem. To the best of

our knowledge, this is the first work that applies contract theory in a two-layer Metaverse system. It is non-trivial to design an incentive mechanism for such systems due to information asymmetry at different layers, i.e., the data collection layer and data delivery layer.

- To solve the resultant iterative contract model, we develop a new variant of MARL systems where we consider that the VSP creates instances for each participant in the contract and interact with each other until reaching a feasible solution that maximizes the profit of the VSP while minimizing the IR and IC violation rates. To the best of our knowledge, this MARL design is the first of its kind.

The structure of the paper is as follows. In Section III we define our system model and provide some preliminaries about semantic information for the Metaverse. In Section IV we formulated the optimization problem as a contract theory problem and develop our learning-based iterative contract model. Finally, we provide numerical results and insightful discussions about our framework in Section V. Section VI concludes the paper.

II. Related Works

A. Metaverse Services

Digital twin modeling of the physical world in the Metaverse has a number of benefits for different application scenarios. In [1], the authors provided a detailed survey about the Metaverse and its applications and challenges from a communication perspectives. As the study of the Metaverse is still in its infancy, only few works have addressed the aspect of wireless resource allocation for the Metaverse. In [9], an IoT-assisted Metaverse sync problem was studied in which an evolutionary game was formulated to enable the IoT devices to select VSPs to work for. However, the volume of the data and the limited bandwidth problem was not addressed. In [10], an iterative algorithm based on transport theory was used to minimize the delivery time between the sensing IoT devices and the VSP, and hence minimize the gap between the real-time state of the physical twin and the state of the digital twin. However, it was assumed that the computation time, bandwidth, rate of the sensing IoT devices are precisely known. However, the sensing IoT devices might belong to different entities and hence, might misreport their private information and cause the whole Metaverse system to collapse. Therefore, there is a need to design incentive mechanism to "incentivize" participating sensing IoT devices to report their private information truthfully.

B. Semantic Communication for the Metaverse

Semantic communication for the Metaverse can be enabled at two different layers. The first layer is the physical layer where the objective is to use some technique, e.g., machine learning (ML), to reduce the number of transmitted bits between the transmitter and the receiver. The second layer is at the application layer where the raw data is reduced in size through ML or other techniques by extracting only the valuable, i.e., semantic, information. The derived semantic information is then transmitted to the receiver through standard wireless transmission technique. In [11], the authors developed a semantic multiverse communication system using generative adversarial networks (GANs). The encoder learns the semantic representations of the data, while the generator learns how to manipulate the extracted semantics for locally rendering the digital twin in the Metaverse. In [12], authors used contest theory to design a semantic-aware sensing information transmission for the Metaverse. Causality was used in [13] to infer cause-effects relationship between the transmitted symbols and the received ones over the wireless communication channel. A game theory based language model was then developed to enable minimalist representation and transmission of the extracted semantics. In [14], the relationship between objects in images was represented as a graph, and in [15] a dataset for this purpose was presented. In our recent work [8], we used the idea presented in [14] and proposed an application-layer semantic communication system for the Metaverse. A reverse auction mechanism was then developed to incentivize the data owners, i.e., sensing IoT devices, to make their bids to sell semantic information to the VSP truthfully.

C. Multi-Dimensional Contracts

To derive the optimal pricing bundles, the system designer needs to have full knowledge about the private information of the participants. To incentivize the participants to reveal their private information, existing works used either Stackelberg games or contract theory. However, the shortcoming of Stackelberg games is that they can be used only for scenarios with a single-dimension private type [16]. Therefore, several works used the framework of contract theory to design incentive mechanisms for problems with multi-dimensional private types. In the area of wireless communications, the idea of extending single-dimensional contract to two-dimensional contract was first proposed in [5], where the authors proposed to use an additional auxiliary variable to transform the two dimensional contract into a single-dimensional contract. The derived contract was then solved by using standard approaches in contract theory to prove the IC and IR properties in addition to the optimality of the derived solution. Motivated by [5], the authors in [6] extended the idea to a three-dimensional contract. However, this approach requires tedious formulation of the problem and several assumptions about the system dynamics to prove that the designed contract is truthful, i.e., does not violate the IC and IR properties. Finally, some existing works used the framework of contract theory to improve the performance of some ML problems, e.g., federated learning as in [17]. However, to the best of our knowledge, no prior work has attempted to use an ML technique, e.g., DRL, to address the aforementioned challenges in contract designs.

Motivated by the limitations mentioned above, we propose a learning-based iterative contract to derive the optimal pricing for the participants in the Metaverse ecosystem. Therefore, our studied Metaverse system can be regarded as an instance of

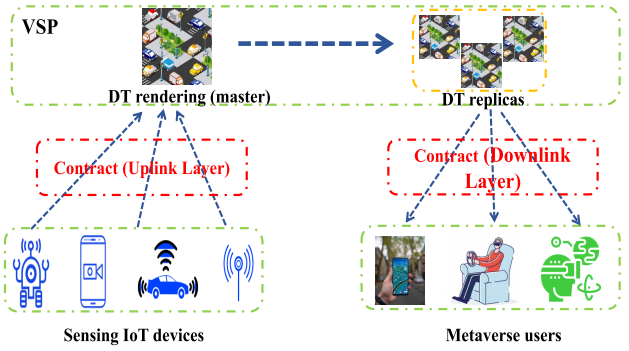


Fig. 1. System model.

a general framework for solving multi-dimensional contracts using DRL.

III. System Model and Preliminaries

We consider a digital market consisting of data owners, a VSP and Metaverse users. The data owners, e.g., IoT devices equipped with sensors, collect data about the physical environment and sell it to the VSP. The VSP then creates the digital twin of the physical environment and commercializes the digital twin to different Metaverse users. A two-layer contract theory-based framework is developed for the VSP to determine prices for purchasing data from the sensing IoT devices and for selling digital twin to the Metaverse users.

A. Metaverse Platform

As illustrated in Fig. 1, we consider a VSP that is collecting sensing data from a set of IoT devices, denoted as $\mathcal{N} = \{1, \ldots, N\}$ in the filed, e.g., vehicles or smartphones. The edge server, which is monitored by the VSP, is responsible for the replication of the physical twin by rendering the received data into a digital twin (DT) of the physical twin. Next, the VSP sends the digital twin to the set of Metaverse users, defined as $\mathcal{M} = \{1, \ldots, M\}$. Each sensing IoT device has a set of sensors to collect geo-spatial data from the surrounding environment and send the data back to the VSP. However, raw data is usually large in size adding further limitations on the required bandwidth and data delivery latency. Therefore, the IoT devices are equipped with machine learning (ML) models to extract only the semantic information from the collected raw data, which is smaller in size, and send the semantic information to the VSP. Nonetheless, if the received data from the IoT devices is outdated, the created digital twin will not be able to reflect real time dynamics of the physical twin. Therefore, the VSP leverages an age of information (AoI) metric to measure and guarantee freshness of the received data from the IoT devices. Once all of the semantic information is received by the VSP, the digital twin is created and distributed to the Metaverse users. In what follows, we discuss preliminaries about semantic information, AoI and their roles in deriving the value of the collected information by the sensing IoT devices and then we describe the delivery model of the digital twin by the VSP to the Metaverse users.

B. Fresh Semantic Information Collection Model

1) Sensing IoT Devices Modeling: Different from traditional crowd-sensing platforms that collect all raw data from data owners directly, the VSP obtains only semantic information from the IoT devices (e.g., semantic mask for each object in an image with its corresponding class or semantic text from voice recording). The incorporation of semantic information into our system is motivated by the following reasons:

- The number of communication channels available to the VSP are limited. Hence, if the VSP allows transmission of raw data by the IoT devices, only few devices will be able to transmit their data which reduces the heterogeneity of the collected data.
- Raw data is large in size in general (e.g., video and images), which can increase the transmission delay, making the rendering of the digital twin very slow and obsolete.
- The quality of the constructed digital twin will be higher as more semantic information about the physical world will be available to the VSP.

Let $\Psi = \{\psi_e : e \in \{1, \ldots, E\}\}$ denote the set of different semantic levels (or scores) available. The similarity score is impacted mainly by the algorithm used by the sensing IoT devices for semantic information extraction as demonstrated in [4] and [8]. We consider the algorithms as types (integers) and they are sorted in an ascending order, i.e., $0 < \psi_1 \leq \psi_2 \leq \cdots \leq \psi_E$. Note that the semantic extraction algorithm used by each sensing IoT device depends on the types of sensors equipped in each IoT device, e.g., camera and radar. Typically, sensing IoT devices with a high semantic score value can provide the VSP with more accurate and rich set of information. Therefore, they are more preferred by the VSP and should receive more payment for their data. However, as the sensing IoT devices are owned by independent parties, their capabilities of extracting the semantic information is different, heterogeneous and private. For instance, two IoT devices might have the same price for selling their semantic information, and the VSP might be indifferent when choosing which IoT device to buy data from. Therefore, if the VSP is aware of the semantic value of each IoT device, i.e., the ability of the IoT device to extract more accurate semantic information, the VSP can then choose the IoT devices that increase the quality of its constructed digital twin.

Nevertheless, the provided semantic information is affected directly by the reliability of the network link between the sensing IoT devices and the VSP. Even if the value of the provided semantic information is high, the link with high bit error rate (BER) can prevent the VSP from receiving the extracted semantic information about the physical world efficiently, and hence making the rendering at the Metaverse obsolete. To mitigate this issue, we consider the radio link transmission rate as a valuation metric for the link quality. The transmission rate can be adjusted by the transmitters through allocating more channels and/or increasing the transmit power to increase the signal-to-noise ratio (SNR) at the receiver. In what follows, we denote $\Lambda = \{\lambda_b : b \in \{1, \ldots, B\}\}$ to be the set of different available transmission rate values which

are also sorted in an ascending order,[1] i.e., $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_B$.

We also consider the age of information (AoI), which is defined as the time elapsed since the generation of the last received data at the source, as an important criterion of the delivered semantic information. Specifically, due to the congestion at the transmission queues, the AoI is directly impacted and becomes larger. It has been shown in [18] that with *first-come-first-served* (FCFS) queues, increasing the refresh rate does not yield a small AoI as this strategy may lead the destination to receive delayed status update because the packets become backlogged in the communication system. These reasons motivates the use of *last-come-first-served* (LCFS) queues with preemption as in [18]. Under LCFS with preemption, the new generated packet is allowed to replace the current packet in service and hence, maintain a low AoI. Let $\Gamma = \{\gamma_c : c \in \{1, \ldots, C\}\}$ denote the set of different refresh rate types. The refresh rate type is related to the average AoI ϖ_γ at the VSP as follows [18]

$$\varpi_\gamma = \frac{1}{\gamma} + \frac{1}{\mu}, \tag{1}$$

where $1/\gamma$ is the mean packet arrival time at the VSP and $1/\mu$ is the mean processing time at the VSP server. From (1) we observe that when μ is considered constant the average AoI is inversely proportional to the refresh rate γ. Therefore, sensing IoT devices which have higher refresh rates bring more utility to the VSP. Based on the findings in [19] and [20], we consider that as the refresh rate increases, the AoI decreases following a non-increasing convex function. Without loss of generality we consider that refresh rate types are sorted in an ascending order similar to the other types, i.e., $0 < \gamma_1 \leq \gamma_2 \leq \cdots \leq \gamma_C$.

In short, each sensing IoT device is differentiated by its three dimensional private information: the semantic score value ψ_e, transmission rate value λ_b and refresh rate value γ_c.[2] Therefore, the utility of a sensing IoT device with type-(λ, γ, ψ) to deliver semantic information to the VSP with volume size $\hat{s}$ is defined as

$$U^\dagger(\hat{s}_{\lambda,\gamma,\psi}) = \pi^\dagger_{\lambda,\gamma,\psi} - \Upsilon^\dagger(\hat{s}_{\lambda,\gamma,\psi}), \tag{2}$$

where $\pi^\dagger_{\lambda,\gamma,\psi}$ is the price associated to data with quality $\hat{s}_{\lambda,\gamma,\psi}$, $\Upsilon^\dagger(\hat{s}_{\lambda,\gamma,\psi})$ refers to the cost for delivering a data with size $\hat{s}$ by an IoT device with type-(λ, γ, ψ) to the VSP, and is defined as

$$\Upsilon^\dagger(\hat{s}_{\lambda,\gamma,\psi}) = \Upsilon^\dagger_0 + T^\dagger(\hat{s}_{\lambda,\gamma,\psi}), \tag{3}$$

where $\Upsilon^\dagger_0$ is a fixed cost and $T^\dagger(\hat{s}_{\lambda,\gamma,\psi})$ is the specific cost for data generated by type-(λ, γ, ψ) IoT device. The cost function reflects both the computation cost (i.e., data collection and semantic information extraction) and communication cost (i.e., channel allocation by the sensing IoT devices).

[1]Note here that the transmission rate has a discrete value due to modulation and coding schemes.

[2]Note here that the transmission rate captures the volume of data that can be transmitted over a period of time while the refresh rate captures the number of times the transmitter sends a new update to the receiver.

2) VSP Modeling: The VSP needs to properly design rewards for each sensing IoT device type. Different from most existing works where some private information of the users are known to the VSP, we are considering a more realistic asymmetric information scenario in which all the private information of the sensing IoT devices are not known to the VSP. In other words, the VSP does not know exactly the type of each IoT device, i.e., its semantic score value ψ_e, its transmission rate value λ_b or its refresh rate value γ_c. To solve the asymmetric information problem, we incorporate contract theory into our model. Specifically, the VSP designs specific contract bundles for each type of the sensing IoT devices with the aim of maximizing its utility, i.e., profit, while ensuring that each sensing IoT device does not deviate from choosing the bundle designed for its true type. The VSP designs a contract bundle, denoted as $\Omega^\dagger = \{\omega_{\lambda,\gamma,\psi} : \lambda \in \Lambda, \gamma \in \Gamma, \psi \in \Psi\}$ that consists of $B \times C \times E$ contract items denoted as $\omega_{\lambda,\gamma,\psi} = \{\hat{s}_{\lambda,\gamma,\psi}, \pi^\dagger_{\lambda,\gamma,\psi}\}$ and characterized by a joint probability mass function $Q^\dagger(\lambda, \gamma, \psi)$ for each IoT device's joint type combination. Hence, the VSP's utility from type-(λ, γ, ψ) IoT device is given by

$$R^\dagger(\omega_{\lambda,\gamma,\psi}) = \sigma f(\hat{s}_{\lambda,\gamma,\psi}) - \pi^\dagger_{\lambda,\gamma,\psi} + (K - \varpi_\gamma), \tag{4}$$

where σ is the revenue coefficient for the VSP and $\sigma f(\hat{s})$ is the revenue of the VSP from the data received from the sensing IoT device with type-(λ, γ, ψ). Motivated by [21], we adopt the α-fairness function to define $f(\hat{s})$ as follows:

$$f(\hat{s}) = \frac{1}{1-\alpha}\hat{s}^{1-\alpha}, \tag{5}$$

where $0 < \alpha < 1$ is a given constant. The last term $(K - \varpi_\gamma)$ in (4) represents the benefit from the AoI. Specifically, K is a constant and $(K - \varpi_\gamma)$ can be interpreted as a satisfactory function from the average AoI [22], where a low average AoI brings a high benefit to the VSP. The overall utility of the VSP from all sensing IoT devices is then formulated as

$$R^\dagger(\Omega^\dagger) = \sum_{\lambda\in\Lambda}\sum_{\gamma\in\Gamma}\sum_{\psi\in\Psi} NQ^\dagger(\lambda,\gamma,\psi)\Big(\sigma f(\hat{s}_{\lambda,\gamma,\psi}) - \pi^\dagger_{\lambda,\gamma,\psi} + (K - \varpi_\gamma)\Big). \tag{6}$$

C. Digital Twin Delivery Model

1) Metaverse Users Modeling: In our model, we define the quality of a digital twin with respect to the Metaverse users in terms of the resolution of the digital twin and the refresh rate per time unit [23]. The resolution captures the size of the transmitted data while the refresh rate captures the freshness of the data. In other words, if a Metaverse user subscribes to a digital twin delivery service with resolution r (e.g., pixel per inch), and refresh rate h (e.g., frame per second (FPS)), the VSP will assert the delivery of the digital twin as requested to the Metaverse user. If the Metaverse user accepts to buy a replica of the digital twin with quality (r, h), the VSP delivers that replica to the Metaverse user and charges with price $\pi^\ddagger(r, h)$. Nonetheless, the Metaverse users have different preferences towards various combinations of resolutions and refresh rates. To present this preference,

we use a valuation function with both resolution and refresh rate parameters. Specifically, each Metaverse user has some private valuation of both resolution and refresh rate, denoted as τ and ϕ, respectively. These private valuation parameters capture both *resolution sensitivity*, i.e., perception, and *refresh rate sensitivity*, i.e., timeliness. Based on the works in [21] and [24], we define the valuation of the Metaverse user with type-(τ, ϕ) to the provided digital twin with resolution r and refresh rate h as

$$V^{\ddagger}(\tau, \phi, r, h) = \tau g_1(r) + \phi g_2(h), \tag{7}$$

where $g_1(\cdot)$ and $g_2(\cdot)$ follow an α-fairness function as earlier described in (5) with changes only to parameter α. The Metaverse user is also required to have enough bandwidth to receive data from the VSP in addition to its internal hardware specifications, e.g., screen refresh rate [25]. Moreover, as the refresh rate h (in FPS) increases, the inter-frame time decreases which is more preferred by the Metaverse user. The Metaverse user needs to trade-off between the quality of the delivered digital twin and the cost. Therefore, the utility of the Metaverse user with type-(τ, ϕ) after purchasing a digital twin with quality (r, h) is defined as

$$U^{\ddagger}(\tau, \phi, r, h) = V^{\ddagger}(\tau, \phi, r, h) - \pi^{\ddagger}(r, h). \tag{8}$$

2) VSP Modeling: To guarantee the delivery of the digital twin with the specified quality, the VSP needs to use a certain number of resources and algorithms which increases the delivery cost as the quality increases. For example, instead of using a single processor or a single queue, to deliver all the digital twin packets to the Metaverse users, the waiting time in the queue can be minimized for each Metaverse user, and hence, minimizing the AoI at the Metaverse user side [20]. This adjustment significantly minimizes the AoI at the Metaverse user side but increases the cost for the VSP. We define the cost to the VSP to deliver the digital twin with quality (r, h) as

$$\Upsilon^{\ddagger}(r, h) = \Upsilon_0^{\ddagger} + T^{\ddagger}(r, h), \tag{9}$$

where $\Upsilon_0^{\ddagger}$ is a fixed cost for the VSP to collect data from the sensing IoT devices and render the digital twin. $T^{\ddagger}(r, h)$ is the specific cost for quality (r, h). Finally, the utility of the VSP for delivering a digital twin with quality (r, h) is defined as the difference between the selling price and the cost, i.e.,

$$R^{\ddagger}(r, h) = \pi^{\ddagger}(r, h) - \Upsilon^{\ddagger}(r, h). \tag{10}$$

IV. Contract Formulation

In this section, we formulate the contract design problem to maximize the utility of the VSP when buying the semantic information from the sensing IoT devices and when selling the constructed digital twin to the Metaverse users. For the contract to be feasible, it has to guarantee both the incentive compatibility (IC) and individual rationality (IR) properties for all types [26]. In what follows, we describe IR and IC properties with respect to the upstream layer, i.e., for the contract between the VSP and the sensing IoT devices, and with respect to the downstream layer, i.e., between the VSP and the Metaverse users. Finally, we propose a DRL-based model to solve the contracts of the upstream and downstream layers, which we call *iterative contract* and is -to the best of our knowledge- an unprecedented method to solve contracts.

A. Upstream Layer (VSP and Sensing IoT Devices)

The VSP obtains historical data about the semantic levels and transmission rates of different sensing IoT devices. The average AoI for each IoT device (and hence, the refresh rate) is derived by the VSP from historical interactions. The VSP then designs a contract by solving problem (13) and broadcasts the designed contract to the IoT devices. Next, each IoT device sends its selected contract item to the VSP, i.e., signs the contract with the VSP. Finally, the IoT devices send their semantic information to the VSP and receive payments as specified in the contract. A feasible contract in an open market must satisfy the IR and IC properties. The IR and IC properties of the upstream layer are defined as follows.

Definition 1: Individual Rationality (IR) for IoT device: An IoT device with type-(λ, γ, ψ) will only accept to sell its semantic information to the VSP if its utility is non-negative, i.e.,

$$U^{\dagger}_{\lambda,\gamma,\psi}(\hat{s}_{\lambda,\gamma,\psi}) \geq 0, \quad \forall \lambda \in \Lambda,\ \forall \gamma \in \Gamma,\ \forall \psi \in \Psi. \tag{11}$$

Definition 2: Incentive Compatibility (IC) for IoT device: The utility of an IoT device with type-(λ, γ, ψ) is maximized only when selecting the contract designed for its true type, i.e.,

$$\begin{aligned} U^{\dagger}_{\lambda,\gamma,\psi}(\hat{s}_{\lambda,\gamma,\psi}) \geq U^{\dagger}_{\lambda,\gamma,\psi}(\hat{s}_{\lambda',\gamma',\psi'}), \quad & \forall \lambda, \lambda' \in \Lambda, \\ & \forall \gamma, \gamma' \in \Gamma,\ \forall \psi, \psi' \in \Psi, \lambda \neq \lambda', \\ & \gamma \neq \gamma', \psi \neq \psi'. \end{aligned} \tag{12}$$

The IR condition ensures the participation of the sensing IoT devices while the IC condition ensures that each sensing IoT device selects the contract designed for its true type. The aim of the VSP is to design a contract $(\hat{\boldsymbol{s}}, \boldsymbol{\pi}^{\dagger})$ to maximize its utility taking into account the IR and IC conditions, which is expressed as follows:

$$\begin{aligned} \mathcal{P}_1: \quad & \max_{(\hat{\boldsymbol{s}}, \boldsymbol{\pi}^{\dagger})} \sum_{\lambda \in \Lambda} \sum_{\gamma \in \Gamma} \sum_{\psi \in \Psi} NQ^{\dagger}(\lambda, \gamma, \psi) \Big(\sigma f(\hat{s}_{\lambda,\gamma,\psi}) \\ & \qquad - \pi^{\dagger}_{\lambda,\gamma,\psi} + (K - \varpi_{\gamma}) \Big) \qquad \text{(13a)} \\ & s.t. \quad (11) \text{ and } (12). \qquad \text{(13b)} \end{aligned}$$

However, the parameters in (11) and (12) are private to the sensing IoT devices and can be misreported to gain higher utility than deserved. Moreover, to solve $\mathcal{P}_1$ we need to address $B \times C \times E$ IR constraints and $(B \times C \times E) \times (B \times C \times E - 1)$ IC constraints, which are all non-convex. Intuitively, such an optimization problem is not straightforward to solve. The classical approach is to first define some lemmas to constrain the pricing function and the types, e.g., monotonicity and pairwise incentive compatibility, and then relax the optimization problem to reduce its complexity. However, these methods are not directly applicable here due to the multi-dimensionality of the contract. Interestingly, some recent works proposed to introduce an auxiliary type to reduce the dimensionality

of the contract and then solve the relaxed problem using dynamic programming or branch and bound techniques [5], [6]. However, these approaches add another layer of difficulty to the problem formulation, which become more tedious, time consuming to adjust and to prove the corresponding lemmas and theorems. Furthermore, the necessary and sufficient conditions for these approaches further tighten the overall assumptions in the system and limit its generality. In this work, we design a DRL-based iterative multi-dimensional contract that is executed over several interactions between the VSP and the sensing IoT devices. The VSP starts with a set of random bundles and converges to the optimal set of bundles, which is the objective of the contract designer. In what follows, we first start by defining the MDP of the upstream layer, then briefly discuss how this MDP is solved.

1) Markov Decision Process: An MDP is defined by a tuple $< \mathcal{S}^{\dagger}, \mathcal{A}^{\dagger}, \mathcal{P}^{\dagger}, r^{\dagger} >$ where $\mathcal{S}^{\dagger}$ is the state space, $\mathcal{A}^{\dagger}$ is the action space, $\mathcal{P}^{\dagger}$ is the state transition probabilities and $r^{\dagger}$ is the immediate reward received by the agent, i.e., the VSP, after performing action $a^{\dagger}$ at state $s^{\dagger}$.

a) State space: The state space of the system at time slot t $(t = 1, 2, \ldots, T)$ is defined as

$$\mathcal{S}^{\dagger(t)} \triangleq \left\{ \mathbf{a}^{\dagger(t-1)}, \boldsymbol{\pi}^{\dagger(t)}, \hat{\boldsymbol{s}}^{(t)}, \mathbf{x}^{\dagger(t)}, \mathbf{y}^{\dagger(t)} \right\}, \tag{14}$$

where $\mathbf{a}^{\dagger(t-1)}$ is the action vector from the previous time slot, $\boldsymbol{\pi}^{\dagger(t)}$ is the price vector at time slot t and $\hat{\boldsymbol{s}}^{(t)}$ is the semantic information size vector at time slot t. $\mathbf{x}^{\dagger(t)}$ and $\mathbf{y}^{\dagger(t)}$ are binary vectors of sensing IoT devices which have their IR an IC violated, respectively. The system state is then defined as a composite variable $\mathbf{s}^{\dagger} \triangleq (\mathbf{a}^{\dagger}, \boldsymbol{\pi}^{\dagger}, \hat{\boldsymbol{s}}, \mathbf{x}^{\dagger}, \mathbf{y}^{\dagger}) \in \mathcal{S}^{\dagger}$.

b) Action space: For better budget allocation, the VSP is able to dynamically adjust the prices and the semantic information size values for each contract bundle. Let $price_k$ denote the price for bundle k and $\eta_{1,k}$ a scalar to adjust the price between two time slots for the contract bundle k. The price is updated as

$$\pi_k^{\dagger,(t+1)} = price_k \times (1 + \eta_{1,k}^{(t)}), \tag{15}$$

where $\eta_{1,k}^{(t)} \in [-range, range]$ and $0 \leq range \leq 1$. The semantic information size values are adjusted similarly. Let $size_k$ denote the semantic information size value for bundle k and $\eta_{2,k}$ a scalar to adjust the size between two time slots for the contract bundle k. The semantic information size value is updated as

$$\hat{s}_k^{(t+1)} = size_k \times (1 + \eta_{2,k}^{(t)}), \tag{16}$$

where $\eta_{2,k}^{(t)} \in [-range, range]$. Based on these definitions of the price and semantic information size adjustments, the action space of the VSP consists of the joint action of reducing, increasing or keeping the current price and semantic information size value for all contract bundles at time slot t. Therefore, the action space is defined by: $\mathcal{A}^{\dagger} \triangleq \left\{ (a', a'') : a', a'' \in \{0, 1, 2\} \right\}$, where $a' = 0$, $a' = 1$ and $a' = 2$ refer to the actions of increasing the semantic information size, decreasing the semantic information size or keeping the current size, respectively. Similarly, $a'' = 0$, $a'' = 1$ and $a'' = 2$ refer to the actions of increasing, decreasing or keeping the current price, respectively. Intuitively, this definition of the action space implies a total of 9 different combination of actions, i.e., (3×3) actions, at each time slot. This strategy significantly reduces the action space size which helps the DRL to converge quickly.

c) Immediate reward: Since our objective in the contract is to maximize (13), we craft the immediate reward function to align with this objective. To incorporate the IC and IR constraints in (13) into the immediate reward function, we design a multi-objective reward function based on weighted sum technique. Specifically, we define the reward function as follows:

$$\begin{aligned} & r^{\dagger}(\mathcal{S}^{\dagger(t)}, a^{\dagger(t)}, \mathcal{S}^{\dagger(t+1)}) \\ &= w_1 \sum_{\lambda \in \Lambda} \sum_{\gamma \in \Gamma} \sum_{\psi \in \Psi} n_{\lambda,\gamma,\psi} \Big(\sigma f(\hat{s}_{\lambda,\gamma,\psi}^{(t)}) - \pi_{\lambda,\gamma,\psi}^{\dagger,(t)} + (K - \varpi_{\gamma}) \Big) \\ &\quad + w_2 \big[\sum \mathbf{x}^{\dagger(t)} - \sum \mathbf{x}^{\dagger(t+1)} \big] \\ &\quad + w_3 \big[\sum \mathbf{y}^{\dagger(t)} - \sum \mathbf{y}^{\dagger(t+1)} \big], \end{aligned} \tag{17}$$

where $w_1 + w_2 + w_3 = 1$ are the weight factors of each term in (17) and $n_{\lambda,\gamma,\psi}$ is the number of sensing IoT devices with type-(λ, γ, ψ). The first term in (17) reflects the objective of maximizing the VSP's revenue. The second and third terms reflect the objective of reducing the number of violations of IR and IC properties, respectively. Note here that rewards are only received after the increment of the timestep.

d) Optimization formulation: The objective is to find a policy $\mathbf{p}^{\dagger *}$ that has the best mapping from states to actions which maximizes the average long-term reward $\mathcal{R}(\mathbf{p}^{\dagger})$. Formally, the optimization problem is defined as

$$\max_{\mathbf{p}^{\dagger}} \quad \mathcal{R}(\mathbf{p}^{\dagger}) = \lim_{\Upsilon \to \infty} \frac{1}{\Upsilon} \sum_{t=1}^{\Upsilon} \mathbb{E}(r_t^{\dagger}(s_t^{\dagger}, \mathbf{p}^{\dagger}(s_t^{\dagger}))), \tag{18}$$

where $r_t^{\dagger}(s_t^{\dagger}, \mathbf{p}^{\dagger}(s_t^{\dagger}))$ is the immediate reward under policy $\mathbf{p}^{\dagger}$ at time t defined in (17).[3]

The standard approach to solve the MDP described earlier is to adopt one of the available single-agent DRL algorithms, e.g., Deep Q-Network (DQN) or Proximal Policy Optimization (PPO) [27]. However, standard single agent DRL algorithms cannot solve the described MDP. Specifically, in single agent DRL and at each time slot, the agent extracts a single action to perform. However, in our MDP there is a need to perform N actions simultaneously. As there are several actions to be executed simultaneously, an attractive approach is to adopt multi-agent reinforcement learning (MARL) [27]. In MARL systems, several agents are trained to work independently to achieve one goal or compete against each other. However, our studied system also differs from these settings as we only want to train the VSP to derive the optimal contract and there are no other agents to train. Inspired by MARL and the work in [28], we develop a novel MARL architecture to solve the optimal iterative contract problem. In what follows, we first continue

[3]Note that the unique ability of our solution is at optimizing for other objectives. Specifically, we might have other objectives that can be simply achieved by modification to the reward function, e.g., maximizing the social welfare of the system.

the formulation of the downstream layer and its corresponding MDP. Next, we describe the details of our proposed learning-based iterative contract.

B. Downstream Layer (VSP and Metaverse Users)

In this layer, the objective of the VSP is to find a set of qualities of the delivered digital twin jointly with their respective prices to maximize its revenue. As earlier described, the quality of the delivered digital twin is measured using the resolution (which reflects the perception) and refresh rate (which reflects the timeliness of the information). Hereafter, we denote the set of available resolutions as $\mathcal{R}$, the set of refresh rates as $\mathcal{H}$, and the set of prices as $\Pi^{\ddagger}$. Here we consider that the different combinations of resolutions and refresh rates are referred to by an auxiliary variable q. For each Metaverse user with type-(τ,ϕ), the VSP assigns a quality $q_{\tau,\phi}$ and charges a price $\pi^{\ddagger}_{\tau,\phi}$. The set of quality-price combinations is denoted as $\Omega^{\ddagger}=\{(q_{\tau,\phi},\pi^{\ddagger}_{\tau,\phi})|\forall\tau\in\Xi,\forall\phi\in\Phi\}$. The IR and IC properties of the downstream layer are defined as follows.

Definition 3: Individual Rationality (IR) for Metaverse user: A Metaverse user with type-(τ,ϕ) will only accept to purchase the digital twin from the VSP if its utility[4] is non-negative, i.e.,

$$V^{\ddagger}(\tau,\phi,q_{\tau,\phi})-\pi^{\ddagger}_{\tau,\phi}\geq 0,\quad \forall\tau\in\Xi,\forall\phi\in\Phi. \tag{19}$$

Definition 4: Incentive Compatibility (IC) for Metaverse user: The utility of a Metaverse user with type-(τ,ϕ) is maximized only when selecting the contract designed for its true type, i.e.,

$$\begin{aligned} &V^{\ddagger}(\tau,\phi,q_{\tau,\phi})-\pi^{\ddagger}_{\tau,\phi}\\ &\quad\geq V^{\ddagger}(\tau,\phi,q_{\tau',\phi'})-\pi^{\ddagger}_{\tau',\phi'},\\ &\quad\forall\tau,\tau'\in\Xi,\ \forall\phi,\phi'\in\Phi,\ \tau'\neq\tau,\ \phi'\neq\phi. \end{aligned} \tag{20}$$

Since the VSP dominates the trading process, we model the digital twin trading as a *monopoly market*, in which the VSP's objective is to maximize its overall utility, which is written as

$$R^{\ddagger}(\Omega^{\ddagger})=\sum_{\tau\in\Xi}\sum_{\phi\in\Phi}MQ^{\ddagger}(\tau,\phi)\left(\pi^{\ddagger}_{\tau,\phi}-\Upsilon^{\ddagger}(q_{\tau,\phi})\right), \tag{21}$$

where $Q^{\ddagger}(\tau,\phi)$ is the joint probability mass function of the Metaverse users having type-(τ,ϕ) and is obtained from previous observations [21]. For instance, each Metaverse user device support a different frame rate and have different valuation towards them, which is a private information for each Metaverse user. Nevertheless, the VSP has some prior knowledge about their probability distribution which is modeled here using $Q^{\ddagger}(\tau,\phi)$. In order for the contract to be feasible, it has to guarantee both IC and IR. Therefore, the optimal contract can be derived by solving the following problem:

$$\mathcal{P}_2:\max_{(\boldsymbol{q}^{\ddagger},\boldsymbol{\pi}^{\ddagger})}\sum_{\tau\in\Xi}\sum_{\phi\in\Phi}MQ^{\ddagger}(\tau,\phi)\left(\pi^{\ddagger}_{\tau,\phi}-\Upsilon^{\ddagger}(q_{\tau,\phi})\right), \tag{22a}$$

$$s.t.\quad (19)\ and\ (20). \tag{22b}$$

[4]Note that we use $V^{\ddagger}(\tau,\phi,q_{\tau,\phi})$ instead of $V^{\ddagger}(\tau,\phi,r,h)$ for notational consistency.

However, the parameters in (19) and (20) are private to the Metaverse users and can be misreported. Similar to the upstream layer problem, this problem is addressed using DRL. Therefore, in what follows, we first start by describing the MDP of the downstream layer. Next, the solution of this MDP and that of the upstream layer is described in detail.

1) Markov Decision Process: The MDP of the downstream layer is defined by the tuple $<\mathcal{S}^{\ddagger},\mathcal{A}^{\ddagger},\mathcal{P}^{\ddagger},r^{\ddagger}>$ where $\mathcal{S}^{\ddagger}$ is the state space, $\mathcal{A}^{\ddagger}$ is the action space, $\mathcal{P}^{\ddagger}$ is the state transition probabilities and $r^{\ddagger}$ is the immediate reward received by the agent, i.e., the VSP, after performing action $a^{\ddagger}$ at state $s^{\ddagger}$.

a) State space: The state space of the system at time slot t $(t=1,2,\ldots,T)$ is defined as

$$\mathcal{S}^{\ddagger(t)}\triangleq\left\{\mathbf{a}^{\ddagger(t-1)},\boldsymbol{\pi}^{\ddagger(t)},\boldsymbol{q}^{(t)},\mathbf{x}^{\ddagger(t)},\mathbf{y}^{\ddagger(t)}\right\}, \tag{23}$$

where $\mathbf{a}^{\ddagger(t-1)}$ is the action vector from the previous time slot, $\boldsymbol{\pi}^{\ddagger(t)}$ is the price vector at time slot t and $\boldsymbol{q}^{(t)}$ is the digital twin quality vector at time slot t. $\mathbf{x}^{\ddagger(t)}$ and $\mathbf{y}^{\ddagger(t)}$ are binary vectors of Metaverse users which have their IR an IC violated, respectively. The system state is then defined as a composite variable $\mathbf{s}^{\ddagger}\triangleq(\mathbf{a}^{\ddagger},\boldsymbol{\pi}^{\ddagger},\boldsymbol{q},\mathbf{x}^{\ddagger},\mathbf{y}^{\ddagger})\in\mathcal{S}^{\ddagger}$.

b) Action space: The action space for the downstream layer is identical to that of the upstream layer with difference only in adjusting the quality instead of the semantic information size.

c) Immediate reward: The reward function of the downstream layer is crafted to maximize (22) while incorporating the IR and IC constraints defined in (19) and (20), respectively. Therefore, the reward function of the downstream layer is formalized as

$$\begin{aligned} &r^{\ddagger}(\mathcal{S}^{\ddagger(t)},a^{\ddagger(t)},\mathcal{S}^{\ddagger(t+1)})\\ &\quad=w'_1\sum_{\tau\in\Xi}\sum_{\phi\in\Phi}n_{\tau,\phi}\left(\pi^{\ddagger,(t)}_{\tau,\phi}-\Upsilon^{\ddagger}(q_{\tau,\phi})\right)\\ &\qquad+w'_2\left[\sum\mathbf{x}^{\ddagger(t)}-\sum\mathbf{x}^{\ddagger(t+1)}\right]\\ &\qquad+w'_3\left[\sum\mathbf{y}^{\ddagger(t)}-\sum\mathbf{y}^{\ddagger(t+1)}\right], \end{aligned} \tag{24}$$

where $w'_1+w'_2+w'_3=1$ are the weight factors and $n_{\tau,\phi}$ is the number of Metaverse users with type-(τ,ϕ).

d) Optimization formulation: The optimization problem of the downstream layer is defined as

$$\max_{\mathbf{p}^{\ddagger}}\quad\mathcal{R}(\mathbf{p}^{\ddagger})=\lim_{\Upsilon\to\infty}\frac{1}{\Upsilon}\sum_{t=1}^{\Upsilon}\mathbb{E}(r^{\ddagger}_t(s^{\ddagger}_t,\mathbf{p}^{\ddagger}(s^{\ddagger}_t))), \tag{25}$$

where $r^{\ddagger}_t(s^{\ddagger}_t,\mathbf{p}^{\ddagger}(s^{\ddagger}_t))$ is the immediate reward under policy $\mathbf{p}^{\ddagger}$ at time t defined in (24).

C. Iterative Contract Design

The proposed learning-based iterative contract is shown in Fig. 2 while Algorithm 1 summarizes the major steps. Here we describe the framework with respect to the upstream layer while its application on the downstream layer is straightforward as clearly seen from the similarity between their MDPs.

Specifically, the algorithm (administered by the VSP) starts by initializing the semantic information size vector $\hat{\boldsymbol{s}}$ and the price vector $\boldsymbol{\pi}^{\dagger}$ based on a uniform distribution from the

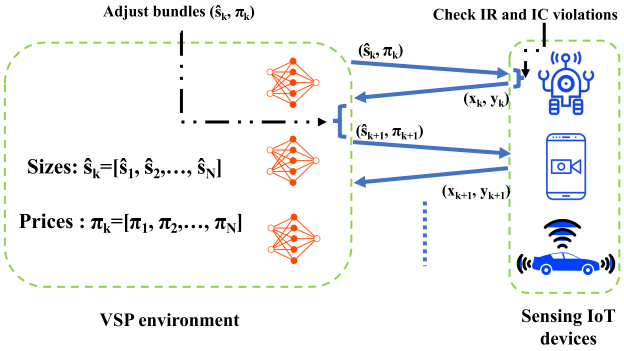


Fig. 2. Proposed learning-based iterative contract.

intervals $[size_k \times (1 - range), size_k \times (1 + range)]$ and $[price_k \times (1 - range), price_k \times (1 + range)]$, respectively. In addition, the binary vectors $\mathbf{x}^{(t)}$ and $\mathbf{y}^{(t)}$ are initially set to 1 for all the vectors' elements. Next, the VSP initializes N single-agent DRL networks to learn the optimal strategy for adjusting the bundles of each sensing IoT device. As there are a variety of DRL algorithms with some algorithms working better in some domains than others, we adopt our previously developed prioritized double deep Q-Learning (PDDQL) algorithm in [29] and refer the reader for the detailed description therein. At the very first iteration of the algorithm, the VSP populates the initial set of bundles directly to the N sensing IoT devices. The number of different types in the multi-dimensional contract is extracted from previous interactions with the sensing IoT devices. Once the sensing IoT devices receive the contract bundles, each sensing IoT device verifies whether its IR or IC is violated based on (11) and (12) and then returns a binary tuple (x, y) to the VSP. The VSP then constructs the full vectors $\mathbf{x}^{(t)}$ and $\mathbf{y}^{(t)}$ and shares their content with each agent in its environment. We refer to this step as augmentation of the MDP as the agents are augmented with information initially unobservable about the states of each other (i.e., IR and IC violations, previous actions, semantic information sizes and prices). At this point, each agent executes the PDDQL algorithm to extract the optimal adjustment to be performed based on the set of actions as earlier defined. As such, we name this algorithm augmented multi-agent PDDQL (MA-PDDQL). Next, the VSP performs the appropriate adjustments for each bundle, i.e., semantic information sizes and prices, and then delivers the new set of bundles to the sensing IoT devices to start the next round. Once we reach a state where $x_i^{(t)} = y_i^{(t)} = 0$ for all the vectors elements, the derived contract is considered feasible and satisfy IR and IC conditions. However, the solution is not necessarily optimal. Several round needs to be executed until no improvement in the VSP's utility/revenue is obtained. For this reason, we call our framework as an *interactive contract* where the optimal contract is derived based on several rounds of interaction between the VSP and the sensing IoT devices.

We should note here some key features in the design of our proposed framework:

- First, a technical challenge to solve is the convergence of the DRL-based contract as the prices of all bundles change simultaneously. Specifically, the very frequent changes in the price or the semantic information size of one bundle affects the strategy of all the participants when choosing their optimal contract bundle. This makes the DRL environment non-stationary and noisy for each agent to learn a stable policy. We address this issue by establishing a virtual communication channel between the learning agents in the MARL environment, which we refer to as augmentation of the MDP. Specifically, after receiving the IR and IC tuples from all the sensing IoT devices, the full vectors $\mathbf{x}^{(t)}$ and $\mathbf{y}^{(t)}$ are created and shared as part of the state of each agent. In addition, each agent in the VSP environment is aware of the current set of bundles and the previously taken actions by all other agents. This augmentation of the observation space for each agent makes the MDP easily learnable and the agents can then learn from collective experiences.
- The MA-PDDQL algorithm requires from the sensing IoT devices only a flag about their IR and IC status and not their private types, which is totally different from existing contract solutions that requires the disclosure of these information (referred to as the revelation principal in contract theory [3]). Some privacy-sensitive participants may be reluctant about engaging in such contracts as their private information might be used for other purposes beyond the contract, e.g., delivering dedicated advertisement as studied in [30]. Our design preserves the privacy of the participants about their private types which is a major usefulness of our proposed framework.
- Finally, note here that as shown in Fig. 2, the computation complexity of our model is calculated based on the number of iterative interactions between the sensing IoT device and the VSP until convergence. The addition of another type (e.g., location of the sensing IoT device) would not have a major impact on the computation complexity as this additional type will be part of the IR and IC equations that are calculated for the total number of different combinations of types. For instance, for 3 types with sizes B, C and E, the total number of IR and IC equations to be calculated is $2 \times B \times C \times E$. The addition of another type with size F will increase this number to $2 \times B \times C \times E \times F$, which does not have a major significance.

Next, we evaluate the proposed learning-based iterative multi-dimensional contract framework.

V. Numerical Evaluation

In this section, we validate the performance of our proposed iterative contract for the Metaverse through extensive simulations. As in [8], we use existing semantic extraction algorithms to process images and radar signals to extract the semantics [4]. As stated before, the structure of our iterative contract in the upstream layer is quite similar to the one in the downstream layer. Therefore, we present here the numerical results for the upstream layer only.

A. Simulation Settings

Unless otherwise stated, the DRL algorithm is trained over 700 episode with 200 iterations on each episode. In addition,

Algorithm 1 Augmented MA-PDDQL Algorithm Pseudo-Code

Input : Initialize semantic information size and prices to random values.
Output: Optimal semantic information sizes and prices $(\hat{s}, \pi^{\dagger})$.

for $t = 1, 2, \ldots$ **to** *convergence* **do**
  Initialize empty action list;
  for $i = 1, 2, \ldots$ **to** N **do**
    Select action for device i based on PDDQL and append to the action list;
  end
  Execute the simultaneous actions from the action list and get the next state;
  for $i = 1, 2, \ldots$ **to** N **do**
    Store the tuple (s_t, a_t, s_{t+1}, r_t) and update the policy of agent i;
  end
end

the weighting factors in the reward function are all set to 0.33. We also consider a total of 27 ($3 \times$ 3x3) different sensing IoT device types. The type of each sensing IoT device is chosen uniformly from the set of possible joint types $(B \times C \times E)$. Furthermore, the types and other variables such as the transmit power and energy consumption of the sensing IoT devices are normalized between 0 and 1. This step is essential and common in deep learning for the algorithms to converge [27]. The description of the double DQN algorithm and the prioritized reply memory technique can be found in our previous work [29]. An important parameter to define is the set from which the semantic information sizes and prices are chosen. We consider that $range = 0.9$ while $n_{1,k}$ and $n_{2,k}$ take values from the discrete set $[-0.9, -0.7, \ldots, 0.7, 0.9]$. The choice of having this set contain 10 elements is part of the fine-tuning of the DRL algorithm. Clearly, the more elements of the set, the higher granularity and better performance is expected. However, this comes at the cost of longer time for the DRL to converge. We should note here that we do not have only a single DRL for which increasing the size of the set would not be that visible. Instead, in our settings we have an MARL system and hence increasing the set with only one element would be visible in the learning time. We conduct our experiments on CARRADA dataset [31] which is a recent, open dataset that contains 30 scenes of synchronized sequences of camera and radar images. More details about the dataset can be found in our early work [8].

B. Benchmarking Scheme

Due to the novelty of our developed MA-PDDQL and the iterative contract design, it is difficult to find existing baselines to compare with. Therefore, to evaluate and show the benefits of our novel augmented MA-PDDQL algorithm, we design a baseline scheme called Naive MA-PDDQL based on [32]. Different from the augmented MA-PDDQL, the naive MA-PDDQL uses a partially observed MDP (POMDP) for each agent. Specifically, the POMDP state is defined as

$$\mathcal{S}_{naive}^{\dagger(t)} \triangleq \left\{ a^{\dagger(t-1)}, \pi^{\dagger(t)}, \hat{s}^{(t)}, x^{\dagger(t)}, y^{\dagger(t)} \right\}, \tag{26}$$

The action set and immediate reward function are also scaled down to the case of single observations. Since the state observation of each agent does not represent the whole system state as earlier defined in the augmented MDP, each agent has only a partial observation.

C. Results

1) Convergence Analysis and Validity of the Feasibility Conditions: To observe the convergence behavior of our learning-based iterative contract, we measure the number of IR and IC violations and the revenue of the VSP at the last iteration of each episode. We observe from Fig. 3(a) that the average reward of the augmented MA-PDDQL stabilizes after 350 episodes. However, the average reward of the naive MA-PDDQL is lower than that of the augmented MA-PDDQL and stops increasing after 100 episode only, indicating that the naive MA-PDDQL is not able to converge. As observed from Fig. 3(b), the number of IR and IC violations for the augmented MA-PDDQL decreases as the algorithm progress in learning. Interestingly, we observe from Fig. 3(b) that there is an improvement in the minimization of the IR violations for the naive MA-PDDQL while no significant change occurs for the number of IC violations. This is justified by the fact that the IR constraint is much simpler than the IC constraint. The IR constraint needs only to guarantee that the utility of the participants is non-negative, while the IC constraint needs to guarantee that the utility of a participant is maximized for the true type of the participant compared to all other types. The latter cannot be learned by the naive MA-PDDQL because of the non-stationarity problem of the POMDP. Specifically, as each agent in the naive MA-PDDQL environment observes only its private state, and thus is unaware of other agents changes of their bundles, it is unable to find an optimal adjustment for its bundle to meet the IC constraint for its respective sensing IoT device.

To dive further in the structure of our learning-based algorithm, we plot in Fig. 3(c) the average revenue of the VSP as the training progress. Remarkably, we observe that the average revenue of the VSP decreases as the training progress, which seems to behave against our main objective, i.e., maximization of revenue of the VSP as stated in $\mathcal{P}_1$. However, we should understand from Fig. 3(b) that in the first few episodes, the obtained revenue of the VSP is achieved while having the IR and IC properties violated for the majority of the participants, i.e., sensing IoT devices. This implies that the majority of the sensing IoT devices will behave untruthfully and select contract items not dedicated for their true types, making the realized utility of the VSP very low compared to the expected one. At the end of the training, the derived VSP utility is achieved with majority of the IR and IC satisfied. Here, we should note that an important difference between our learning-based contract and existing works on contract theory is the satisfaction of the feasibility conditions, i.e., IR and IC, under information asymmetry. In classical

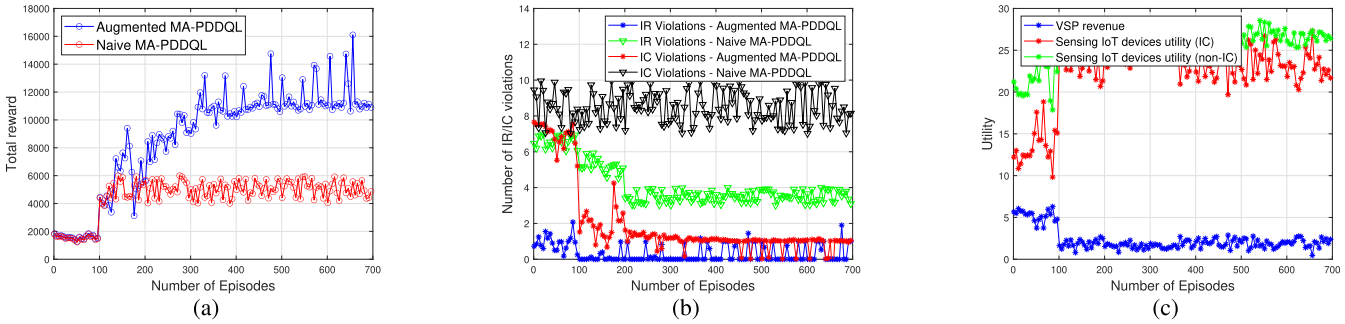


Fig. 3. (a) Total reward for each episode. (b) Average number of IR and IC violations. (c) Average revenue of the VSP and sensing IoT devices.

approaches, the IR and IC constraints need to be satisfied for all the participants, i.e., sensing IoT devices. However, in our framework we aim towards the minimization of their occurrence.

We also plot how the utility of the sensing IoT devices changes as the training progresses in Fig. 3(c). The red color refer to the utility of the sensing IoT devices in the case that they have chosen their designated bundle (i.e., IC preserved). The green color, however, refers to the case when the sensing IoT devices choose the bundle that maximize their utility. In both scenarios, we observe that the utility of the sensing IoT devices increases then stabilize after episode 100 while the revenue of the VSP decreases and then stabilizes, which is counter-intuitive. To explain this behavior, we first note that based on the objective function in $\mathcal{P}_1$, we should only maximize the revenue of the VSP. From (11), it seems that the optimal strategy for the VSP is to set the price for each contract item equal to their valuation by their corresponding sensing IoT devices. In this case, the utility of the sensing IoT devices will be equal to zero. Based on this observation, we expect the revenue of the VSP to increase and the utilities of the participants to decrease. However, as the objective function of problem $\mathcal{P}_1$ is adjusted in the reward function of the MDP, an action that further minimizes or maintains the number of IR and IC violations is given a positive reward (see the last term in (17)). Therefore, the derived bundles are not pushed towards minimizing the gap between the provided prices and the private valuations of each sensing IoT device, which justifies the increase of the utility of the sensing IoT devices.

2) Impact of the Weighting Factors: Motivated by the previous results from Fig. 3(c), we further study the impact of the weighting factors in the reward function on the performance of our framework. In this experiment, we set three different scenarios for the values of w_1, w_2 and w_3 and observe how the VSP revenue and the IR and IC violations changes. Specifically, we consider the following scenarios: $(w_1, w_2, w_3) = (0.33, 0.33, 0.33)$, $(w_1, w_2, w_3) = (0.33, 0.01, 0.66)$ and $(w_1, w_2, w_3) = (0.01, 0.01, 0.98)$. The results are shown in Fig. 4. Interestingly, we observe from Fig. 4(b) that putting a weight of 0.01 on the IR term gives better results for the IR violation compared to the case of setting a higher weight 0.33. This is explained by the fact that IR is satisfied for the majority of the sensing IoT devices, which is dependant on the sets of semantic information size and prices that the MA-PDDQL is learning on. We should also note that putting more weight on the IC term helps reducing the IR violation further and hence, minimizes the influence of the weight term of the IR term. Specifically, during the IC property verification, each sensing IoT device compares its utility when choosing its true type with the case of choosing any other type. Therefore, if its IC property is not violated, its utility is unlikely to be negative.

We also plot in Fig. 5(a) and Fig. 5(b) the average number of IR and IC violations when changing the weighting factors. Specifically, we measure the probability of IR and IC violations to be under 10 % from episode 350 to episode 700. The results are shown in a box plot where on each box, the central mark indicates the median value and the bottom and top values of the box indicate the 25th and 75th percentiles, respectively. Fig. 5(a) validates the previous results that the IR violation rate diminishes as more weight is given to the IC term. Furthermore, we observe from Fig. 5(b) that when more weight is given to the IC term in the reward function, the majority of the IC violation rates are bellow 1 violation only on average. This results indicates that the derived solution to the contract problem is unlikely to violate the feasibility conditions. We also plot in Fig. 5(c) the time complexity of our DRL-based solution. We observe that the MA-PDDQL has a linear complexity of $O(N)$ with respect to the number of devices. This is because at each iteration, the algorithm executes the PDDQL of each agent sequentially.

3) Impact of the Number of Participants and the Number of Contract Items: In this experiment, we vary the number of combinations of the three-dimensional contract types and observe how the VSP revenue and the sensing IoT devices utilities change. The experiment is conducted on different number of participants, i.e., different N sensing IoT devices, as shown in Fig. 6(a) and Fig. 6(b). We set three different scenarios for the number of contract items: 8(2×2x2), 27(3×3x3) and 64(4×4x4). We observe that as the number of participating sensing IoT devices increases, the revenue of the VSP and the utility of the sensing IoT devices increase. This result is expected because more participating sensing IoT devices bring more semantic information to the VSP and hence the VSP gets higher utility. Interestingly, we observe from Fig. 6(b) that as the number of contract items increases, the utilities of the participating sensing IoT devices decreases. This is

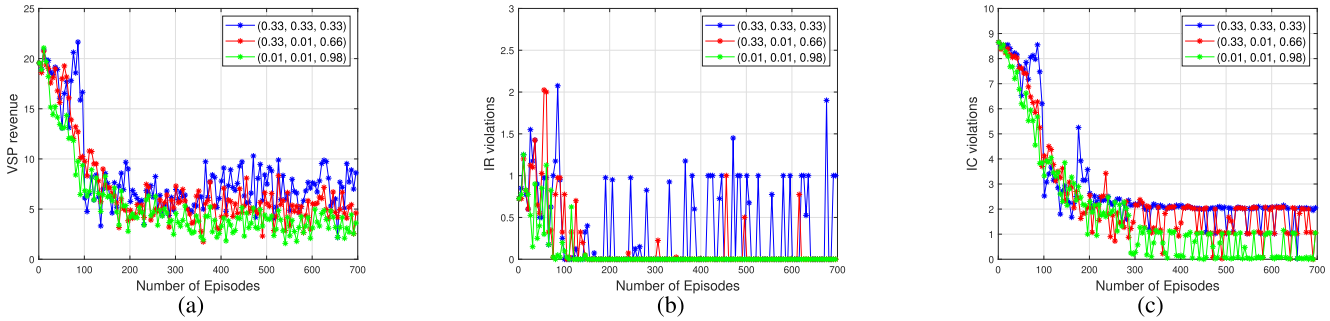


Fig. 4. Impact of the weighting factors: (a) average VSP revenue. (b) and (c) average number of IR and IC violations.

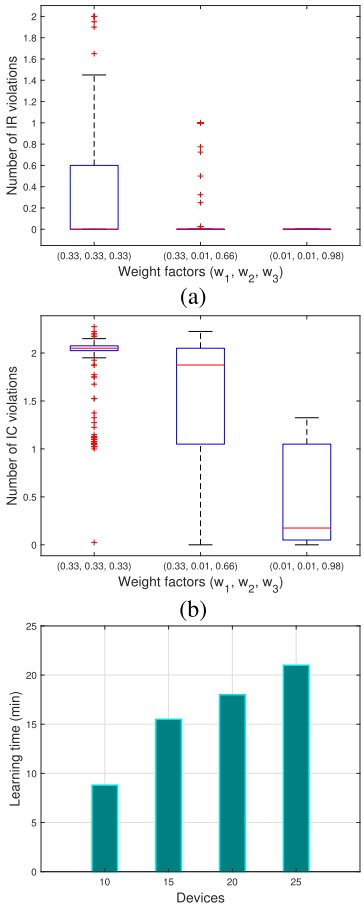


Fig. 5. (a) and (b) average number of IR and IC violations, respectively, when changing the weight factors. (c) Learning time.

due to the fact that as the contract designer is able to derive more specific contract items for each participant based on their private information at a low level of precision, e.g., semantic value or AoI, it can extract more profit, which is reflected by the decrease in the utility of the participants. However, the profit that the VSP receives is marginal as observed from Fig. 6(a), which is due the IR and IC constraints that have to be minimized. The algorithm adjust the contract bundles to satisfy IR and IC with less importance to the maximization of the VSP's revenue.

4) Sensitivity to the Distribution of Types: To further push our proposed framework to the limits, we train the model on a specific distribution of the types and then plug in other distribution and observe how the system reacts. Change of the distribution between training time and testing time is a common problem in DRL, see for example [33], and its important to evaluate our framework for this change. During training time, the joint type of each sensing IoT device is chosen uniformly from the different types' sets. However, at test time, we change the distribution of the joint type by changing the probabilities of each element in each set of types, i.e., Ψ, Λ and Γ. The number of sensing IoT devices is set to 10. The following results are that of the augmented MA-PDDQL algorithm after convergence. The algorithm is given a set of tuples (semantic information sizes and prices) drawn from random values and the objective is to adjust the tuples to find an optimal solution that maximizes the revenue of the VSP and minimizes the number of IR and IC violations. We consider three different scenarios. The first one is to consider the type of contract items drawn from the same distribution of the training time, i.e., uniform distribution. The second scenario is to consider the testing distribution drawn from a set with more weight on the lower types of the sets Ψ, Λ and Γ. In the third scenario, we consider larger weights are given to higher types values. Fig. 6(c) shows the results of this experiment.

We observe that the revenue of the VSP is marginally affected by the change of the distribution. However, the utilities of the sensing IoT devices when the lower types are giving larger weights are greater than that of the case of equal weights. Moreover, the utilities of the sensing IoT devices when the higher types are giving larger weight is less than that of the case of equal weights. We explain this behavior by observing that the DRL-based model is trained on types drawn from a uniform distribution. When faced with devices with lower types only (on all of the three dimensions), the model is not able to optimally minimize the gap between the cost and the price which makes the utilities of lower types devices higher. Similarly, this makes the utility of higher type devices less than that if the model was trained on the same distribution. However, note that the main objective of the VSP is not to minimize the utility of the sensing IoT devices. Instead its main objective, as shown in the objective function of $\mathcal{P}_1$ is to maximize its revenue while guaranteeing IR and IC, which is successfully achieved in all scenarios.

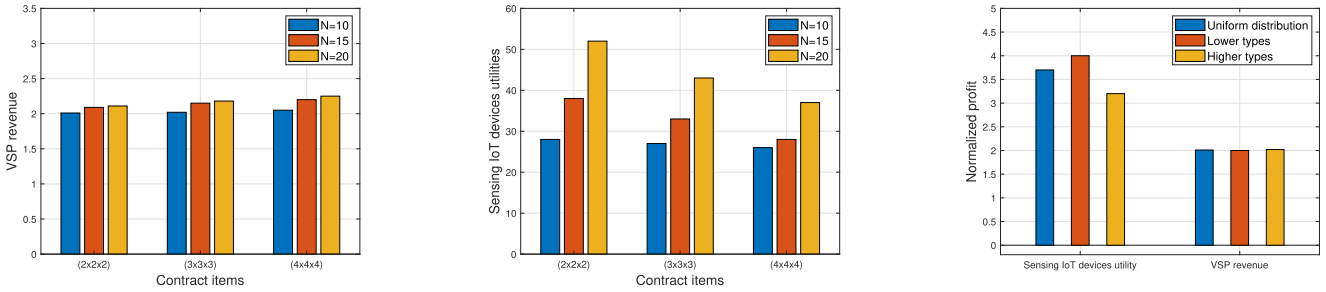


Fig. 6. (a) and (b) VSP revenue and Utilities of the sensing IoT devices for different contract items. (c) Impact of changes in the distribution of the joint types.

VI. Conclusion

In this paper, we design a semantic aware truthful mechanism for the Metaverse based on contract theory and MARL. Specifically, we design a two-layer Metaverse ecosystem where in the first layer, the VSP hires sensing IoT devices to obtain semantic information about the physical environment and render the digital twin. In the second layer, the VSP delivers the constructed digital twin to the Metaverse users. We then use contract theory to design the pricing bundles on both layers. We design a novel architecture for the contract in which the VSP interacts with the participants, i.e., sensing IoT devices or Metaverse users, to derive the optimal set of bundles. This interaction is conducted by using a new variant of MARL that we develop where the VSP creates DRL instances for each participant and sets the objective to maximize its revenue while minimizing the IR and IC violation rates. The simulation results show that our designed framework achieves good performance in terms of maximizing the profit of the VSP while not requiring several assumptions about the system model. As a future work, it is interesting to explore the strategy of joint optimization of the upstream layer and the downstream layer in a single MDP as it is expected to better help the VSP increase its profit and the profit of the Metaverse users.

References

[1] M. Xu et al., "A full dive into realizing the edge-enabled metaverse: Visions, enabling technologies, and challenges," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 1, pp. 656–700, 1st Quart., 2023.

[2] R. Minerva, G. M. Lee, and N. Crespi, "Digital twin in the IoT context: A survey on technical features, scenarios, and architectural models," *Proc. IEEE*, vol. 108, no. 10, pp. 1785–1824, Oct. 2020.

[3] P. Bolton and M. Dewatripont, *Contract Theory*, 1st ed. Cambridge, MA, USA: MIT Press, 2005.

[4] A. Ouaknine, A. Newson, P. Pérez, F. Tupin, and J. Rebut, "Multi-view radar semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15651–15660.

[5] Z. Wang, L. Gao, and J. Huang, "Multi-cap optimization for wireless data plans with time flexibility," *IEEE Trans. Mobile Comput.*, vol. 19, no. 9, pp. 2145–2159, Sep. 2020.

[6] Z. Xiong, J. Kang, D. Niyato, P. Wang, H. V. Poor, and S. Xie, "A multi-dimensional contract approach for data rewarding in mobile networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 9, pp. 5779–5793, Sep. 2020.

[7] M. Xu, D. Niyato, J. Kang, Z. Xiong, C. Miao, and D. I. Kim, "Wireless edge-empowered metaverse: A learning-based incentive mechanism for virtual reality," in *Proc. IEEE Int. Conf. Commun.*, May 2022, pp. 5220–5225.

[8] L. Ismail, D. Niyato, S. Sun, D. I. Kim, M. Erol-Kantarci, and C. Miao, "Semantic information market for the metaverse: An auction based approach," in *Proc. IEEE Future Netw. World Forum (FNWF)*, Oct. 2022, pp. 628–633.

[9] Y. Han, D. Niyato, C. Leung, C. Miao, and D. I. Kim, "A dynamic resource allocation framework for synchronizing metaverse with IoT service and data," in *Proc. IEEE Int. Conf. Commun.*, May 2022, pp. 1196–1201.

[10] O. Hashash, C. Chaccour, W. Saad, K. Sakaguchi, and T. Yu, "Towards a decentralized metaverse: Synchronized orchestration of digital twins and sub-metaverses," 2022, *arXiv:2211.14686*.

[11] J. Park, J. Choi, S.-L. Kim, and M. Bennis, "Enabling the wireless metaverse via semantic multiverse communication," 2022, *arXiv:2212.06908*.

[12] J. Wang, H. Du, Z. Tian, D. Niyato, J. Kang, and X. Shen, "Semantic-aware sensing information transmission for metaverse: A contest theoretic approach," *IEEE Trans. Wireless Commun.*, vol. 22, no. 8, pp. 5214–5228, Aug. 2023, doi: 10.1109/TWC.2022.3232565.

[13] C. Kurisummoottil Thomas and W. Saad, "Neuro-symbolic causal reasoning meets signaling game for emergent semantic communications," 2022, *arXiv:2210.12040*.

[14] K. Tang, H. Zhang, B. Wu, W. Luo, and W. Liu, "Learning to compose dynamic tree structures for visual contexts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6612–6621.

[15] R. Krishna et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," 2016, *arXiv:1602.07332*.

[16] Y. Liu, M. Tian, Y. Chen, Z. Xiong, C. Leung, and C. Miao, "A contract theory based incentive mechanism for federated learning," in *Federated and Transfer Learning*. Cham, Switzerland: Springer, 2022, pp. 117–137.

[17] J. Kang, Z. Xiong, D. Niyato, S. Xie, and J. Zhang, "Incentive mechanism for reliable federated learning: A joint optimization approach to combining reputation and contract theory," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 10700–10714, Dec. 2019.

[18] S. K. Kaul, R. D. Yates, and M. Gruteser, "Status updates through queues," in *Proc. 46th Annu. Conf. Inf. Sci. Syst. (CISS)*, Mar. 2012, pp. 1–6.

[19] R. D. Yates, "Status updates through networks of parallel servers," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 2281–2285.

[20] A. M. Bedewy, Y. Sun, and N. B. Shroff, "Minimizing the age of information through queues," *IEEE Trans. Inf. Theory*, vol. 65, no. 8, pp. 5215–5232, Aug. 2019.

[21] Z. Xiong, J. Zhao, Y. Zhang, D. Niyato, and J. Zhang, "Contract design in hierarchical game for sponsored content service market," *IEEE Trans. Mobile Comput.*, vol. 20, no. 9, pp. 2763–2778, Sep. 2021.

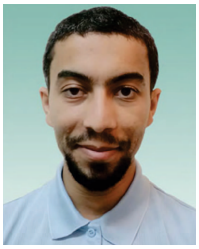
[22] X. Zhou, W. Wang, N. U. Hassan, C. Yuen, and D. Niyato, "Age of information aware content resale mechanism with edge caching," *IEEE Trans. Commun.*, vol. 69, no. 8, pp. 5269–5282, Aug. 2021.

[23] M. F. Bado, D. Tonelli, F. Poli, D. Zonta, and J. R. Casas, "Digital twin for civil engineering systems: An exploratory review for distributed sensing updating," *Sensors*, vol. 22, no. 9, p. 3168, Apr. 2022.

[24] L. Zhang, W. Wu, and D. Wang, "Sponsored data plan: A two-class service model in wireless data networks," in *Proc. ACM SIGMETRICS Int. Conf. Meas. Modeling Comput. Syst.* New York, NY, USA, Jun. 2015, pp. 85–96.

[25] Y. Sun et al., "CS2P: Improving video bitrate selection and adaptation with data-driven throughput prediction," in *Proc. ACM SIGCOMM Conf.*, Aug. 2016, pp. 272–285.

[26] L. Gao, X. Wang, Y. Xu, and Q. Zhang, "Spectrum trading in cognitive radio networks: A contract-theoretic modeling approach," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 4, pp. 843–855, Apr. 2011.
[27] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.
[28] H. Mao, M. Alizadeh, I. Menache, and S. Kandula, "Resource management with deep reinforcement learning," in *Proc. 15th ACM Workshop Hot Topics Netw.*, Nov. 2016, pp. 50–56.
[29] I. Lotfi, D. Niyato, S. Sun, H. T. Dinh, Y. Li, and D. I. Kim, "Protecting multi-function wireless systems from jammers with backscatter assistance: An intelligent strategy," *IEEE Trans. Veh. Technol.*, vol. 70, no. 11, pp. 11812–11826, Nov. 2021.
[30] B. Hu, Q. Yan, and Y. Zheng, "Tracking location privacy leakage of mobile ad networks at scale," in *Proc. IEEE Conf. Comput. Commun. Workshops*, Apr. 2018, pp. 1–2.
[31] A. Ouaknine, A. Newson, J. Rebut, F. Tupin, and P. Pérez, "CARRADA dataset: Camera and automotive radar with range-angle-Doppler annotations," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 5068–5075.
[32] S. Gronauer and K. Diepold, "Multi-agent deep reinforcement learning: A survey," *Artif. Intell. Rev.*, vol. 55, no. 2, pp. 895–943, Apr. 2021.
[33] F. Fang, W. Liang, Y. Wu, Q. Xu, and J.-H. Lim, "Improving generalization of reinforcement learning using a bilinear policy network," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2022, pp. 991–995.



Ismail Lotfi (Member, IEEE) received the B.Eng. degree from the Oran University of Science and Technology, Algeria, in 2015, and the master's degree in emerging networks, security and multimedia from Oran 1 University, Algeria, in 2017. He is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. During the master's degree, he was an Exchange Student with Masaryk University, Czech Republic, working on wireless network security. His main research interests include the area of resource management and optimization in communication networks.

Dusit Niyato (Fellow, IEEE) received the B.Eng. degree from the King Mongkut's Institute of Technology Ladkrabang, Thailand, in 1999, and the Ph.D. degree in electrical and computer engineering from the University of Manitoba, Canada, in 2008. He is currently a Professor with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. His research interests include the area of energy harvesting for wireless communication, the Internet of Things (IoT), and sensor networks.

Sumei Sun (Fellow, IEEE) is currently a Distinguished Institute Fellow and the Acting Executive Director of the Institute for Infocomm Research (I2R), Agency for Science, Technology, and Research (A*STAR), Singapore. Her current research interests include next-generation wireless communications, cognitive communications and networks, and the Industrial Internet of Things. She is the Editor-in-Chief of IEEE OPEN JOURNAL OF VEHICULAR TECHNOLOGY, the Chair of IEEE TRANSACTIONS ON MACHINE LEARNING IN COMMUNICATIONS AND NETWORKING Steering Committee, a member of the board of governors of the IEEE Vehicular Technology Society, and the Member-at-Large of IEEE Communications Society.

Dong In Kim (Fellow, IEEE) received the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 1990. He was a tenured Professor with the School of Engineering Science, Simon Fraser University, Burnaby, BC, Canada. He is currently a Distinguished Professor with the College of Information and Communication Engineering, Sungkyunkwan University, Suwon, South Korea. He was the Founding Editor-in-Chief of IEEE WIRELESS COMMUNICATIONS LETTERS from 2012 to 2015. He was selected the 2019 recipient of the IEEE ComSoc Joseph LoCicero Award for Exemplary Service to Publications. He was the General Chair for IEEE ICC 2022 in Seoul.

Xuemin (Sherman) Shen received the Ph.D. degree in electrical engineering from Rutgers University, New Brunswick, NJ, USA, in 1990. He is currently a University Professor with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research interests include network resource management, network security, the Internet of Things, and 5G and beyond.